Layered Group Sparse Beamforming for Cache-Enabled Green Wireless Networks

Xi Peng, *Student Member, IEEE*, Yuanming Shi, *Member, IEEE*,
Jun Zhang, *Senior Member, IEEE*, and Khaled B. Letaief, *Fellow, IEEE*

Abstract

The exponential growth of mobile data traffic is driving the deployment of dense wireless networks, which will not only impose heavy backhaul burdens, but also generate considerable power consumption. Introducing caches to the wireless network edge is a potential and cost-effective solution to address these challenges. In this paper, we will investigate the problem of minimizing the network power consumption of cache-enabled wireless networks, consisting of the base station (BS) and backhaul power consumption. The objective is to develop efficient algorithms that unify adaptive BS selection, backhaul content assignment and multicast beamforming, while taking account of user QoS requirements and backhaul capacity limitations. To address the NP-hardness of the network power minimization problem, we first propose a generalized layered group sparse beamforming (LGSBF) modeling framework, which helps to reveal the layered sparsity structure in the beamformers. By adopting the reweighted $\ell_1/\,\ell_2$-norm technique, we further develop a convex approximation procedure for the LGSBF problem, followed by a three-stage iterative LGSBF framework to induce the desired sparsity structure in the beamformers. Simulation results validate the effectiveness of the proposed algorithm in reducing the network power consumption, and demonstrate that caching plays a more significant role in networks with higher user densities and less power-efficient backhaul links.

Index Terms

Wireless caching, content-centric wireless networks, multicasting beamforming, layered group sparse beamforming, convex approximation, network power minimization, green communications.

X. Peng, J. Zhang and K. B. Letaief are with the Dept. of ECE at the Hong Kong University of Science and Technology, Hong Kong (email: {xpengab, eejzhang, eekhaled}@ust.hk). K. B. Letaief is also affiliated with Hamad bin Khalifa University, Doha, Qatar (e-mail: kletaief@hbku.edu.qa). Y. Shi is with the School of Information Science and Technology, ShanghaiTech University, Shanghai, China (e-mail: shiym@shanghaitech.edu.cn).

This work is supported by the Hong Kong Research Grant Council under Grant No. 16200214.

I. Introduction

To cater for the unprecedented explosion of mobile data traffic [1], cell densification has been regarded as a key mechanism for further wireless evolution [2]. To effectively manage co-channel interference in dense cellular networks, coordinated multipoint (CoMP) technology, i.e., cooperation among base stations (BSs), has been proposed [3]. However, it requires data sharing among cooperative BSs, which will yield considerable backhaul traffic. Current small cell backhaul solutions, such as xDSL [4] and non-line-of-sight microwave [5], are far from adequate to provide sufficient data rate and thus make the current networks vulnerable to congestion. Caching frequently requested content at the wireless network edge, especially at small BSs [6], has been recently proposed as a cost-effective approach to lower the latency for content delivery and alleviate the heavy burden on backhaul links. Remarkably, caching also has the prominent advantage in improving the network energy efficiency. Since local caching brings the content closer to mobile users (MUs) and enables content delivery without using backhaul links, BS transmit power and backhaul power can be substantially reduced.

Energy efficiency, as an essential concern in green cellular networks, has attracted global attention [7] since it is related to maintaining profitability for cellular operators, as well as reducing the overall environmental effects. Most previous investigations on energy efficiency of cellular networks either ignored the backhaul power consumption [8] or employed simplified models to measure it [9]. As cellular networks will evolve to be progressively dense and heterogeneous, backhaul power consumption will play an increasingly important role in total network power consumption [10]. It is inspiring that caching can be very effective in fundamentally reducing backhaul power consumption. Owing to the recent technology development of caching hardware [11], massive backhaul data can be reduced with energy-efficient caches. Also, the frequent reuse of cached contents implies the potential of cache-enabled networks in energy saving. In this paper, we will investigate network power minimization for cache-enabled wireless networks, by taking both the BS and backhaul power consumption into consideration.

A. Related works

Caching popular contents at small BSs has been attracting a lot of attention. The idea of femtocaching was first proposed in [12] to alleviate backhaul loads for small BSs with low-capacity backhaul links. The caching content could be uncoded or coded, and a coded caching scheme can achieve a global caching gain as discussed in [6]. However, these initial studies

assumed no interference among different communication links, and did not take the impact of wireless channels into account. It was proposed in [13], [14] that caching at BSs will not only provide load balancing gain, but also bring interference cancellation gain and interference alignment gain. Follow-up papers [15]–[17] have shed light on cache-aided wireless communications and interference management under various performance metrics. Aiming at minimizing the download delay, distributed caching algorithms were designed in [15], [16]. The tradeoff between the small BS density and total cache size under a certain outage probability was investigated in [17]. Cooperation among multi-antenna cache-enabled BSs [18]–[20] is promising since caching can reduce the backhaul requirement. Full cooperation was considered in [18] to minimize total transmit power. Dynamic clustering and partial cooperation were adopted in [19]. Moreover, by employing the cloud processing and edge caching, cooperative transmission and low delivery latency can be achieved at the same time [20].

There is a growing concern on energy efficiency in wireless networks. Previous works include transmit power minimization via coordinated beamforming [21]–[24] and adaptive selection of active BSs [7], [25]–[27]. After introducing edge caches, similar approaches have been extended to the cache-enabled wireless networks [18], [28]. With cell densification, backhaul power consumption will become a significant component of the total network power consumption [29]. In [30], energy efficiency for cache-aided networks was optimized by assuming constant transmit power for small cell BSs and wireless backhaul nodes. In [31], caching content placement and multicast association were optimized in order to minimize the overall energy cost. But it only considered the backhaul power of the macro BS and did not count small BSs. In order to minimize the network power consumption, joint beamforming and backhaul data assignment problem was investigated in [9], [19], [24], [32]–[34]. But a comprehensive consideration of traffic-dependent backhaul power consumption, active BS selection, multicast beamforming, and backhaul data assignment is still missing.

There are some preliminary studies on developing sparsity-based approaches for designing wireless networks. Inspired by the success of sparse signal processing techniques such as compressed sensing [35], [36], more structured sparsity patterns have been exploited, including group sparsity [37], overlapping group sparsity [38], and layered group sparsity [39], [40], which yield efficient algorithms. Recent years have witnessed an increasing prevalence of applying sparse optimization to design wireless networks, such as the individual sparsity-inducing norm applied for user admission in [34] and link admission control in [41], and the group sparsity-inducing

norm applied for active remote radio head selection of Cloud-RAN in [33]. Sparse optimization is further applied to joint beamforming and backhaul data assignment design in caching networks [19], [32], which may provide potential solutions for 5G wireless networks. As will be revealed in this paper, network energy minimization in cache-enabled wireless networks involves more complicated sparsity structures, and thus more thorough investigations will be needed.

B. Contributions

The main objective of this work is to minimize the network power consumption for cache-enabled wireless networks, which mainly consists of the BS and backhaul power consumption. In this problem, coupled with the non-convex combinatorial composite objective function, there are non-convex quadratic QoS constraints due to multicast transmission, as well as the challenging ℓ_0-norm per-BS backhaul capacity constraints. As a result, it is a mixed-integer non-linear programming problem, and is NP-hard. In this paper, we propose a systematic framework to develop low-complexity algorithms to solve this challenging problem. Specifically, our main contributions are listed as follows:

1) We adopt a realistic model to evaluate the total network power consumption, incorporating practical power consumption models for BSs and backhaul links. In particular, we allow the BS sleep mode, and consider a traffic-dependent backhaul power consumption model, which is essential to investigate backhaul-limited networks. To make the network power minimization problem tractable, we propose a layered group sparse beamforming (LGSBF) modeling framework, which is able to jointly select active BSs, assign backhaul data, and determine the multicast beamformers. This generalized structured sparse formulation unifies existing approaches [19], [21], [23], [33], and will assist the problem analysis and efficient algorithm design.
2) The LGSBF formulation reveals that adaptive BS selection (i.e., the decision for the active BS set) and backhaul assignment (i.e., the delivery of uncached content via backhaul links) can be achieved by controlling the sparsity structure in multicast beamformers. To solve the problem, we first propose to convexify the original problem via structured group sparsity-inducing norm minimization. The second algorithmic contribution is an iterative search procedure that can effectively determine BS selection and backhaul assignment. Finally, coordinated multicast beamforming is adopted to determine the overall beamformers.

3) Simulation results are provided to demonstrate the effectiveness of our proposed algorithm, and show the performance gain compared with existing approaches, including the coordinated beamforming algorithm [42] and two sparse multicast beamforming algorithms [19], [34]. Moreover, we observe that the network performance can be effectively enhanced by employing edge caching, which shows the potential of caches as effective and efficient alternatives for high-capacity backhaul links. In particular, it is shown that caching can reduce the network power consumption more effectively in networks with higher user densities and with less power-efficient backhaul links.

C. Organization and Notations

The rest of the paper is organized as follows. Section II presents the system model. Section III provides the problem formulation and problem analysis. In Section IV, the LGSBF framework is proposed to minimize the network power consumption. Simulation results are demonstrated in Section V. Finally, Section VI concludes the paper.

Throughout this paper, vectors and matrices are denoted by lower-case and upper-case bold letters, respectively. The ℓ_p-norm is represented by $\|\cdot\|_p$. The indicator function is denoted as $\boldsymbol{I}(\cdot)$, where $\boldsymbol{I}(e) = 1$ if event e is true, and $\boldsymbol{I}(e) = 0$ otherwise. We use $(\cdot)^{\mathsf{T}}$, $(\cdot)^{\mathsf{H}}$, $\mathrm{Tr}(\cdot)$ and $\mathrm{Re}\{\cdot\}$ to denote transpose, Hermitian transpose, trace and real part operators, respectively. Calligraphy letters are used to denote sets.

II. System Model

In this section, we will introduce the communication model, caching and backhaul models, as well as the power consumption model. Then the main performance metrics will be presented.

A. Communication Model

We consider a downlink multicast network consisting of N_U single-antenna MUs cooperatively served by N_B multi-antenna BSs, where the j-th BS has L_j antennas. Each BS is equipped with a cache storage and connected to the central controller via a capacity-limited backhaul link. The central controller has access to the whole data library containing N_F pieces of equal-size content objects. Let $\mathcal{J} = \{1, \ldots, N_B\}$, $\mathcal{K} = \{1, \ldots, N_U\}$ and $\mathcal{F} = \{1, \ldots, N_F\}$ denote the sets of BSs, MUs and content objects, respectively. At the beginning of each interval, each MU makes a content request which follows a content popularity distribution. The MUs requesting the

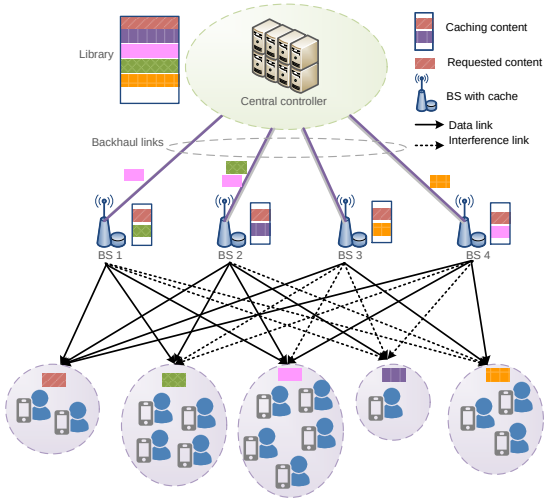


Fig. 1. System model. MUs requesting the same content form a group served by a cluster of BSs via multicast transmission. The requested content is either cached at serving BSs or retrieved from the central controller via corresponding backhaul links.

