Joint Data Assignment and Beamforming for Backhaul Limited Caching Networks

Xi Peng, Juei-Chin Shen, Jun Zhang and Khaled B. Letaief, *Fellow*, IEEE
Dept. of ECE, The Hong Kong University of Science and Technology
E-mail: {xpengab, eejcshen, eejzhang, eekhaled}@ust.hk

***Abstract*—Caching at wireless access points is a promising approach to alleviate the backhaul burden in wireless networks. In this paper, we consider a cooperative wireless caching network where all the base stations (BSs) are connected to a central controller via backhaul links. In such a network, users can get the required data locally if they are cached at the BSs. Otherwise, the user data need to be assigned from the central controller to BSs via backhaul. In order to reduce the network cost, i.e., the backhaul cost and the transmit power cost, the data assignment for different BSs and the coordinated beamforming to serve different users need to be jointly designed. We formulate such a design problem as the minimization of the network cost, subject to the quality of service (QoS) constraint of each user and the transmit power constraint of each BS. This problem involves mixed-integer programming and is highly complicated. In order to provide an efficient solution, the connection between the data assignment and the sparsity-introducing norm is established. Low-complexity algorithms are then proposed to solve the joint optimization problem, which essentially decouple the data assignment and the transmit power minimization beamforming. Simulation results show that the proposed algorithms can effectively minimize the network cost and provide near optimal performance.**

***Index Terms*—Caching networks, data assignment, backhaul cost, power cost, sparsity-introducing norm**

I. Introduction

Recent years have witnessed an exponential growth of mobile data traffic, especially mobile video streaming. The increasing video traffic raises new challenges for mobile operators due to its highly demanding requirements in terms of high data rate, low latency and moderate delay jitter [1]. In particular, it has a strong demand on high-capacity backhaul links, which cannot be satisfactorily met in current networks. Recently, introducing caching to wireless access points has been proposed as a favorable approach to reduce the deployment cost of wireless networks, as it can help reduce the capacity requirement of backhaul links [2], [3]. Thus it will increase the network scalability to counteract the fierce increase in multimedia traffic.

In downlink multicell networks without caching, base station (BS) cooperation is a powerful technique to increase the network capacity, enabled by information exchange among BSs with the support of backhaul links. There are two main approaches for the downlink cooperation: Joint processing (JP) and coordinated beamforming (CB) [4]. JP has a superior performance to CB at the price of higher backhaul overhead caused by sharing user data among all the BSs via backhaul links. To reduce the backhaul cost, some recent proposals considered partial coordinated transmission, where the BSs can form (possibly overlapping) coordinated clusters of different sizes to perform JP [5]–[7]. With caching at BSs, we can further reduce the backhaul cost. Consider the following example. When all the payloads of users are cached at each BS, BSs can cooperatively perform full JP to provide reliable and high-data-rate transmission to multiple users without additional backhaul cost. In practice, due to the limited size of the local cache, only part of the user data can be stored at each BS. In this case, partial coordinated transmission can be appealed, where the backhaul acts as a supplement to deliver additional user data to BSs.

To increase the possibility for a user to access its interested video file at the local cache, caching contents are usually designed to follow some video popularity distribution, e.g., the Zipf distribution [8]. However, files may be cached at the BSs whose channels to the served users are poor. In this case, even though BSs who share the requested data are clustered together to deliver service, data transmission may not be reliable or high transmit power will be needed. Hence, it is necessary to allocate the requested files via backhaul to BSs who have good channels to the served users. Such allocation via backhaul will raise the backhaul cost. The *network cost* mainly consists of the above-mentioned two parts of cost, i.e., the transmit power cost and the backhaul cost. How to strike a balance between achieving good QoS and minimizing the network cost is a major issue in multicell caching networks.

In this paper, we will investigate the joint data assignment and beamforming design in caching networks, with the objective of minimizing the network cost. The backhaul cost is measured by the number of files forwarded through backhaul links, which is proportional to the required backhaul capacity as well as the power consumption on backhaul links. We consider partial coordination, where each user will be served by a cluster of BSs and its data need to be made available at each BS of this cluster via backhaul if its data are not in the local cache. The joint design can be formulated as a mixed-integer nonlinear programming (MINLP) problem, which is difficult to solve. Therefore, we break down the problem into two manageable sub-problems: Data assignment

This work is supported by the Hong Kong Research Grant Council under Grant No. 610311.