same content are grouped together and served by a cluster of BSs using multicast transmission. During each interval, the number of multicast groups is N_G ($1 \leq N_G \leq \min\{N_U, N_F\}$), and the set of groups is denoted as $\mathcal{M} = \{1, \ldots, N_G\}$. The set of MUs in group m is denoted as $\mathcal{G}_m, \forall m \in \mathcal{M}$. Since each MU is assumed to request one piece of content during an interval, we have $\mathcal{G}_m \cap \mathcal{G}_i = \emptyset$, for $m \neq i$, and $\sum_{m=1}^{N_G} |\mathcal{G}_m| = N_U$. When BS j caches the content requested by group m, BS j can directly transmit the local content to group m. Otherwise, the uncached content has to be retrieved from the central controller to BS j via the corresponding backhaul link and then transmitted to group m. The system model is illustrated in Fig. 1. The propagation channel from the j-th BS to the k-th MU is denoted as $\mathbf{h}_{kj} \in \mathbb{C}^{L_j}, \forall k, j$, and the transmit beamforming vector from the j-th BS to the multicast group m is denoted as $\mathbf{v}_{jm} \in \mathbb{C}^{L_j}, \forall j, m$. The transmit signal at the j-th BS is given by

$$\mathbf{x}_j = \sum_{m=1}^{N_G} \mathbf{v}_{jm} s_m, \tag{1}$$

where $s_m \in \mathbb{C}$ stands for the encoded information symbol for the multicast group m with $\mathbb{E}\left[|s_m|^2\right] = 1$. The received signal at the MU $k \in \mathcal{G}_m$ is given by

$$y_{km} = \sum_{j=1}^{N_B} \mathbf{h}_{kj}^{\mathsf{H}} \mathbf{v}_{jm} s_m + \sum_{i=1, i\neq m}^{N_G} \sum_{j=1}^{N_B} \mathbf{h}_{kj}^{\mathsf{H}} \mathbf{v}_{ji} s_i + n_k, \forall k \in \mathcal{G}_m, \forall m \in \mathcal{M}, \tag{2}$$

where $n_k \sim \mathcal{CN}(0, \sigma_k^2)$ is the additive Gaussian noise at the k-th MU. Assume that all MUs adopt single user detection and thus treat interference as noise. The signal-to-interference-plus-noise ratio (SINR) at MU $k \in \mathcal{G}_m$ is given by

$$\mathrm{SINR}_k = \frac{\left|\mathbf{h}_k^{\mathsf{H}} \mathbf{v}_m\right|^2}{\sum_{i \neq m}^{N_G} \left|\mathbf{h}_k^{\mathsf{H}} \mathbf{v}_i\right|^2 + \sigma_k^2}, \forall k \in \mathcal{G}_m, \forall m \in \mathcal{M}, \tag{3}$$

where $\mathbf{h}_k = \left[\mathbf{h}_{k1}^{\mathsf{H}}, \mathbf{h}_{k2}^{\mathsf{H}}, \ldots, \mathbf{h}_{kN_B}^{\mathsf{H}}\right]^{\mathsf{H}} \in \mathbb{C}^N$ with $N = \sum_{j=1}^{N_B} L_j$, represents the channel vector from all the BSs to the k-th MU, and $\mathbf{v}_m = \left[\mathbf{v}_{1m}^{\mathsf{H}}, \mathbf{v}_{2m}^{\mathsf{H}}, \ldots, \mathbf{v}_{N_B m}^{\mathsf{H}}\right]^{\mathsf{H}} \in \mathbb{C}^N$ represents the beamforming vector from all the BSs to group m. Let $\mathbf{v} = [\tilde{\mathbf{v}}_j]_{j=1}^{N_B} \in \mathbb{C}^{N_G N}$ denote the aggregate beamforming vector with $\tilde{\mathbf{v}}_j = [\mathbf{v}_{jm}]_{m=1}^{N_G} \in \mathbb{C}^{N_G L_j}$ as the beamforming vector from the j-th BS to all multicast groups, i.e.,

$$\mathbf{v} = \left[\underbrace{\mathbf{v}_{11}^H, \mathbf{v}_{12}^H, \ldots, \mathbf{v}_{1N_G}^H}_{\tilde{\mathbf{v}}_1^H}, \ldots, \underbrace{\mathbf{v}_{j1}^H, \ldots, \mathbf{v}_{jm}^H, \ldots, \mathbf{v}_{jN_G}^H}_{\tilde{\mathbf{v}}_j^H}, \ldots, \underbrace{\mathbf{v}_{N_B 1}^H, \mathbf{v}_{N_B 2}^H, \ldots, \mathbf{v}_{N_B N_G}^H}_{\tilde{\mathbf{v}}_{N_B}^H}\right]^H. \tag{4}$$

To keep the analysis simple, we assume that each BS has the same number of antennas, i.e., $L_j = L, \forall j \in \mathcal{J}$. Define the target SINR vector as $\gamma = [\gamma_1, \ldots, \gamma_{N_G}]$, where γ_m stands for the lowest received SINR threshold for the users in group m. In order to decode the message successfully, any user $k \in \mathcal{G}_m$, should satisfy the following QoS constraint

$$\mathrm{SINR}_k \geq \gamma_m, \forall k \in \mathcal{G}_m, \forall m. \tag{5}$$

Denote the maximum transmit power of the j-th BS as P_j^{TX}, and transmit power constraints are given by

$$\sum_{m=1}^{N_G} \|\mathbf{v}_{jm}\|_2^2 \leq P_j^{\mathrm{TX}}, \forall j \in \mathcal{J}. \tag{6}$$

B. Caching and Backhaul Models

Caching networks operate in two phases, i.e., the prefetching phase and the delivery phase. In the prefetching phase, BSs fetch some contents from the file library of the central controller and store them at local caches, which usually happens during off-peak time. In the delivery phase (usually the busy hours), MUs may request arbitrary content in the file library. Since some desired contents have already been cached locally in the prefetching phase, only the rest of the requested content objects need to be delivered to BSs via backhaul links.

Define a caching matrix $\mathbf{C} = [c_{f,j}] \in \{0,1\}^{N_F \times N_B}$, where $c_{f,j} = 1$ means that the f-th content is cached at the j-th BS. Assume that MUs in group m request content $q_m \in \{1, \ldots, N_F\}$, and

$c_{q_m,j} = 1$ means that the content requested by MUs in group m is cached at the j-th BS. The transmit association status matrix is denoted as $\mathbf{T} = [t_{jm}] \in \{0,1\}^{N_B \times N_G}$, where $t_{jm} = 1$ means that the j-th BS serves group m and $t_{jm} = 0$ means the opposite. Let $\mathbf{N}^{\mathrm{BA}} = [n_{jm}] \in \{0,1\}^{N_B \times N_G}$ denote the backhaul data assignment matrix, where $n_{jm} = 1$ means that the content requested by the m-th user group will be assigned to the j-th BS via its backhaul link. It is not difficult to obtain that

$$n_{jm} = t_{jm}\left(1 - c_{q_m,j}\right). \tag{7}$$

Therefore, only when $t_{jm} = 1$ and $c_{q_m,j} = 0$, it will spawn backhaul traffic to retrieve the requested content, i.e., $n_{jm} = 1$, and otherwise we have $n_{jm} = 0$.

For ease of discussion, we consider fixed and feasible target SINR requirements as in [19]. The transmission data rate for group m is given by $R_m = B_0 \log_2\left(1 + \gamma_m\right)\ (\mathrm{bps})$, where B_0 is the available bandwidth. The data rate (i.e., the traffic load) of backhaul link j is then given by

$$R_j^{\mathrm{BH}} = \sum_{m=1}^{N_G} R_m n_{jm}\ (\mathrm{bps}), \forall j \in \mathcal{J}. \tag{8}$$

Since the capacity of each backhaul link is limited, we consider the following backhaul capacity constraints

$$R_j^{\mathrm{BH}} \leq C_j^{\mathrm{BH}}, \forall j \in \mathcal{J}. \tag{9}$$

C. Power Consumption Model

We focus on the network power consumption of the delivery phase, for which the signal processing and optimization are much more challenging than the prefetching phase. Owing to the advances in caching hardwares [11], caches have been made very energy-efficient. Moreover, once the cache placement is finished in the prefetching phase, it will remain unchanged for a period of time, e.g., several days or weeks, since the file popularity evolves slowly, while user requests happen much more frequently. Therefore, the frequent reuse of cached contents can save considerable backhaul power consumption, which makes the power consumption in the prefetching phase negligible. Also, the cache placement is usually conducted during off-peak hours when the electricity resource is abundant and with a low price. Therefore, we focus on the network power consumption for the delivery phase, and omit the power consumption for caching. The main components of network power consumption, i.e, BS power consumption and backhaul power consumption, will be modeled as follows.

1) BS Power Consumption Model: We adopt the empirical linear model [25] to describe the power consumption of the j-th BS:

$$P_j^{\mathrm{BS}} = \begin{cases} P_{A,j}^{\mathrm{BS}} + \delta_j P_j^{\mathrm{out}}, & \text{if } 0 < P_j^{\mathrm{out}} \leq P_j^{\mathrm{TX}} \\ P_{S,j}^{\mathrm{BS}}, & \text{if } P_j^{\mathrm{out}} = 0 \end{cases}, \tag{10}$$

where $P_{A,j}^{\mathrm{BS}}$ ($P_{S,j}^{\mathrm{BS}}$) stands for the active (sleep) mode power consumption, δ_j represents the slope of the load-dependent power consumption, and P_j^{out} is the BS transmit power, i.e., $P_j^{\mathrm{out}} = \sum_{m=1}^{N_G} \|\mathbf{v}_{jm}\|_2^2 = \|\tilde{\mathbf{v}}_j\|_2^2$. Although a BS's power consumption can be arbitrarily close to zero Watt in the deepest sleep level, it may cause an undesirable long delay to wake up the BS from this low power mode [26]. In practice, when a BS has no transmission tasks, a less deep sleep mode is usually adopted, where only some well-selected parts of the hardware may be inactivated, in order to fasten the activation process. As a result, it is typical to have $P_{S,j}^{\mathrm{BS}} \neq 0$. For instance, according to the survey on BS power consumption [25], for a 2-antenna pico-BS, the typical values are $P_{A,j}^{\mathrm{BS}} = 6.8$ W, $P_{S,j}^{\mathrm{BS}} = 4.3$ W and $\delta_j = 4$. Let $\mathcal{A} \subseteq \mathcal{J}$ and $\mathcal{Z} \subseteq \mathcal{J}$ denote the sets of active BSs and inactive BSs, respectively. Then, the total BS power consumption is given by

$$\hat{p}_1 = \sum_{j \in \mathcal{A}} \left(P_{A,j}^{\mathrm{BS}} + \delta_j \sum_{m=1}^{N_G} \|\mathbf{v}_{jm}\|_2^2 \right) + \sum_{j \in \mathcal{Z}} P_{S,j}^{\mathrm{BS}}. \tag{11}$$

Based on the BS power consumption model, we conclude that it is essential to put BSs into sleep mode whenever possible in order to save the power consumption.

2) Backhaul Transport Power Consumption Model: The total backhaul transport power consumption is given by

$$\hat{p}_2 = \sum_{j=1}^{N_B} P_j^{\mathrm{BH}}, \tag{12}$$

where P_j^{BH} is the power consumption of the backhaul link corresponding to BS j. Similar to the BS power consumption model, we need to consider both active and sleep modes for backhaul links. The power consumption of an active backhaul link turns out to be traffic-dependent [43]. Therefore, the backhaul transport power consumption is expressed as

$$P_j^{\mathrm{BH}} = \begin{cases} P_{A,j}^{\mathrm{BH}} + \frac{R_j^{\mathrm{BH}}}{C_j^{\mathrm{BH}}} P_j^{\max}, & \text{if } 0 < R_j^{\mathrm{BH}} \leq C_j^{\mathrm{BH}} \\ P_{S,j}^{\mathrm{BH}}, & \text{if } R_j^{\mathrm{BH}} = 0 \end{cases}, \forall j \in \mathcal{J}, \tag{13}$$

where C_j^{BH} denotes the maximum data rate (i.e., capacity) of the backhaul link, $P_j^{\max}$ represents the backhaul power consumption when supporting the maximum data rate, and $E_j^{\mathrm{BH}} \triangleq$

$P_j^{\max}/C_j^{\mathrm{BH}}$ is the backhaul transport energy coefficient. For a backhaul link, typical values are $P_{A,j}^{\mathrm{BH}} = 3.85$ W, $P_{S,j}^{\mathrm{BH}} = 0.75$ W. The typical value for E_j^{BH} is around $10^{-7}\,\mathrm{J/bit}$ for microwave backhaul link [43], and around $10^{-5}\,\mathrm{J/bit}$ for copper DSL [4]. The power consumption of all backhaul links can be calculated as

$$\hat{p}_2 = \sum_{j\in\mathcal{A}} \left(P_{A,j}^{\mathrm{BH}} + E_j^{\mathrm{BH}} R_j^{\mathrm{BH}}\right) + \sum_{j\in\mathcal{Z}} P_{S,j}^{\mathrm{BH}}. \tag{14}$$

Combining formula (8), (11) and (14), we have the total network power consumption as

$$\tilde{p}\left(\mathcal{A}, \mathbf{T}, \mathbf{v}\right) = \hat{p}_1 + \hat{p}_2 \tag{15}$$

$$= \sum_{j\in\mathcal{A}} \delta_j \sum_{m=1}^{N_G} \|\mathbf{v}_{jm}\|_2^2 + \sum_{j\in\mathcal{A}} \sum_{m=1}^{N_G} E_j^{\mathrm{BH}} R_m n_{jm} + \sum_{j\in\mathcal{A}} P_j^{\mathrm{D}} + \sum_{j\in\mathcal{J}} \left(P_{S,j}^{\mathrm{BS}} + P_{S,j}^{\mathrm{BH}}\right), \tag{16}$$

where

$$P_j^{\mathrm{D}} = \left(P_{A,j}^{\mathrm{BS}} - P_{S,j}^{\mathrm{BS}}\right) + \left(P_{A,j}^{\mathrm{BH}} - P_{S,j}^{\mathrm{BH}}\right) \tag{17}$$

is the difference of static state power consumption between active and sleep modes for BS j and its corresponding backhaul link, and is named as the *relative power consumption* for simplification. As a matter of fact, usually we have $P_{A,j}^{\mathrm{BS}} > P_{S,j}^{\mathrm{BS}}$ and $P_{A,j}^{\mathrm{BH}} > P_{S,j}^{\mathrm{BH}}$, and thus $P_j^{\mathrm{D}} > 0$. Let $\beta_{jm} = E_j^{\mathrm{BH}} R_m$ denote the backhaul power consumption for BS j for serving user group m. Since constant terms will not influence the optimization design, we can equivalently minimize the re-defined network power consumption instead of (16):

$$p\left(\mathcal{A}, \mathbf{N}^{\mathrm{BA}}, \mathbf{v}\right) = \sum_{j\in\mathcal{A}} \delta_j \sum_{m=1}^{N_G} \|\mathbf{v}_{jm}\|_2^2 + \sum_{j\in\mathcal{A}} \sum_{m=1}^{N_G} \beta_{jm} n_{jm} + \sum_{j\in\mathcal{A}} P_j^{\mathrm{D}}, \tag{18}$$

which consists of BS transmit power consumption, traffic-dependent backhaul power consumption, and relative power consumption of active BSs and corresponding backhaul links.

III. Problem Formulation and Analysis

In this section, we will first formulate the network power minimization problem, which will then be analyzed and reformulated to reveal the layered group sparsity structure in the optimization variables. Based on (18), there are three strategies minimizing the network power consumption: i) to reduce the relative power consumption by switching off as many BSs and corresponding backhaul links as possible; ii) to reduce the transmit power consumption of BSs with coordinated beamforming by having more active BSs; and iii) to reduce the traffic-dependent backhaul power consumption by minimizing backhaul delivery of uncached content. Obviously,

these strategies cannot be achieved at the same time. Hence, the network power consumption minimization problem will be a joint design across BS selection, backhaul data assignment and coordinated transmit beamforming.

A. Problem Formulation

In this work, we assume that perfect channel state information (CSI) $\{\mathbf{h}_k\}$, cache placement $\mathbf{C}$, and overall user requests $\{c_{q_m,j}\}$ are known a priori at the central controller. Considering MU QoS requirements, BS transmit power constraints and per-BS backhaul capacity constraints, we formulate the network power consumption minimization problem as a joint active BS selection, backhaul data assignment and transmit beamforming design problem:

$$\mathscr{P}\colon \underset{\mathcal{A},\{n_{jm}\},\{\mathbf{v}_{jm}\}}{\text{minimize}} \quad p\left(\mathcal{A}, \mathbf{N}^{\mathrm{BA}}, \mathbf{v}\right) \tag{19}$$

$$\text{subject to} \quad \frac{\left|\mathbf{h}_k^{\mathsf{H}}\mathbf{v}_m\right|^2}{\sum_{i\neq m}^{N_G}\left|\mathbf{h}_k^{\mathsf{H}}\mathbf{v}_i\right|^2+\sigma_k^2} \geq \gamma_m, \forall k \in \mathcal{G}_m, \forall m \in \mathcal{M} \tag{19a}$$

$$\sum_{m=1}^{N_G}\|\mathbf{v}_{jm}\|_2^2 \leq P_j^{\mathrm{TX}}, \forall j \in \mathcal{J} \tag{19b}$$

$$\sum_{m=1}^{N_G} R_m n_{jm} \leq C_j^{\mathrm{BH}}, \forall j \in \mathcal{J} \tag{19c}$$

$$\mathbf{N}^{\mathrm{BA}} = [n_{jm}] \in \{0,1\}^{N_B \times N_G}. \tag{19d}$$

In the following subsection, we will analyze problem $\mathscr{P}$, which will motivate us to reformulate it for developing low-complexity algorithms.

B. Problem Analysis

In this subsection, we will identify the main challenges of the network power minimization problem $\mathscr{P}$. We first consider the case with a given active BS set $\mathcal{A}$ and a given backhaul data assignment matrix $\mathbf{N}^{\mathrm{BA}}$, resulting in a transmit power minimization problem given by

$$\mathscr{P}\left(\mathcal{A}, \mathbf{N}^{\mathrm{BA}}\right)\colon \underset{\{\mathbf{v}_{jm}\}}{\text{minimize}} \quad \sum_{j\in\mathcal{A}}\delta_j\sum_{m=1}^{N_G}\|\mathbf{v}_{jm}\|_2^2 \tag{20}$$

$$\text{subject to} \quad (19\text{a}), (19\text{b}),$$

which is a multicast beamforming problem as discussed in [44].

The above analysis implies that once the optimal $\mathcal{A}$ and $\mathbf{N}^{\mathrm{BA}}$ are identified, the solution $\mathbf{v}$ can be determined by solving problem $\mathscr{P}\left(\mathcal{A}, \mathbf{N}^{\mathrm{BA}}\right)$. Thus, problem $\mathscr{P}$ can be solved by searching over all the possible active BS sets and all possible $\mathbf{N}^{\mathrm{BA}}$'s, i.e.,

$$p^{\star} = \underset{Q \in \{A, \ldots, N_B\}}{\text{minimize}} \quad p^{\star}(Q), \tag{21}$$

where $A \geq 1$ is the minimum number of active BSs to meet the QoS constraints, and $p^{\star}(Q)$ is determined by

$$p^{\star}(Q) = \underset{\substack{\mathcal{A} \subseteq \mathcal{J}, |\mathcal{A}| = Q \\ \mathbf{N}^{\mathrm{BA}} \in \{0,1\}^{N_B \times N_G}}}{\text{minimize}} \quad p^{\star}\left(\mathcal{A}, \mathbf{N}^{\mathrm{BA}}\right), \tag{22}$$

where $p^{\star}\left(\mathcal{A}, \mathbf{N}^{\mathrm{BA}}\right)$ is the optimal value of problem $\mathscr{P}\left(\mathcal{A}, \mathbf{N}^{\mathrm{BA}}\right)$ and $|\mathcal{A}|$ is the cardinality of set $\mathcal{A}$. Since the number of subsets $\mathcal{A}$ of size a is $\binom{N_B}{a}$ and we need to search over $2^{N_B N_G}$ possible $\mathbf{N}^{\mathrm{BA}}$'s for each subset $\mathcal{A}$, the complexity of the overall search procedure will grow exponentially with $N_B(N_G+1)$, which makes this approach unscalable. Therefore, the key to solve the problem is to effectively determine $\mathcal{A}^{\star}$ and $\mathbf{N}^{\mathrm{BA}\star}$. This problem needs to be reformulated to develop more efficient algorithms.

C. Layered Group Sparse Beamforming Formulation

In the following, we will reformulate the original problem. First, let us present several key observations, aiming at exploiting the unique structure of the problem, which will help us address the main challenges. The original objective can be decomposed into three parts, i.e.,

$$p\left(\mathcal{A}, \mathbf{N}^{\mathrm{BA}}, \mathbf{v}\right) = T(\mathbf{v}) + F_1(\mathcal{A}) + F_2\left(\mathcal{A}, \mathbf{N}^{\mathrm{BA}}\right), \tag{23}$$

where $T(\mathbf{v}) = \sum_{j=1}^{N_B} \sum_{m=1}^{N_G} \delta_j \|\mathbf{v}_{jm}\|_2^2$ is the BS transmit power consumption, $F_1(\mathcal{A}) = \sum_{j \in \mathcal{A}} P_j^{\mathrm{D}}$ is the relative power consumption, and $F_2\left(\mathcal{A}, \mathbf{N}^{\mathrm{BA}}\right) = \sum_{j \in \mathcal{A}} \sum_{m=1}^{N_G} \beta_{jm} n_{jm}$ is the backhaul power consumption. We will show that F_1 and F_2 can be expressed as functions of the aggregate beamforming vector $\mathbf{v}$, which are able to indicate the group sparsity of $\mathbf{v}$ at different layers.

1) BS-layer Group Sparsity of $\mathbf{v}$*:* All the coefficients in a given vector $\tilde{\mathbf{v}}_j = [\mathbf{v}_{jm}]_{m=1}^{N_G} \in \mathbb{C}^{N_G L_j}$ form a *BS-layer* group and $\sum_{j=1}^{N_B} \boldsymbol{I}\left(\|\tilde{\mathbf{v}}_j\|_2 > 0\right)$ can be considered as a group sparsity measure of $\mathbf{v}$. When the j-th BS is switched off, all the coefficients in vector $\tilde{\mathbf{v}}_j$ will be set to zero, i.e., $\tilde{\mathbf{v}}_j = \mathbf{0}$. It is possible that multiple BSs can be switched off and the corresponding beamformers will be set to zero, which means that $\mathbf{v}$ has a *BS-layer* group sparsity structure. It is observed

that if $\|\tilde{\mathbf{v}}_j\|_2 > 0$, then we have $j \in \mathcal{A}$, and if $\|\tilde{\mathbf{v}}_j\|_2 = 0$, we have $j \in \mathcal{Z}$. Therefore, for a given beamformer $\mathbf{v}$, the relative power consumption $F_1(\mathcal{A})$ can be rewritten as

$$F_1(\mathbf{v}) = \sum_{j=1}^{N_B} P_j^{\mathrm{D}} I\left(\|\tilde{\mathbf{v}}_j\|_2 > 0\right). \tag{24}$$

2) Data Assignment-layer Group Sparsity of $\mathbf{v}$*:* The backhaul data assignment matrix $\mathbf{N}^{\mathrm{BA}}$ can be fully specified with the knowledge of the beamformer $\mathbf{v}$ as

$$n_{jm} = (1 - c_{q_m,j})\, \boldsymbol{I}\left(\|\mathbf{v}_{jm}\|_2 > 0\right). \tag{25}$$