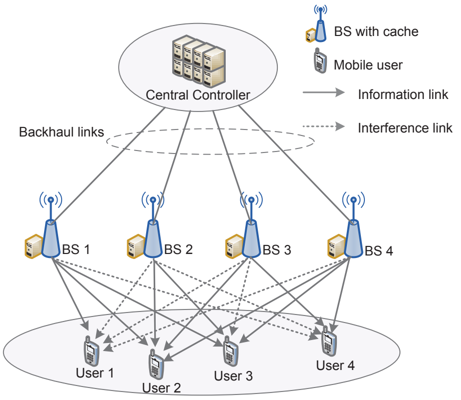

Figure 1. An example of cooperative wireless caching networks.

for minimizing the backhaul cost and coordinated beamforming for minimizing the transmit power. We propose two low-complexity algorithms by exploiting the relationship between the desired beamformer and the sparsity-introducing norm. Through simulations, it will be shown that our proposed design framework can significantly reduce the network cost. Meanwhile, the proposed algorithms provide performance close to the exhaustive search.

II. System Model

A. System Model

We consider a downlink multicell network consisting of a central controller, B base stations, each with N_t transmit antennas, and U single-antenna mobile users. The central controller has access to the whole data library containing F files. All BSs are connected to the central controller through backhaul links, as shown in Fig. 1. Each BS is equipped with a cache, which can store a certain number of files, depending on its storage capacity. For simplicity, we assume all the files are of the same size. Global channel state information (CSI) and the knowledge of cache and request status are all available at the central controller, based on which it will design the beamforming vectors and determine the data assignment for each backhaul link.

Let $\mathcal{Q}_i$ denote the cluster of BSs serving the *i*th user, and thus these BSs should acquire the data signal for the *i*th user, either from the local cache or through the backhaul link. For example, in Fig. 1, the BS clusters for different users are $\mathcal{Q}_1 = \{1\}$, $\mathcal{Q}_2 = \{1,2,3,4\}$, $\mathcal{Q}_3 = \{1,2,4\}$, and $\mathcal{Q}_4 = \{3,4\}$. If BS *j* in $\mathcal{Q}_i$ does not cache the file required by user *i*, it will obtain the file from the central controller via backhaul. Each BS belonging to $\mathcal{Q}_i$ can contribute to the received SINR of the *i*th user, whereas may incur the backhaul cost of getting data from the central controller.

The baseband transmit signal from the *j*th BS is given as $\mathbf{x}_j = \sum_{i=1}^{U} \mathbf{w}_{ij} s_i$, where s_i is an independent complex scalar representing the data symbol for the *i*th user with $E\left[|s_i|^2\right] = 1$, and $\mathbf{w}_{ij} \in \mathbb{C}^{N_t}$ denotes the transmit beamforming vector from the *j*th BS to the *i*th user. Define $\mathbf{w} = \left[\mathbf{w}_{11}^H, \ldots, \mathbf{w}_{1B}^H, \ldots, \mathbf{w}_{U1}^H, \ldots, \mathbf{w}_{UB}^H\right]^H \in \mathbb{C}^{BUn_T}$ as the aggregate beamforming vector. The baseband received signal at the *i*th user can be written as

$$y_i = \underbrace{\sum_{j \in \mathcal{Q}_i} \mathbf{h}_{ij}^H \mathbf{w}_{ij} s_i}_{\text{desired signal}} + \underbrace{\sum_{k \neq i}^{U} \sum_{j \in \mathcal{Q}_k} \mathbf{h}_{ij}^H \mathbf{w}_{kj} s_k}_{\text{intercell interference}} + n_i, \forall i, \qquad (1)$$

where $\mathbf{h}_{ij} \in \mathbb{C}^{N_t}$ is the channel vector from the *j*th BS to the *i*th user and $n_i \sim \mathcal{CN}\left(0, \sigma_i^2\right)$ is the additive Gaussian noise.

We assume that all mobile users adopt single user detection and thus treat interference as noise. The signal-to-interference-plus-noise ratio (SINR) at the *i*th user is given by

$$\mathrm{SINR}_i = \frac{\left|\sum_{j \in \mathcal{Q}_i} \mathbf{h}_{ij}^H \mathbf{w}_{ij}\right|^2}{\sum_{k \neq i}^{U} \left|\sum_{j \in \mathcal{Q}_k} \mathbf{h}_{ij}^H \mathbf{w}_{kj}\right|^2 + \sigma_i^2}, \forall i. \qquad (2)$$

The transmit power constraint of each BS is $\sum_{i=1}^{U} \|\mathbf{w}_{ij}\|_2^2 \leq P_j$, $\forall j$, where P_j is the maximum transmit power of BS *j*.