From (25), we observe that for a given user group m, when BS j does not serve it, i.e., $\mathbf{v}_{jm} = \mathbf{0}$, there is no need to assign the content requested by this user group to BS j, and n_{jm} will be set to zero; when the content requested by user group m happens to be cached at BS j, i.e., $c_{q_m,j} = 1$, regardless of whether BS j serves this user group or not, there is no need to assign the content requested by this user group to BS j, and hence n_{jm} will always be set to zero. It is likely that we can reduce the number of backhaul data assignments and the corresponding n_{jm} values will be set to zero, from which we can infer that the backhaul data assignment matrix $\mathbf{N}^{\mathrm{BA}}$ has a sparsity structure. In addition, we observe

$$\left\|\mathbf{N}^{BA}\right\|_0 \leq \sum_{j=1}^{N_B} \sum_{m=1}^{N_G} \boldsymbol{I}\left(\|\mathbf{v}_{jm}\|_2 > 0\right), \tag{26}$$

which means that minimizing $\sum_{j=1}^{N_B} \sum_{m=1}^{N_G} \boldsymbol{I}\left(\|\mathbf{v}_{jm}\|_2 > 0\right)$ can imply the minimization of $\left\|\mathbf{N}^{\mathrm{BA}}\right\|_0$. All the coefficients in a given vector $\mathbf{v}_{jm} \in \mathbb{C}^{L_j}$ form a group and $\sum_{j=1}^{N_B} \sum_{m=1}^{N_G} \boldsymbol{I}\left(\|\mathbf{v}_{jm}\|_2 > 0\right)$ can be considered as another group sparsity measure of $\mathbf{v}$. Since this measure is related to the backhaul data assignment, it can be regarded to represent the "*data assignment-layer*" group sparsity. Hence, the backhaul power consumption $F_2\left(\mathcal{A}, \mathbf{N}^{\mathrm{BA}}\right)$ can be rewritten as

$$F_2(\mathbf{v}) = \sum_{j=1}^{N_B} \sum_{m=1}^{N_G} \beta_{jm}\,(1 - c_{q_m,j})\, \boldsymbol{I}\left(\|\mathbf{v}_{jm}\|_2 > 0\right) I\left(\|\tilde{\mathbf{v}}_j\|_2 > 0\right). \tag{27}$$

For a given BS j, $\{\mathbf{v}_{jm}\}$ are non-overlapping subgroups of $\tilde{\mathbf{v}}_j$, and $\|\mathbf{v}_{jm}\|_2 > 0$ is a sufficient condition for $\|\tilde{\mathbf{v}}_j\|_2 > 0$. As a result, $F_2(\mathbf{v})$ can be simplified as

$$F_2(\mathbf{v}) = \sum_{j=1}^{N_B} \sum_{m=1}^{N_G} \beta_{jm}\,(1 - c_{q_m,j})\, \boldsymbol{I}\left(\|\mathbf{v}_{jm}\|_2 > 0\right). \tag{28}$$

TABLE I

GENERALIZED GROUP SPARSE BEAMFORMING MODEL

Parameters	Problem	Algorithm
$\lambda_1 = 0, \lambda_2 = 0$	Transmit power minimization	Coordinated beamforming [21]
$\lambda_1 > 0, \lambda_2 = 0$	BS selection	Group sparse beamforming [19], [23], [33]
$\lambda_1 = 0, \lambda_2 > 0$	Backhaul data assignment	
$\lambda_1 > 0, \lambda_2 > 0$	BS selection + backhaul data assignment	Proposed layered group sparse beamforming

Based on the above discussions, the network power minimization problem $\mathscr{P}$ can be equivalently reformulated as the following group sparse beamforming problem:

$$\mathscr{P}^{LGSBF}\text{: } \underset{\mathbf{v}}{\text{minimize}} \quad p^{LGSBF}(\mathbf{v}) = T(\mathbf{v}) + F_1(\mathbf{v}) + F_2(\mathbf{v}) \tag{29}$$

$$\text{subject to} \quad \sum_{m=1}^{N_G} R_m (1 - c_{q_m,j}) \boldsymbol{I}\left(\|\mathbf{v}_{jm}\|_2 > 0\right) \leq C_j^{\mathrm{BH}}, \forall j \in \mathcal{J}, \tag{29a}$$

$$(19\mathrm{a}), (19\mathrm{b}).$$

The equivalence between problem $\mathscr{P}$ and problem $\mathscr{P}^{LGSBF}$ means that if $\mathbf{v}^\star$ is a solution to problem $\mathscr{P}^{LGSBF}$, then $\left(\mathcal{A}^\star, \left\{n_{jm}^\star\right\}, \left\{\mathbf{v}_{jm}^\star\right\}\right)$ with $\mathcal{A}^\star = \left\{j \,\middle|\, \left\|\tilde{\mathbf{v}}_j^\star\right\|_2 > 0, j \in \mathcal{J}\right\}$ and $n_{jm}^\star = (1 - c_{q_m,j}) \boldsymbol{I}\left(\left\|\mathbf{v}_{jm}^\star\right\|_2 > 0\right)$ is a solution to problem $\mathscr{P}$, and vice versa.

The incorporation of two group-sparsity measures in our problem formulation generalizes those in previous works [19], [33] which considered only one group-sparsity measure. Notice that all these group sparse beamforming problems can be unified in the following generalized group structured optimization problem:

$$\underset{\mathbf{v}}{\text{minimize}} \quad T(\mathbf{v}) + \lambda_1 \sum_{j=1}^{N_B} \alpha_j I\left(\|\tilde{\mathbf{v}}_j\|_2 > 0\right) + \lambda_2 \sum_{j=1}^{N_B} \sum_{m=1}^{N_G} \eta_{jm} \boldsymbol{I}\left(\|\mathbf{v}_{jm}\|_2 > 0\right) \tag{30}$$

$$\text{subject to} \quad (19\mathrm{a}), (19\mathrm{b}), (29\mathrm{a}),$$

where $T(\mathbf{v}) = \sum_{j=1}^{N_B} \sum_{m=1}^{N_G} \delta_j \|\mathbf{v}_{jm}\|_2^2$ is a smoothed convex function, and $\lambda_k \geq 0, \forall k \in \{1, 2\}$ are regularization parameters for groups at different layers. When $\{\lambda_k\}$ take variant combinations, the model falls into different problems, as shown in Table I. In our formulation, we have $\lambda_1 > 0, \lambda_2 > 0$, which means that we incorporate multiple sparsity-inducing regularizers into the objective function, and therefore enable joint optimization of BS selection and backhaul data assignment, which generalizes the previous works. Specifically, the entries of $\mathbf{v}$ are partitioned

into different groups at two layers: i) the BS-layer where the beamforming coefficients sent from each BS form a group (the number of groups of this layer is N_B), and ii) the data assignment-layer where the beamforming coefficients associated with one BS and one user group are considered as a group (the number of groups of this layer is $N_B N_G$). Furthermore, the previous works [19], [33] failed to take the per-BS backhaul capacity constraints into consideration, which restricts their practical applications. In order to solve problem $\mathscr{P}^{LGSBF}$, we are confronted with several unique challenges which are highlighted as follows.

3) Combinatorial Objective Function: There are two indicator functions in $p^{LGSBF}(\mathbf{v})$, acting as two group-sparsity measures inducing group sparsity at different layers to the problem. Moreover, the variables in the two group-sparsity measures are non-separable. All the existing group sparse beamforming methods [19], [24], [33], [34] can only deal with one group-sparsity measure and are not applicable to our problem.

4) Non-convex Quadratic QoS Constraints: The non-convex quadratic QoS constraints are yielded by the physical-layer multicast beamforming problem. Consider problem $\mathscr{P}\left(\mathcal{A}, \mathbf{N}^{\mathrm{BA}}\right)$ as an example. On one hand, a ready approach to deal with these constraints is to apply a semidefinite relaxation (SDR) technique [44], and relax problem $\mathscr{P}\left(\mathcal{A}, \mathbf{N}^{\mathrm{BA}}\right)$ into a semidefinite programming (SDP) problem by removing the rank-one constraints, at the price of lifting the variables to higher dimensions. On the other hand, the non-convex quadratic QoS constraints can also be rewritten into the DC form [19]. Compared to the SDP transformation, the DC transformation will not incur loss of optimality since it does not involve rank-one constraints. Moreover, the number of variables in the SDP transformation is almost the square of that in problem $\mathscr{P}\left(\mathcal{A}, \mathbf{N}^{\mathrm{BA}}\right)$, while the number of variables in the DC transformation nearly remains the same as that in problem $\mathscr{P}\left(\mathcal{A}, \mathbf{N}^{\mathrm{BA}}\right)$.

5) Non-convex per-BS Backhaul Capacity Constraints: Besides the aforementioned difficulties, the discrete indicator function $\boldsymbol{I}\left(\|\mathbf{v}_{jm}\|_2 > 0\right)$ in per-BS backhaul capacity constraint, which characterizes whether BS j serves user group m, makes the problem much more challenging. A key observation is that the indicator function can be equivalently expressed as an ℓ_0-norm of a scalar. The ℓ_0-norm stands for the number of nonzero entries in a vector, and reduces to an indicator function in the scalar case. By using ideas from previous literature [24], we may further approximate the non-convex ℓ_0-norm by a convex reweighted ℓ_1-norm.

Based on the challenges identified above, we will propose a low-complexity algorithm to solve the problem efficiently based on the formulation $\mathscr{P}^{LGSBF}$ in the following section.

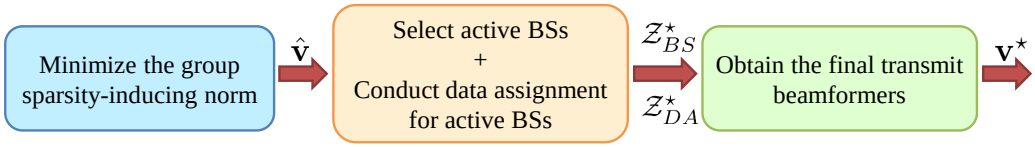


Fig. 2. Proposed generalized three-stage LGSBF framework.

IV. Layered Group Sparse Beamforming Framework

In this section, based on the formulation $\mathscr{P}^{LGSBF}$, we will develop a low-complexity algorithm. The main motivation is to induce group sparsity in the aggregate beamformer $\mathbf{v}$ at both the BS-layer and the data assignment-layer to minimize the total network power consumption. The proposed framework has three stages, as shown in Fig. 2. At the first stage, we solve a reweighted group sparsity-inducing norm minimization problem, so as to induce a group sparsity structure in the aggregate beamformer. Then, in the second stage, based on the approximately sparse beamformer obtained from the first stage, we will conduct a two-layer iterative search procedure, which can efficiently identify the active BSs and backhaul data assignment, respectively. In the last stage, with the knowledge of the active BS set and backhaul data assignment, coordinated multicast beamforming will be adopted to obtain the final beamformers. The details will be presented in the following subsections.

A. Preliminaries and Motivation of LGSBF framework

In order to induce group sparsity in the aggregate beamformer $\mathbf{v}$, we first replace the indicator functions by ℓ_0-norm, which is thereafter relaxed into the mixed ℓ_1/ℓ_p-norm ($p>1$) [45]. The mixed ℓ_1/ℓ_2-norm and ℓ_1/ℓ_∞-norm are two commonly used norms (also called regularizers) for inducing group sparsity. The mixed ℓ_1/ℓ_2-norm is the most common choice and known as the group least-absolute selection and shrinkage operator (group Lasso). In this study, we also adopt $p=2$, and obtain a convex approximation for the objective $p^{LGSBF}(\mathbf{v})$ as

$$\widehat{p}(\mathbf{v}) = \sum_{j=1}^{N_B}\sum_{m=1}^{N_G} \delta_j \left\|\mathbf{v}_{jm}\right\|_2^2 + \sum_{j=1}^{N_B} P_j^{\mathrm{D}} \tilde{\omega}_j \left\|\tilde{\mathbf{v}}_j\right\|_2 + \sum_{j=1}^{N_B}\sum_{m=1}^{N_G} \beta_{jm}\left(1-c_{q_m,j}\right)\omega_{jm}\left\|\mathbf{v}_{jm}\right\|_2, \tag{31}$$

where $\{\tilde{\omega}_j\}, j=1,\ldots,N_B$, and $\{\omega_{jm}\}, j=1,\ldots,N_B, m=1,\ldots,N_G$, are positive weights. Compared with existing sparse beamforming methods dealing with only one sparsity-inducing regularizer [19], [34], the problem $\mathscr{P}^{LGSBF}$ is even more complicated since its objective function

has incorporated two sparsity-inducing regularizers. The multiple sparsity-inducing regularizers indicate that the solution has a layered group sparse pattern, based on which we name the proposed framework as a *layered group sparse beamforming* (LGSBF) framework. Moreover, in our problem, we are facing noncovex constraints as discussed in Section III, which add more challenges.