B. Cache Model and User Request

Define the cache matrix $\mathbf{C} = [c_{ij}] \in \{0,1\}^{F \times B}$, where $c_{ij} = 1$ means the *i*th file is cached in the *j*th BS and $c_{ij} = 0$ indicates the opposite. The user request matrix is denoted as $\mathbf{Q} = [q_{ij}] \in \{0,1\}^{F \times U}$, where $q_{ij} = 1$ means the *i*th file is requested by the *j*th user and $q_{ij} = 0$ means the opposite. We assume that each time each user requires one file and each file is requested by at most one user. The central controller has the knowledge of both $\mathbf{C}$ and $\mathbf{Q}$ and hence can obtain the cache association matrix $\mathbf{L} = [l_{ij}] \in \{0,1\}^{U \times B}$, where $l_{ij} = 1$ means that the data requested by the *i*th user is cached in the *j*th BS. Thus non-zero elements in $\mathbf{L}$ indicate that corresponding links may have backhaul cost.

After obtaining the cache association $\mathbf{L}$ and CSI, the central controller will make a strategic decision on *data assignment* as well as *beamforming*. Denote the data assignment matrix as $\mathbf{N} = [n_{ij}] \in \{0,1\}^{U \times B}$, where $n_{ij} = 1$ means that the data requested by the *i*th user will be assigned to the *j*th BS via backhaul. The cooperation status matrix is denoted as $\mathbf{T} = [t_{ij}] \in \{0,1\}^{U \times B}$, where $t_{ij} = 1$ indicates that the *j*th BS has obtained the data requested by the *i*th user. Therefore, the ℓ_0-norm of N, i.e., $\|\mathbf{N}\|_0$, is the number of files via backhaul, which measures the backhaul cost as mentioned in Section I.

III. Problem Formulation and Analysis

In this section, we will first formulate the network cost minimization problem, which involves joint design of data assignment and beamforming. We will then analyze the problem and provide some insights for the following algorithm design.

A. Problem Formulation

In terms of the caching network cost, we consider two parts: The backhaul cost and the transmit power cost. According to the definition in Subsection II-B, the backhaul cost is measured by $\|\mathbf{N}\|_0$. On the other hand, the transmit power of all BSs is given by $\sum_{i=1}^{U}\sum_{j=1}^{B}\left\|\mathbf{w}_{ij}\right\|_2^2$. We are seeking to reduce the backhaul and transmit power cost simultaneously. However, two aspects conflict with each other. More backhaul data can improve the cooperation of BSs so that they can provide a higher beamforming gain, which can help reduce transmit power. On the other hand, allocating more data via backhaul will uplift the backhaul cost. As a result, the network cost minimization problem requires a joint design of data assignment and cooperative transmit beamforming. We define the network cost as the weighted sum of the backhaul and transmit power cost. Since such two parts are of different units and magnitudes, a normalized weighted sum is adopted and thus the network cost is given by $C_{\text{N}} = (1-\lambda)\, C_{\text{B}} + \lambda C_{\text{P}}$, where $0 < \lambda < 1$ is the parameter controlling the tradeoff between two parts of cost. C_{B} and C_{P} are the normalized backhaul cost and transmit power cost, defined as

$$C_{\text{B}} = \frac{\left\|\mathbf{N}\right\|_0}{\left\|\mathbf{1}-\mathbf{L}\right\|_0} \text{ and } C_{\text{P}} = \frac{\sum_{i=1}^{U}\sum_{j=1}^{B}\left\|\mathbf{w}_{ij}\right\|_2^2}{\sum_{j=1}^{B} P_j}, \tag{3}$$

which are the ratios of the actual cost to the maximal allowed cost. The corresponding SINR target of the *i*th user is denoted as γ_i. Then the network cost minimization problem can be formulated as

$$\begin{aligned}
\mathcal{P}_0 : \underset{\{n_{ij}\},\{\mathbf{w}_{ij}\}}{\text{minimize}} \quad & C_{\text{N}} && \text{(P0)}\\
\text{subject to} \quad & \text{SINR}_i \geq \gamma_i, \forall i && \text{(C1)}\\
& \sum_{i=1}^{U}\left\|\mathbf{w}_{ij}\right\|_2^2 \leq P_j, \forall j && \text{(C2)}\\
& n_{ij} + l_{ij} \leq 1, \forall i,j && \text{(C3)}\\
& n_{ij} \in \{0,1\}, \forall i,j. && \text{(C4)}
\end{aligned}$$

Problem $\mathcal{P}_0$ is an MINLP problem, which is highly complicated [9]. In the following subsection, we will analyze the difficulty of solving $\mathcal{P}_0$, which motivates us to reformulate it and develop low-complexity algorithms afterward.

B. Problem Analysis

We notice that constraint (C1) of $\mathcal{P}_0$ has a complicated form with the presence of $\mathcal{Q}_k$. Based on $\mathbf{T}$, we define an inactive set $\mathcal{S}_{\mathbf{T}} = \{(i,j)\,|t_{ij}=0\}$, and then rewrite constraint (C1) as

$$\frac{\left|\sum_{j=1}^{B}\mathbf{h}_{ij}^H\mathbf{w}_{ij}\right|^2}{\sum_{k\neq i}^{U}\left|\sum_{j=1}^{B}\mathbf{h}_{ij}^H\mathbf{w}_{kj}\right|^2+\sigma_i^2} \geq \gamma_i, \forall i \qquad \text{(C1a)}$$

$$\mathbf{w}_{ij} = \mathbf{0}, \forall (i,j) \in \mathcal{S}_{\mathbf{T}}. \qquad \text{(C1b)}$$

We first consider the case with a given cooperation status matrix $\mathbf{T}$ for $\mathcal{P}_0$. Once $\mathbf{T}$ is known, the cooperation topology (i.e., $\{\mathcal{Q}_k\}$) is fixed. Hence we can obtain the transmit power minimization beamforming problem

$$\begin{aligned}
\mathcal{P}_{\text{MP}} : \underset{\{\mathbf{w}_{ij}\}}{\text{minimize}} \quad & \sum_{i=1}^{U}\sum_{j=1}^{B}\left\|\mathbf{w}_{ij}\right\|_2^2\\
\text{subject to} \quad & \text{(C1a), (C1b), (C2)}
\end{aligned}$$

with the optimal value denoted as $p^\star(\mathbf{T})$. $\mathcal{P}_{\text{MP}}$ can be shown to be a second-order cone programming (SOCP) problem [10] and can be solved efficiently using the interior-point method with computational complexity as $\mathcal{O}\left(B^{3.5}U^{3.5}n_T^{3.5}\right)$ [11].

The above analysis implies that once the optimal $\mathbf{T}^\star$ is identified, the optimal $\mathbf{w}^\star$ can be immediately determined by solving $\mathcal{P}_{\text{MP}}$. Although $\mathbf{T}^\star$ can be found by conducting an exhaustive search, i.e.,

$$\mathbf{T}^\star = \underset{\mathbf{T}\in\{0,1\}^{U\times B}}{\arg\min}\; p^\star(\mathbf{T}), \tag{4}$$

searching over 2^{BU} possible $\mathbf{T}$'s means that the total number of SOCP problems required to solve grows exponentially with BU, making this approach unscalable. Therefore, the key step in solving $\mathcal{P}_0$ is to effectively determine $\mathbf{T}^\star$. Before moving on, let us present a few observations. First, $\mathbf{N}=\mathbf{T}-\mathbf{L}$ can be fully specified with the knowledge of $\mathbf{w}$ as

$$n_{ij} = \begin{cases} I\left(\left\|\mathbf{w}_{ij}\right\|_2 > 0\right), & \text{if } (i,j) \notin \mathcal{S}_{\mathbf{L}}\\ 0, & \text{if } (i,j) \in \mathcal{S}_{\mathbf{L}} \end{cases} \tag{5}$$

with $\mathcal{S}_{\mathbf{L}} = \{(i,j)\,|l_{ij}=1\}$ and $I(\cdot)$ denoting an indicator function. In addition,

$$\left\|\mathbf{N}\right\|_0 \leq \left\|\mathbf{w}\right\|_{0,2} \leq \left\|\mathbf{T}\right\|_0 = \left\|\mathbf{N}\right\|_0 + \left\|\mathbf{L}\right\|_0, \tag{6}$$

where the mixed ℓ_0/ℓ_2-norm of $\mathbf{w}$ is defined as $\left\|\mathbf{w}\right\|_{0,2} \triangleq \sum_{i=1}^{U}\sum_{j=1}^{B} I\left(\left\|\mathbf{w}_{ij}\right\|_2 > 0\right)$. As each $\mathbf{w}_{ij}$ can be regarded as a group, $\left\|\mathbf{w}\right\|_{0,2}$ is referred to as a measure of group sparsity of $\mathbf{w}$. These two observations will facilitate the construction of $\mathbf{T}^\star$, or equivalently $\mathbf{N}^\star$.