B. Stage I: Group Structured Sparsity Inducing Norm Minimization

In this subsection, we propose a convex relaxation for problem $\mathscr{P}^{LGSBF}$. To start with, we adopt the DC transformation to deal with the non-convex quadratic QoS constraints, which are rewritten as

$$\gamma_k \left(\sum_{i \neq m}^{N_G} \left| \mathbf{h}_k^{\mathsf{H}} \mathbf{v}_i \right|^2 + \sigma_k^2 \right) - \left| \mathbf{h}_k^{\mathsf{H}} \mathbf{v}_m \right|^2 \leq 0, \forall k \in \mathcal{G}_m, \forall m \in \mathcal{M}. \tag{32}$$

Then, we need to address the indicator functions in both the objective function and backhaul capacity constraints. As stated in (31), we use the convex surrogate $\widehat{p}\left(\mathbf{v}\right)$ for the objective function $p^{LGSBF}\left(\mathbf{v}\right)$ by employing the mixed $\ell_1/\,\ell_2$-norm to approximate the nonconvex ℓ_0-norm. The indicator function can be equivalently expressed as an ℓ_0-norm of another scalar $\left\|\mathbf{v}_{jm}\right\|_2^2$ instead of $\left\|\mathbf{v}_{jm}\right\|_2$, i.e.,

$$I\left(\left\|\tilde{\mathbf{v}}_j\right\|_2 > 0\right) = \left\|\left\|\tilde{\mathbf{v}}_j\right\|_2^2\right\|_0, \text{ and } \boldsymbol{I}\left(\left\|\mathbf{v}_{jm}\right\|_2 > \mathbf{0}\right) = \left\|\left\|\mathbf{v}_{jm}\right\|_2^2\right\|_0, \tag{33}$$

which allows us to extend the mixed $\ell_1/\,\ell_2$-norm approximation. In order to further enhance sparsity, we employ an iterative re-weighted $\ell_1/\,\ell_2$-minimization, inspired by the reweighted ℓ_1-minimization proposed in [46]. The surrogate objective is rewritten as

$$\tilde{p}\left(\mathbf{v}\,|\tilde{\omega}_j, \omega_{jm}\right) = \sum_{j=1}^{N_B}\sum_{m=1}^{N_G} \delta_j \left\|\mathbf{v}_{jm}\right\|_2^2 + \sum_{j=1}^{N_B} P_j^{\mathrm{D}} \tilde{\omega}_j \left\|\tilde{\mathbf{v}}_j\right\|_2^2 + \sum_{j=1}^{N_B}\sum_{m=1}^{N_G} \beta_{jm}\left(1 - c_{q_m,j}\right)\omega_{jm}\left\|\mathbf{v}_{jm}\right\|_2^2, \tag{34}$$

and the problem is reformulated as

$$\mathscr{P}^{DC}: \underset{\mathbf{v}}{\text{minimize}} \quad \tilde{p}\left(\mathbf{v}\,|\tilde{\omega}_j, \omega_{jm}\right) \tag{35}$$

$$\text{subject to} \quad \sum_{m=1}^{N_G} R_m\left(1 - c_{q_m,j}\right)\omega_{jm}\left\|\mathbf{v}_{jm}\right\|_2^2 \leq C_j^{\mathrm{BH}}, \forall j \in \mathcal{J} \tag{35a}$$

$$(19\mathrm{b}), (32),$$

where $\tilde{\omega}_j$ is a weight associated with the j-th BS, and ω_{jm} is a weight associated with the j-th BS and the m-th user group. Similar to [46], we develop the iterative weight update rules as

$$\tilde{\omega}_j = \frac{1}{\|\tilde{\mathbf{v}}_j\|_2^2 + \tau}, \text{ and } \omega_{jm} = \frac{1}{\|\mathbf{v}_{jm}\|_2^2 + \tau}, \forall j \in \mathcal{J}, \forall m \in \mathcal{M}, \tag{36}$$

with $\tilde{\mathbf{v}}_j$ and $\mathbf{v}_{jm}$ obtained from the previous iteration and a small constant parameter $\tau > 0$. Since the beamformer $\tilde{\mathbf{v}}_j$ (or $\mathbf{v}_{jm}$) with a lower transmit power usually has less impact, its transmit power should be encouraged to be further reduced, and eventually forced to zero, in order to switch off this BS (and its backhaul data delivery). Consequently, we are motivated to design weight updating rules (36) where $\tilde{\omega}_j$ and ω_{jm} are inversely proportional to the transmit power. The small parameter $\tau > 0$ is introduced to provide stability, and to ensure that a zero-valued component $\tilde{\mathbf{v}}_j$ (or $\mathbf{v}_{jm}$) does not strictly prohibit a nonzero estimate at the next step. Similar heuristic updating rules were also adopted in [24].

It is observed that problem $\mathscr{P}^{DC}$ has a convex objective function, as well as DC constraints and convex constraints, and thus falls into the category of the general DC programming problems which take the following form:

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & f_0(\mathbf{x}) - h_0(\mathbf{x}) \\ \text{subject to} \quad & f_i(\mathbf{x}) - h_i(\mathbf{x}) \leq 0, i = 1, \ldots, m, \end{aligned} \tag{37}$$

where $f_i(\cdot)$ and $h_i(\cdot)$, for $i = 0, \ldots, m$, are convex functions. The concave-convex procedure (CCCP) [47] has been developed to reach a local minimum of DC programming problems with a guaranteed convergence, where $\mathbf{x}_t$ can be updated by solving the convex subproblem:

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & g_0(\mathbf{x}\,|\mathbf{x}_t) \\ \text{subject to} \quad & g_i(\mathbf{x}\,|\mathbf{x}_t) \leq 0, i = 1, \ldots, m, \end{aligned} \tag{38}$$

where

$$g_i(\mathbf{x}\,|\mathbf{x}_t) = f_i(\mathbf{x}) - \left[h_i(\mathbf{x}_t) + \nabla h_i(\mathbf{x}_t)^T (\mathbf{x} - \mathbf{x}_t)\right], \tag{39}$$

for all $i = 0, \ldots, m$. To be specific, for problem $\mathscr{P}^{DC}$, the subproblem in the t-th iteration of the CCCP takes the following form:

$$\underset{\mathbf{v}}{\text{minimize}} \quad \tilde{p}\left(\mathbf{v} \left| \tilde{\omega}_j^{[t]}, \omega_{jm}^{[t]} \right.\right) \tag{40}$$

$$\text{subject to} \quad \begin{aligned} &\gamma_m \left(\sum_{i \neq m}^{N_G} \left| \mathbf{h}_k^{\mathsf{H}} \mathbf{v}_i \right|^2 + \sigma_k^2 \right) - 2\mathrm{Re}\left\{ \left(\mathbf{v}_m^{[t]} \right)^{\mathsf{H}} \mathbf{h}_k \mathbf{h}_k^{\mathsf{H}} \mathbf{v}_m \right\} \\ &\quad + \left(\mathbf{v}_m^{[t]} \right)^{\mathsf{H}} \mathbf{h}_k \mathbf{h}_k^{\mathsf{H}} \mathbf{v}_m^{[t]} \leq 0, \forall k \in \mathcal{G}_m, \forall m \in \mathcal{M} \end{aligned} \tag{40a}$$

$$\sum_{m=1}^{N_G} R_m \left(1 - c_{q_m,j}\right) \omega_{jm}^{[t]} \left\| \mathbf{v}_{jm} \right\|_2^2 \leq C_j^{\mathrm{BH}}, \forall j \in \mathcal{J} \tag{40b}$$

$$\sum_{m=1}^{N_G} \left\| \mathbf{v}_{jm} \right\|_2^2 \leq P_j^{\mathrm{TX}}, \forall j \in \mathcal{J}, \tag{40c}$$

where the coefficients

$$\tilde{\omega}_j^{[t]} = \frac{1}{\left\| \tilde{\mathbf{v}}_j^{[t]} \right\|_2^2 + \tau}, \text{ and } \omega_{jm}^{[t]} = \frac{1}{\left\| \mathbf{v}_{jm}^{[t]} \right\|_2^2 + \tau} \tag{41}$$

are updated with the solution $\mathbf{v}^{[t]}$ obtained in the previous iteration. In the t-th iteration, we obtain the solution $\mathbf{v}^{[t+1]}$ by solving problem (40). Problem (40) is a convex quadratically constrained quadratic program (QCQP), which can be regarded as a special case of a second-order cone program (SOCP) and can be readily solved by interior-point methods with complexity as $\mathcal{O}\left(N_G^{3.5} N_B^{3.5} L^{3.5}\right)$ [48]. To solve our problem efficiently, we need to carefully choose an initial feasible point for the CCCP algorithm. Therefore, an initialization step is proposed by solving a transmit power minimization problem $\mathscr{P}_0$ with SDR technique [44], i.e.,

$$\mathscr{P}_0 : \underset{\{\mathbf{W}_m\}}{\text{minimize}} \quad \sum_{m=1}^{N_G} \mathrm{Tr}\left(\mathbf{W}_m\right) \tag{42}$$

$$\text{subject to} \quad \frac{\mathrm{Tr}\left(\mathbf{W}_m \mathbf{H}_k\right)}{\sum_{i=1, i \neq m}^{N_G} \mathrm{Tr}\left(\mathbf{W}_i \mathbf{H}_k\right) + \sigma_k^2} \geq \gamma_m, \forall k \in \mathcal{G}_m, \forall m \in \mathcal{M} \tag{42a}$$

$$\sum_{m=1}^{N_G} \mathrm{Tr}\left(\mathbf{W}_m \mathbf{J}_j\right) \leq P_j, \forall j \in \mathcal{J} \tag{42b}$$

$$\sum_{m=1}^{N_G} R_m \left(1 - c_{q_m,j}\right) \omega_{jm} \mathrm{Tr}\left(\mathbf{W}_m \mathbf{J}_j\right) \leq C_j^{\mathrm{BH}}, \forall j \in \mathcal{J} \tag{42c}$$

$$\mathbf{W}_m \succeq 0, \forall m \in \mathcal{M}, \tag{42d}$$

where we define two matrices $\mathbf{W}_m = \mathbf{v}_m \mathbf{v}_m^H \in \mathbb{C}^{N \times N}, \forall m \in \mathcal{M}$ and $\mathbf{H}_k = \mathbf{h}_k \mathbf{h}_k^H \in \mathbb{C}^{N \times N}, \forall k \in \mathcal{K}$, to lift the quadratic constraints into higher dimensions. Moreover, we define a set of selective

matrices $\mathbf{J}_j \in \{0,1\}^{N\times N}, \forall j \in \mathcal{J}$, with $\mathbf{J}_j = \mathrm{diag}\left(\mathbf{0}_{(j-1)L}, \mathbf{1}_L, \mathbf{0}_{(N_B-j)L}\right)$ as a diagonal matrix. All $\{\mathbf{W}_m\}$ are rank-one constrained. If the solution $\{\mathbf{W}_m\}$ are all rank-one, the feasible beamformers $\{\mathbf{v}_m\}$ obtained by applying the eigenvalue decomposition (EVD) on $\{\mathbf{W}_m\}$ can be directly employed as the initial feasible point for the CCCP algorithm. If the solutions $\{\mathbf{W}_m\}$ are not rank-one, $\{\mathbf{v}_m\}$ are obtained through randomizing and scaling. If problem $\mathscr{P}_0$ is infeasible, the original problem $\mathscr{P}$ is infeasible and the optimization has to terminate.

After solving problem $\mathscr{P}^{DC}$, we will obtain the sparse beamforming vector $\hat{\mathbf{v}}$ as the output of the first stage, as shown in Fig. 2. The algorithm solving the group sparsity-inducing norm minimization problem for the first stage is presented as Algorithm 1, which will converge to local minima or saddle points of problem $\mathscr{P}^{DC}$ [47], [49], if it is feasible.

Algorithm 1 The Group Sparsity-Inducing Norm Minimization Algorithm

Step 1: Find an initial feasible point $\left\{\mathbf{v}^{[0]}\right\}$ by solving problem $\mathscr{P}_0$;

Step 2: Initialize$\left\{\omega_{jm}^{[0]}\right\}$, $\left\{\tilde{\omega}_{j}^{[0]}\right\}$, and set the iteration counter as $t = 0$;

Step 3: Repeat

1) Solve problem (40), and obtain the beamformer $\mathbf{v}^{[t+1]}$;
2) Set $t = t + 1$, and update the weights according to (41);

Step 4: Until stopping criterion is met and obtain the beamformer $\hat{\mathbf{v}}$;

End

C. Stage II: Iterative Search Procedure

Inducing the sparsity structure in the solution is critical to problem $\mathscr{P}^{LGSBF}$. As illustrated in Fig. 3, the layered group sparse pattern can be mapped to a hierarchy tree. Generally, the solution $\hat{\mathbf{v}}$ obtained from Stage I is not strictly sparse, and thus we propose to trim the entries of $\hat{\mathbf{v}}$ to obtain the group sparse solution. Although backhaul capacity constraints can help us filter some sparsity patterns (i.e, prune some nodes in the hierarchy tree), finding the optimal sparsity pattern in $\mathbf{v}$ brings about high computational complexity. In this subsection, we develop an efficient search procedure to identify active BSs and backhaul data assignment.

1) BS Ordering and Selection: With the knowledge of the input $\hat{\mathbf{v}}$, the next step is to determine the active BS set. After giving proper priorities to BSs, we can obtain an ordering list to switch them off. Previous works have considered different ways to calculate the priorities. For example,

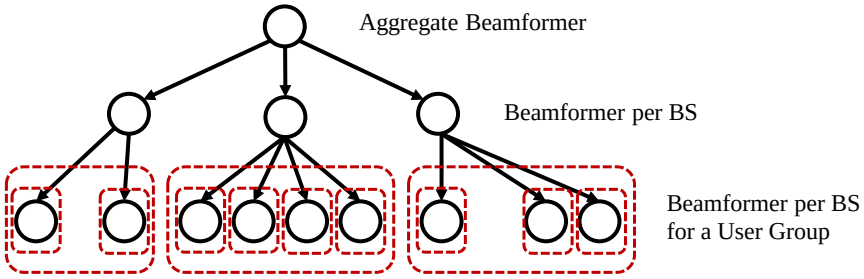


Fig. 3. The layered group sparse pattern in the aggregate beamformer.

Mehanna *et al.* [50] directly mapped the group-sparsity obtained by the group-sparsity inducing norm minimization to their application, i.e., the transmit antennas with smaller coefficients in the group were determined to be turned off with a higher priority. Following this idea, in our setting, the priorities might be given as $\tilde{\theta}_j = \|\tilde{\mathbf{v}}_j\|_2, \forall j$, which implies that the BS with a lower transmit beamforming gain should be encouraged to be switched off. However, such a direct mapping might bring performance degradation, as shown in [33]. To get a better performance, it is essential to consider not only the transmit beamforming gain but also other key system parameters indicating the impact of the BSs on the network performance. Similar to the one employed in [33], to assign priorities to BSs, we propose the following ordering criteria that incorporates channel power gain, BS power amplifier efficiency, relative power consumption, backhaul power consumption, caching status and beamforming gain, that is,

$$\tilde{\theta}_j = \sqrt{\frac{\tilde{\kappa_j}}{\delta_j \left(P_j^D + \sum_{m=1}^{N_G} \beta_{jm} \left(1 - c_{q_m,j}\right)\right)}} \|\tilde{\mathbf{v}}_j\|_2, \forall j, \tag{43}$$

where $\tilde{\kappa_j} = \sum_{k=1}^{N_U} \|h_{kj}\|^2$ is the channel gain from the j-th BS to all MUs. The BS with a higher priority (i.e., smaller $\tilde{\theta}_j$) will be switched off before the one with a lower priority (i.e., larger $\tilde{\theta}_j$). This ordering criteria implies that the BS with a lower channel power gain, lower BS power amplifier efficiency (i.e., higher δ_j), higher relative transport link power consumption, higher backhaul power consumption and lower cache hit ratio should have a higher priority to be switched off.

Once BS j is decided to be switched off, all its corresponding beamforming coefficients will be set to zero, i.e, $\tilde{\mathbf{v}}_j = \mathbf{0}$. Based on the ordering criteria rule, we sort the coefficients in ascending order. Each time a BS is decided to be switched off, the inactive BS set $\mathcal{Z}_{BS}$ will be updated

and we check a feasibility problem

$$\mathscr{F}_1\left(\mathcal{Z}_{BS}\right): \text{find} \quad \mathbf{v} \tag{44}$$

$$\text{subject to} \quad \sum_{m=1}^{N_G} \|\mathbf{v}_{jm}\|_2^2 \leq P_j^{\text{TX}}, \forall j \notin \mathcal{Z}_{BS} \tag{44a}$$

$$\tilde{\mathbf{v}}_j = \mathbf{0}, \text{if } j \in \mathcal{Z}_{BS} \tag{44b}$$

$$(32), (35\text{a}),$$

which is a DC programming, and can be solved by the CCCP algorithm.

2) Backhaul Data Assignment for Active BSs: If problem $\mathscr{F}_1\left(\mathcal{Z}_{BS}\right)$ is feasible, the next question is to determine the backhaul data assignment for the active BSs in order to further reduce the power consumption. Similar to the design idea for (43), we calculate priorities of backhaul data assignment for the active BSs by taking the aforementioned key system parameters into consideration. Consequently, we propose the following ordering criteria to determine which backhaul data assignment should be turned off, i.e.,

$$\theta_{jm} = \begin{cases} \sqrt{\frac{\kappa_{jm}}{\delta_j\left(P_j^D + \beta_{jm}(1-c_{q_m,j})\right)}}\, \|\mathbf{v}_{jm}\|_2, & \text{if } j \notin \mathcal{Z}_{BS} \\ 0, & \text{if } j \in \mathcal{Z}_{BS} \end{cases}, \tag{45}$$

where $\kappa_{jm} = \sum_{k \in \mathcal{G}_m} \|h_{kj}\|^2$ is the channel gain from the j-th BS to the MUs in the m-th user group. Based on the ordering criteria, we delete a piece of data assignment each time and update the inactive data assignment set $\mathcal{Z}_{DA}$. With $\mathcal{Z}_{BS}$ and $\mathcal{Z}_{DA}$, the subproblem that we need to solve takes the following form:

$$\mathscr{F}_2\left(\mathcal{Z}_{BS}, \mathcal{Z}_{DA}\right): \text{find} \quad \mathbf{v} \tag{46}$$

$$\text{subject to} \quad \mathbf{v}_{jm} = \mathbf{0}, \forall\, (j,m) \in \mathcal{Z}_{DA}$$

$$(32), (35\text{a}), (44\text{a}), (44\text{b}),$$

which is also a DC program, and can be solved by the CCCP algorithm.

Realizing that switching off as many BSs as possible may not result in a minimum total network power consumption, we are motivated to adopt a conservative strategy to determine the final active BS set and backhaul data assignment. To obtain the minimum network power consumption, we iteratively search over all possible $\mathcal{Z}_{BS}$ and $\mathcal{Z}_{DA}$, and record the corresponding network power. By comparing all the recorded values, we can determine $(\mathcal{Z}_{BS}^{\star}, \mathcal{Z}_{DA}^{\star})$ that

corresponds to the minimal network power consumption. Overall, the iterative search method can be accomplished via solving no more than $\frac{N_G N_B(N_B+1)}{2}$ DC problems.

D. Stage III: Obtain Transmit Beamformers

With the obtained inactive BS set $\mathcal{Z}_{BS}^{\star}$ and inactive data assignment set $\mathcal{Z}_{DA}^{\star}$, we can obtain the final beamforming vector by solving the following problem:

$$\begin{aligned} \mathscr{P}^{Final}\left(\mathcal{Z}_{BS}^{\star}, \mathcal{Z}_{DA}^{\star}\right): \underset{\mathbf{v}}{\text{minimize}} \quad & \sum_{j=1}^{N_B}\sum_{m=1}^{N_G} \delta_j \left\|\mathbf{v}_{jm}\right\|_2^2 \\ \text{subject to} \quad & \sum_{m=1}^{N_G} \left\|\mathbf{v}_{jm}\right\|_2^2 \leq P_j^{\mathrm{TX}}, \forall j \notin \mathcal{Z}_{BS}^{\star} \\ & \tilde{\mathbf{v}}_j = \mathbf{0}, \forall j \in \mathcal{Z}_{BS}^{\star} \\ & \mathbf{v}_{jm} = \mathbf{0}, \forall \left(j,m\right) \in \mathcal{Z}_{DA}^{\star} \\ & \left(32\right), \left(35\mathrm{a}\right), \end{aligned} \tag{47}$$

which is also a DC program. In principle, problem (47) can be globally solved via the branch-and bound algorithm by extending the method developed in [51]. Such global optimization algorithms have high computational complexity, and cannot be applied in dense networks. Therefore, the CCCP algorithm is adopted to efficiently obtain a local optimal solution. The overall iterative LGSBF algorithm is summarized in Algorithm 2. Note that the proposed approach provides a general framework for a multi-layer GSBF problem, where various group sparsity-inducing algorithms, e.g., the smoothed ℓ_p-minimization [34], can be applied in Stage I.

E. Complexity and Convergence Analysis

It has been shown that for the iterative search procedure, the number of general DC problems to be solved is no more than $\frac{N_G N_B(N_B+1)}{2}$. To obtain a local optimal solution for general DC programs, at each iteration of the CCCP-based algorithm, we need to solve a convex QCQP (or equivalently SOCP) problem with a complexity of $\mathcal{O}\left(N_G^{3.5} N_B^{3.5} L^{3.5}\right)$ by interior-point methods, which constitutes the main computational complexity of the proposed LGSBF algorithm. For large-scale networks, other approaches for solving large-sized SOCPs, e.g., the alternating direction method of multipliers (ADMM) method [52], need to be explored. For unconstrained DC programs with differentiable objectives, it could converge superlinearly [49], while the convergence rate of general DC programs is still an open problem.

Algorithm 2 The Iterative LGSBF Algorithm

Step 1: Solve problem $\mathscr{P}^{LGSBF}$ by applying Algorithm 1: **if** it is infeasible, **go to End**; otherwise, obtain $\hat{\mathbf{v}}$;

Step 2: Calculate the ordering criterion (43), and sort the values in the ascending order $\tilde{\theta}_{\pi_1} \leq \cdots \leq \tilde{\theta}_{\pi_{N_B}}$;

Step 3: Initialize $\mathcal{Z}_{BS}^{[0]} = \emptyset$, and $i = 0$;

Step 4: Solve the optimization problem $\mathscr{F}_1\left(\mathcal{Z}_{BS}^{[i]}\right)$

1) **If** $\mathscr{F}_1\left(\mathcal{Z}_{BS}^{[i]}\right)$ is feasible,
   a) Calculate the ordering criterion (45), and sort the values in the ascending order $\tilde{\theta}_{\varpi_1} \leq \cdots \leq \tilde{\theta}_{\varpi_{N_B N_G}}$;
   b) Initialize $\mathcal{Z}_{DA}^{[0]} = \emptyset$, and $k = 0$;
   c) **Repeat** Solve the optimization problem $\mathscr{F}_2\left(\mathcal{Z}_{BS}^{[i]}, \mathcal{Z}_{DA}^{[k]}\right)$, update the set $\mathcal{Z}_{DA}^{[k+1]} = \mathcal{Z}_{DA}^{[k]} \cup \{\varpi_{k+1}\}$ and $k = k + 1$;
   d) **Until** infeasible, obtain $\mathcal{S}_{\mathcal{K}}{}^{[i]} = \{0, 1, \ldots, k-1\}$;
   e) Update the set $\mathcal{Z}_{BS}^{[i+1]} = \mathcal{Z}_{BS}^{[i]} \cup \{\pi_{i+1}\}$ and $i = i + 1$, **go to Step 4**;
2) **If** $\mathscr{F}_1\left(\mathcal{Z}_{BS}^{[i]}\right)$ is infeasible, obtain $\mathcal{S}_{\mathcal{I}} = \{0, 1, \ldots, i-1\}$, **go to Step 5**;

Step 5: Obtain the optimal inactive BS set $\mathcal{Z}_{BS}^{\star}$ and inactive data assignment set $\mathcal{Z}_{DA}^{\star}$ by solving $(\mathcal{Z}_{BS}^{\star}, \mathcal{Z}_{DA}^{\star}) = \underset{i \in \mathcal{S}_{\mathcal{I}}, k \in \mathcal{S}_{\mathcal{K}}{}^{[i]}}{\arg\min}\ p^{\star}\left(\mathcal{Z}_{BS}^{[i]}, \mathcal{Z}_{DA}^{[k]}\right)$;

Step 6: Obtain beamformers by solving problem $\mathscr{P}^{Final}\left(\mathcal{Z}_{BS}^{\star}, \mathcal{Z}_{DA}^{\star}\right)$;

End

V. Simulation results

In this section, we simulate the performance of the proposed algorithm. We consider a hexagonal multicell network, where each BS is located at the center of a hexagonal cell whose radius is set to be 500 m, and MUs are uniformly and independently distributed in the network, excluding an inner circle of 50 m around each BS. The channel between the j-th BS and the k-th user is modeled as $\mathbf{h}_{kj} = 10^{-L\left(d_{kj}\right)/20}\sqrt{\varphi_{kj} s_{kj}}\mathbf{g}_{kj}$, where $L\left(d_{kj}\right)$ is the path-loss at distance d_{kj}, s_{kj} is the shadowing coefficient, φ_{kj} is transmit antenna power gain and $\mathbf{g}_{kj}$ is the small scale fading coefficient. We adopt the standard cellular network parameters as presented in Table II.