In $\mathcal{P}_0$, minimizing C_{N} indicates that both $\left\|\mathbf{N}\right\|_0$ and $\left\|\mathbf{w}\right\|_2^2$ have to be kept as small as possible once the constraints (C1)~(C4) are met. To some extent, the minimization of $\left\|\mathbf{w}\right\|_{0,2}$ also implies minimizing $\left\|\mathbf{N}\right\|_0$ because of (6). However, this does not guarantee the minimization of $\left\|\mathbf{w}\right\|_2^2$. The way to circumvent this difficulty is to introduce a relaxed group sparsity measure, i.e.,

$$\left\|\mathbf{w}\right\|_{1,2} \triangleq \sum_{i=1}^{U}\sum_{j=1}^{B}\left\|\mathbf{w}_{ij}\right\|_2. \tag{7}$$

This mixed ℓ_1/ℓ_2-norm has been shown to be a suitable surrogate for the ℓ_0/ℓ_2-norm [6], [12]. Moreover, $\left\|\mathbf{w}\right\|_2^2$ is upper bounded by $\left\|\mathbf{w}\right\|_{1,2}^2$ [11]. Thus, by solving $\mathcal{P}_0$ with the objective function replaced by $\left\|\mathbf{w}\right\|_{1,2}$ and with $\mathbf{T}$ temporarily assumed to be $\mathbf{1}$, we expect to obtain a solution $\tilde{\mathbf{w}}^\star$that can act as a reasonable approximation of $\mathbf{w}^\star$. Following this, a fair estimate of $\mathbf{N}^\star$, denoted as $\tilde{\mathbf{N}}^\star$, can be obtained from $\tilde{\mathbf{w}}^\star$ and the function (5).

IV. Group Sparse Beamforming Algorithms

The discussion in the previous section suggests a method for obtaining $\tilde{\mathbf{w}}^\star$ and $\tilde{\mathbf{N}}^\star$ as approximations of $\mathbf{w}^\star$ and $\mathbf{N}^\star$. However, $\tilde{\mathbf{w}}^\star$ and $\tilde{\mathbf{N}}^\star$ are not guaranteed to be feasible for $\mathcal{P}_0$. In this section, we will develop two low-complexity algorithms to refine $\tilde{\mathbf{w}}^\star$ and $\tilde{\mathbf{N}}^\star$ and ensure their feasibility, which

```
Algorithm 1 : Full Group Sparse Beamforming Algorithm
Step 1: Solve the full group-sparsity optimization problem $\mathcal{P}_{\text{F-GSBF}}$. Obtain the beamformer $\tilde{\mathbf{w}}$.
Step 2: Initialize $\tilde{\mathbf{w}}^\star = \tilde{\mathbf{w}}$ and $\tilde{\mathbf{N}}^\star = \mathbf{1} - \mathbf{L}$.
Step 3: Obtain the set $E_{\mathbf{N}}$ from $\tilde{\mathbf{N}}^\star$. Use the selection criterion (8) to select $\tilde{n}^\star_{i^*j^*}$ and set $\tilde{n}^\star_{i^*j^*} = 0$.
Step 4: Update $\mathbf{T}$ from (9). Solve the minimum power optimization problem $\mathcal{P}_{\text{MP}}$.
    1) If $\mathcal{P}_{\text{MP}}$ is feasible, obtain the updated $\tilde{\mathbf{w}}^\star$ and $\tilde{\mathbf{N}}^\star$, go to Step 3.
    2) If $\mathcal{P}_{\text{MP}}$ is infeasible, go to Step 5.
Step 5: Obtain the desired solution $\mathbf{w}^\star = \tilde{\mathbf{w}}^\star$ and $\mathbf{N}^\star = \tilde{\mathbf{N}}^\star$.
End
```

are inspired by the group sparse beamforming framework proposed in [12].

A. Full Group Sparse Beamforming Algorithm

The first proposed algorithm is called as the full group sparse beamforming (F-GSBF) algorithm. It is a two-stage approach, which is described as follows.