The file library contains 100 pieces of content, whose popularity follows a Zipf distribution with parameter $\gamma_z = 1.2$. In this popularity model, a small γ_z implies a flat popularity distribution,

TABLE II

SIMULATION PARAMETERS

Parameter	Value
Transmit antenna power gain φ_{kj}	$10\,\mathrm{dBi}$
Path-loss at distance d_{kj} (km)	$148.1 + 37.6 \log_{10}(d_{kj})$
Standard deviation of log-norm shadowing σ_s	$8\,\mathrm{dB}$
Small scale fading distribution $\mathbf{g}_{kj}$	$\mathcal{CN}(0, \mathbf{I})$
Noise power spectral density σ_k^2	$-172\ \mathrm{dBm/\,Hz}$
Bandwidth B_0	$10\,\mathrm{MHz}$
Maximum BS transmit power P_j^{TX}	$1\,\mathrm{W}$
Slope of the load-dependent power consumption δ_j	4

while a large γ_z means the opposite. BSs are assumed to have equal cache sizes. We shall briefly show the role of cache by varying the cache size and caching strategies via simulations. Herein, we consider two widely-employed heuristic caching strategies, i.e., the most popular caching (MPC) [53] and probabilistic caching (ProbC) [19], [54]. For MPC, each BS caches as many popular files as possible in accordance with the file popularity rank in the descending order. As for ProbC, each BS randomly caches files with the same probabilities as their request probabilities. Assume that the SINR requirements for different user groups are the same, i.e., $\gamma_m = \gamma, \forall m \in \{1, \ldots, N_G\}$.

A. Network Power Consumption

Consider a network with $N_B = 7$ BSs, each of which has two antennas, and $N_U = 15$ single-antenna MUs. We set the relative power consumption as $P_j^D = [5.6 + j - 1]\,\mathrm{W}, \forall j \in \mathcal{J}$, backhaul energy coefficient as $E_j^{\mathrm{BH}} = 1 \times 10^{-7}\,\mathrm{J/\,bit}$, and per-BS backhaul capacity as $500\,\mathrm{Mbps}$. Each BS has a cache size of 10 files [54], i.e., $c_{fj} = 1, \forall f = 1, \ldots, 10, \forall j \in \mathcal{J}$.

The proposed algorithm is compared with the following algorithms:

- Coordinated beamforming (CB) algorithm: In this algorithm [42], all BSs are in the active mode and only the total BS transmit power consumption is minimized.
- Sparse multicast beamforming algorithm with adaptive BS selection: This algorithm [34] develops a procedure to switch off as many BSs as possible. The non-convex smoothed ℓ_p-norm is adopted to replace the convex mixed $\ell_1/\,\ell_2$-norm in the objective function. The non-convex quadratic forms of beamforming vectors in the objective function and the

non-convex quadratic QoS constraints are relaxed by leveraging SDR technique. Then an iterative reweighted-ℓ_2 algorithm is employed to solve the problem.

- Sparse multicast beamforming algorithm with adaptive backhaul content assignment: In this algorithm [19], the number of backhaul content assignments is minimized. Smooth approximated functions are employed to approximate the ℓ_0-norm terms and a generalized CCCP algorithm is applied to solve the problem. The arctangent function is adopted since [19] shows it gives the best approximation performance.

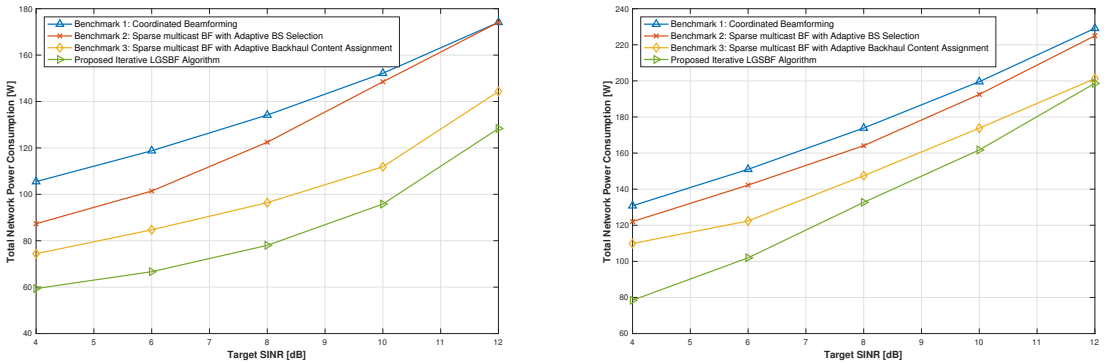
Fig. 4 demonstrates the total network power consumption with different target SINR values under MPC and ProbC, respectively. It shows that the proposed iterative LGSBF algorithm outperforms existing algorithms with both caching strategies, which confirms the effectiveness of the proposed algorithm. When the target SINR increases, it is observed that the gap between different algorithms becomes smaller, while benchmark 2 converges to CB faster than benchmark 3 and the proposed algorithm. It is because that more and more BSs need to be switched on to support the increasing QoS requirements, which decreases the benefit of active BS selection. Whereas, a careful design for backhaul content assignment can still help reduce network power consumption, since it can bring some cooperation chance for BSs and avoid unnecessary backhaul consumption at the same time.

Remark 1. The proposed algorithm achieves a better performance than those of [34] and [19] by considering a two-layer adaptive selection for both active BS and actual backhaul assignment instead of the existing one-layer approaches. This indicates that the joint adaptive decision of BS selection and backhaul content assignment can effectively reduce network power consumption for a wide range of target SINRs.

Remark 2. Comparing two caching strategies, we observe that MPC performs better than ProbC in reducing network power consumption, and the gap becomes larger when the target SINR increases. In general, MPC provides better performance than ProbC for normal network settings, and similar findings are also observed in [19].

B. Impact of Cache Size

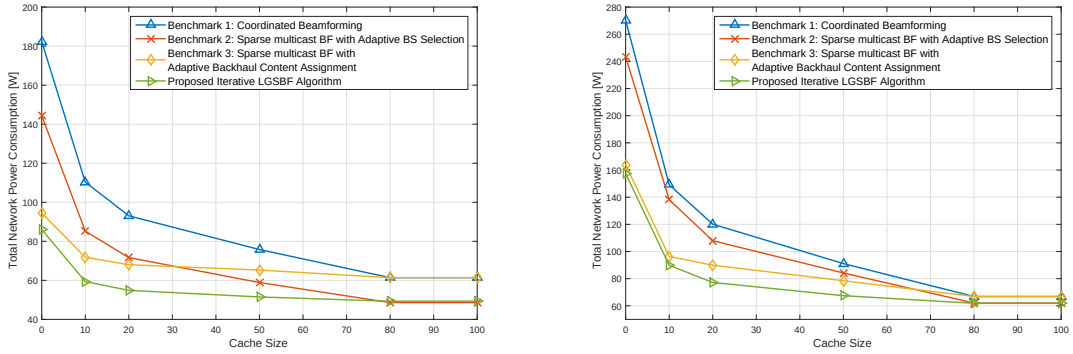
In Fig. 5(a) and Fig. 5(b), we compare the performance of the proposed algorithm with benchmarks in terms of the tradeoff between total network power consumption and per-BS cache size under target SINR $= 5\,\mathrm{dB}$ and target SINR $= 10\,\mathrm{dB}$, respectively. Other settings are the same as those in Fig. 4. From Fig. 5, it is observed that the proposed algorithm outperforms



(a) With MPC strategy.

(b) With ProbC strategy.

Fig. 4. Network power consumption versus target SINR.



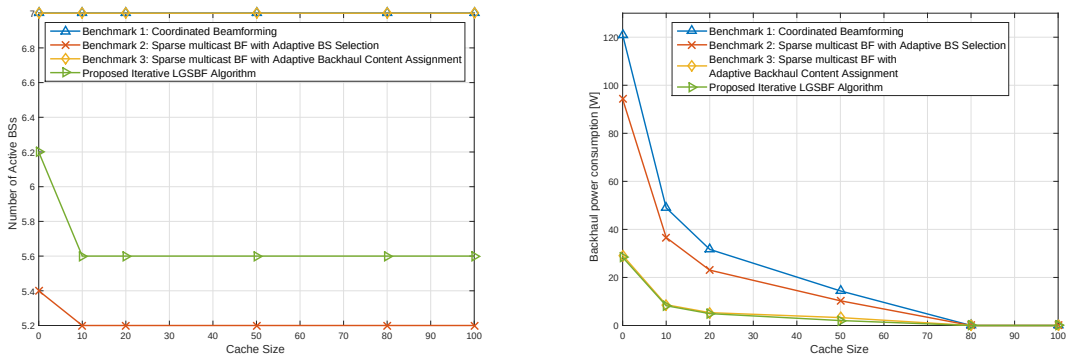
(a) Target SINR = 5 dB.

(b) Target SINR = 10 dB.

Fig. 5. The tradeoff between the total network power consumption and the cache size of each BS.

benchmarks under different cache sizes. Moreover, the advantage of adaptive backhaul content assignment will gradually be surpassed by adaptive BS selection when the cache size increases; at a higher target SINR regime, the crosspoint will occur at a larger cache size. Besides this, the proposed algorithm achieves a better performance in a low SINR regime. It can be inferred that the increase in the cache size allows more BSs to be switched off, especially when the QoS requirement is comparatively low.

With the same network setting as in Fig. 5(a), the details of the impact of caching on the BSs and backhaul links are demonstrated in Fig. 6. This figure shows that the CB algorithm, which intends to minimize the BS transmit power consumption, has the highest backhaul power consumption. This is because all the BSs are active in the CB algorithm in order to achieve



(a) The number of active BSs versus cache size.

(b) Backhaul power consumption versus cache size.

Fig. 6. The impact of the cache size.

the highest beamforming gain. Moreover, by comparing Benchmark 2, Benchmark 3 and the proposed algorithm, it can be inferred that minimizing either the number of active BSs or the number of backhaul content delivery cannot be the optimal strategy to save power. Since both the backhaul power consumption and BS power consumption hold a nontrivial share, a joint adaptive BS selection, backhaul data assignment and power minimization beamforming is crucial for minimizing the total network power consumption.

C. Impact of the Number of Mobile Users and Backhaul Energy Coefficient

We also investigate the impact of other important network parameters, i.e., the number of MUs and backhaul energy coefficient, as shown in Fig. 7. The figure demonstrates that when the number of MUs increases, the performance gap between the zero-cache case and full-cache case becomes larger. On the other hand, Fig. 7 also shows that the performance gap between the zero-cache case and full-cache case is larger for the network with a higher backhaul energy coefficient. Actually, different backhaul energy coefficients represent different types of backhaul links: a higher backhaul energy coefficient stands for less power-efficient backhaul links, and vice versa. To enhance the performance in total network power consumption, the operators can either upgrade the backhaul links, which is expensive, or simply install cost-effective caches. From the simulation, we can infer that caches will play a more significant part in networks with higher user densities, and less power-efficient backhaul links.

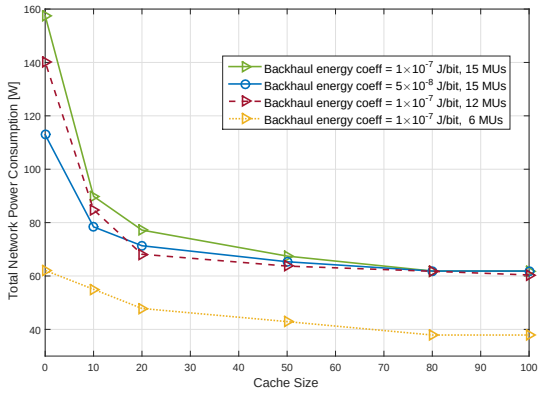


Fig. 7. Network power consumption versus cache size under different MU densities.

VI. Conclusions

In this study, we developed an effective framework to minimize the total network power consumption of cache-enabled wireless networks. The proposed LGSBF formulation generalized existing works on group sparse beamforming, for which an effective algorithm was developed. The proposed algorithm can significantly reduce the total network power consumption via a joint design of adaptive BS selection, backhaul content assignment and multicast beamforming. From the simulations, the proposed LGSBF framework was demonstrated to outperform existing algorithms by striking a balance between the BS power consumption and backhaul power consumption. Furthermore, it was shown that caching tends to play a more significant part in networks with higher user densities and less power-efficient backhaul links. For future research directions, it would be interesting to optimize the caching placement in the prefetching phase, and incorporate it when minimizing the total network power consumption. It is also important but challenging to develop more efficient distributed algorithms for practical implementation in large-scale networks.

References

[1] Cisco Systems Inc., "Cisco visual networking index: Global mobile data traffic forecast update, 2016-2021," White Paper, Feb. 2017.

[2] N. Bhushan, J. Li, D. Malladi, R. Gilmore, D. Brenner, A. Damnjanovic, R. T. Sukhavasi, C. Patel, and S. Geirhofer, "Network densification: The dominant theme for wireless evolution into 5G," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 82–89, Feb. 2014.

[3] D. Lee, H. Seo, B. Clerckx, E. Hardouin, D. Mazzarese, S. Nagata, and K. Sayana, "Coordinated multipoint transmission and reception in LTE-advanced: Deployment scenarios and operational challenges," *IEEE Commun. Mag.*, vol. 50, no. 2, pp. 148–155, Feb. 2012.

[4] D. Schien, P. Shabajee, M. Yearworth, and C. Preist, "Modeling and assessing variability in energy consumption during the use stage of online multimedia services," *J. Ind. Ecol.*, vol. 17, no. 6, pp. 800–813, Dec. 2013.

[5] X. Ge, H. Cheng, M. Guizani, and T. Han, "5G wireless backhaul networks: Challenges and research advances," *IEEE Netw.*, vol. 28, no. 6, pp. 6–11, Nov. 2014.

[6] K. Shanmugam, N. Golrezaei, A. Dimakis, A. 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.

[7] X. Wang, A. V. Vasilakos, M. Chen, Y. Liu, and T. T. Kwon, "A survey of green mobile networks: Opportunities and challenges," *Mobile Netw. Appl.*, vol. 17, no. 1, pp. 4–20, 2012.

[8] F. Richter, A. J. Fehske, P. Marsch, and G. P. Fettweis, "Traffic demand and energy efficiency in heterogeneous cellular mobile radio networks," in *Proc. IEEE Vehicular Technology Conference (VTC)*, Taipei, Taiwan, May 2010, pp. 1–6.

[9] 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.

[10] S. Tombaz, P. Monti, K. Wang, A. Vastberg, M. Forzati, and J. Zander, "Impact of backhauling power consumption on the deployment of heterogeneous mobile networks," in *IEEE Global Commun. Conf. (GLOBECOM)*, Houston, TX, Dec. 2011, pp. 1–5.

[11] N. Choi, K. Guan, D. C. Kilper, and G. Atkinson, "In-network caching effect on optimal energy consumption in content-centric networking," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Ottawa, Canada, Jun. 2012, pp. 2889–2894.

[12] N. Golrezaei, K. Shanmugam, A. G. Dimakis, A. F. Molisch, and G. Caire, "Femtocaching: Wireless video content delivery through distributed caching helpers," in *Proc. IEEE Int. Conf. Computer Commun. (INFOCOM)*, Orlando, FL, Mar. 2012, pp. 1107–1115.

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

[14] ——, "Cache-aided interference channels," *arXiv preprint*, Jun. 2017. [Online]. Available: http://arxiv.org/abs/1510.06121v2

[15] J. Li, Y. Chen, Z. Lin, W. Chen, B. Vucetic, and L. Hanzo, "Distributed caching for data dissemination in the downlink of heterogeneous networks," *IEEE Trans. Commun.*, vol. 63, no. 10, pp. 3553–3568, Oct. 2015.

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

[17] E. Baştug, M. Bennis, M. Kountouris, and M. Debbah, "Cache-enabled small cell networks: Modeling and tradeoffs," *EURASIP J. Wireless Commun. Netw.*, vol. 2015, no. 1, pp. 41–51, Feb. 2015.

[18] 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.

[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, Sept. 2016.

[20] A. Sengupta, R. Tandon, and O. Simeone, "Fog-aided wireless networks for content delivery: Fundamental latency trade-offs," *arXiv preprint*, Jun. 2017. [Online]. Available: http://arxiv.org/abs/1605.01690v5

[21] J. Zhang, R. Chen, J. G. Andrews, A. Ghosh, and R. W. Heath, "Networked MIMO with clustered linear precoding," *IEEE Trans. Wireless Commun.*, vol. 8, no. 4, pp. 1910–1921, Apr. 2009.

[22] C. T. K. Ng and H. Huang, "Linear precoding in cooperative MIMO cellular networks with limited coordination clusters," *IEEE J. Select. Areas Commun.*, vol. 28, no. 9, pp. 1446–1454, Dec. 2010.

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

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

[25] G. Auer, V. Giannini, C. Desset, I. Godor, P. Skillermark, M. Olsson, M. Imran, D. Sabella, M. Gonzalez, O. Blume, and A. Fehske, "How much energy is needed to run a wireless network?" *IEEE Wireless Commun.*, vol. 18, no. 5, pp. 40–49, Oct. 2011.

[26] P. Frenger, P. Moberg, J. Malmodin, Y. Jading, and I. Godor, "Reducing energy consumption in lte with cell DTX," in *IEEE Vehicular Technology Conference (VTC)*, May 2011, pp. 1–5.

[27] J. Wu, S. Zhou, and Z. Niu, "Traffic-aware base station sleeping control and power matching for energy-delay tradeoffs in green cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 8, pp. 4196–4209, Aug. 2013.

[28] H. Yao, C. Fang, C. Qiu, C. Zhao, and Y. Liu, "A novel energy efficiency algorithm in green mobile networks with cache," *EURASIP J. Wireless Commun. Netw.*, vol. 2015, no. 1, pp. 139–147, May 2015.

[29] J. Ghimire and C. Rosenberg, "Impact of limited backhaul capacity on user scheduling in heterogeneous networks," in *Proc. IEEE Wireless Commun. Networking Conf. (WCNC)*, Apr. 2014, pp. 2480–2485.

[30] I. Atzeni, M. Maso, I. . Ghamnia, M. Debbah, and E. Başţug, "Flexible cache-aided networks with backhauling," in *Proc. IEEE Int. Workshop Signal Process. Advances in Wireless Commun. (SPAWC)*, Sapporo, Japan, Jul. 2017.

[31] K. Poularakis, G. Iosifidis, V. Sourlas, and L. Tassiulas, "Exploiting caching and multicast for 5G wireless networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 4, pp. 2995–3007, Apr. 2016.

[32] X. Peng, J.-C. Shen, J. Zhang, and K. B. Letaief, "Joint data assignment and beamforming for backhaul limited caching networks," in *Proc. IEEE Int. Symp. Personal Indoor and Mobile Radio Comm. (PIMRC)*, Washington, DC, Sept. 2014, pp. 1370–1374.

[33] 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.

[34] Y. Shi, J. Cheng, J. Zhang, B. Bai, W. Chen, and K. B. Letaief, "Smoothed L_p-minimization for green Cloud-RAN with user admission control," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 1022–1036, Apr. 2016.

[35] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[36] E. J. Candes and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?" *IEEE Trans. Inf. Theory*, vol. 52, no. 12, pp. 5406–5425, Dec. 2006.

[37] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. R. Statist. Soc. B*, vol. 68, pp. 49–67, 2006.

[38] L. Jacob, G. Obozinski, and J.-P. Vert, "Group lasso with overlap and graph lasso," in *Proc. Int. Conf. Machine Learning (ICML)*, 2009, pp. 433–440.

[39] J. Wang and J. Ye, "Two-layer feature reduction for sparse-group lasso via decomposition of convex sets," *Neural Inf. Process. Syst.*, pp. 2132–2140, 2014.

[40] S. Gao, L. T. Chia, and I. W. H. Tsang, "Multi-layer group sparse coding-for concurrent image classification and annotation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, Jun. 2011, pp. 2809–2816.

[41] Y.-F. Liu, Y.-H. Dai, and S. Ma, "Joint power and admission control: Non-convex Lq approximation and an effective polynomial time deflation approach," *IEEE Trans. Signal Process.*, vol. 63, no. 14, pp. 3641–3656, Jul. 2015.

[42] H. Dahrouj and W. Yu, "Coordinated beamforming for the multicell multi-antenna wireless system," *IEEE Trans. Wireless Commun.*, vol. 9, no. 5, pp. 1748–1759, May 2010.

[43] A. Fehske, P. Marsch, and G. Fettweis, "Bit per joule efficiency of cooperating base stations in cellular networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM) Workshop*, Miami, FL, Dec. 2010, pp. 1406–1411.

[44] N. Sidiropoulos, T. Davidson, and Z.-Q. Luo, "Transmit beamforming for physical-layer multicasting," *IEEE Trans. Signal Process.*, vol. 54, no. 6, pp. 2239–2251, Jun. 2006.

[45] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski, "Optimization with sparsity-inducing penalties," *Found. Trends Mach. Learn.*, vol. 4, no. 1, pp. 1–106, Jan. 2012.

[46] E. J. Candes, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted ℓ_1 minimization," *J. Fourier Anal. Appl.*, vol. 14, no. 5, pp. 877–905, Dec. 2008.

[47] A. Yuille and A. Rangarajan, "The concave-convex procedure," *Neural Computation*, vol. 15, no. 4, pp. 915–936, Apr. 2003.

[48] S. P. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[49] G. R. Lanckriet and B. K. Sriperumbudur, "On the convergence of the concave-convex procedure," *Neural Inf. Process. Syst.*, pp. 1759–1767, 2009.

[50] O. Mehanna, N. D. Sidiropoulos, and G. B. Giannakis, "Joint multicast beamforming and antenna selection," *IEEE Trans. Signal Process.*, vol. 61, no. 10, pp. 2660–2674, May 2013.

[51] C. Lu and Y.-F. Liu, "An efficient global algorithm for single-group multicast beamforming," *IEEE Trans. Signal Process.*, vol. 65, no. 14, pp. 3761–3774, Jul. 2017.

[52] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations Trends Mach. Learning*, vol. 3, pp. 1–122, Jul. 2011.

[53] H. Ahlehagh and S. Dey, "Video-aware scheduling and caching in the radio access network," *IEEE/ACM Trans. Netw.*, vol. 22, no. 5, pp. 1444–1462, Oct. 2014.

[54] C. Bernardini, T. Silverston, and O. Festor, "A comparison of caching strategies for content centric networking," in *Proc. Global Commun. Conf. (GLOBECOM)*, San Diego, CA, Dec. 2015, pp. 1–6.