1) Group-sparsity norm minimization: Solve an SOCP problem at this stage, i.e., $\mathcal{P}_{\text{F-GSBF}}$, which is written as

$$\mathcal{P}_{\text{F-GSBF}} : \underset{\{\mathbf{w}_{ij}\}}{\text{minimize}} \quad \|\mathbf{w}\|_{1,2}$$
$$\text{subject to} \quad \text{(C1a), (C2)}.$$

2) Search Procedure: After obtaining the approximate sparse beamformer $\tilde{\mathbf{w}}$ by solving $\mathcal{P}_{\text{F-GSBF}}$, the next task is to determine $\tilde{\mathbf{N}}^\star$. Instead of using (5) to specify $\tilde{\mathbf{N}}^\star$, we first assume that $\tilde{\mathbf{w}}^\star = \tilde{\mathbf{w}}$ and $\tilde{\mathbf{N}}^\star = \mathbf{1} - \mathbf{L}$, and then iteratively update them. Each time an element from the set $E_{\mathbf{N}} = \left\{ \tilde{n}^\star_{ij} \in \tilde{\mathbf{N}}^\star \middle| \tilde{n}^\star_{ij} = 1 \right\}$ is selected and set to zero. The selection criterion is given by

$$(i^*, j^*) = \underset{(i,j)}{\arg\min} \left\{ \|\mathbf{h}_{ij}\|_2 \|\tilde{\mathbf{w}}_{ij}\|_2 \middle| \ \tilde{n}^\star_{ij} \in E_{\mathbf{N}} \right\}, \quad (8)$$

which means that $\tilde{n}^\star_{i^*j^*}$ will be selected if it has relatively small contribution to the received signal power. After having $\tilde{n}^\star_{i^*j^*} = 0$, the matrix $\mathbf{T}$ in $\mathcal{P}_{\text{MP}}$ is specified as

$$t_{ij} = \begin{cases} 1, & \text{if } \tilde{n}^\star_{ij} + l_{ij} = 1, \\ 0, & \text{otherwise.} \end{cases} \quad (9)$$

If $\mathcal{P}_{\text{MP}}$ is feasible, we will regard its solution as an updated $\tilde{\mathbf{w}}^\star$ and modify $\tilde{\mathbf{N}}^\star$ with $\tilde{n}^\star_{i^*j^*} = 0$. Also, we will go back to set another element in $E_{\mathbf{N}}$ equal to zero, and then solve $\mathcal{P}_{\text{MP}}$, and so on. If there is no feasible solution for $\mathcal{P}_{\text{MP}}$, we claim that the present pair of $\tilde{\mathbf{w}}^\star$ and $\tilde{\mathbf{N}}^\star$ is the desired solution. The F-GSBF algorithm is summarized in Alg. 1.

B. Partial Group Sparse Beamforming Algorithm

As $\|\mathbf{w}_{\text{P}}\|_{0,2} = \|\mathbf{N}\|_0$, where $\mathbf{w}_{\text{P}} = \{[\mathbf{w}_{ij}]| \ (i,j) \notin \mathcal{S}_{\mathbf{L}}\}$ is a partial vector of $\mathbf{w}$, we can focus on the minimization of $\|\mathbf{w}_{\text{P}}\|_{1,2}$, which is the relaxed measure of $\|\mathbf{w}_{\text{P}}\|_{0,2}$. Meanwhile, the other partial vector of $\mathbf{w}$, denoted as $\mathbf{w}'_{\text{P}} = \{[\mathbf{w}_{ij}]| \ (i,j) \in \mathcal{S}_{\mathbf{L}}\}$, also has to be considered since it contributes to the transmit power cost. Based on the above discussion, we formulate the partial group-sparsity optimization problem as

Table I

SIMULATION PARAMETERS

Parameter	Value
BS transmit antenna power gain φ	9 dBi
Standard deviation of log-norm shadowing σ_ζ	8 dB
Distribution of the small scale fading $\mathbf{g}_{ij}$	$\mathcal{CN}(0,1)$
Maximum transmit power of each BS P_{max}	1 W
Noise power σ_i^2 over 10 MHz bandwidth	−102 dBm

$$\mathcal{P}_{\text{P-GSBF}} : \underset{\{\mathbf{w}_{ij}\}}{\text{minimize}} \quad \|\mathbf{w}_{\text{P}}\|_{1,2} + \|\mathbf{w}'_{\text{P}}\|_{2,2}$$
$$\text{subject to} \quad \text{(C1a), (C2)}.$$

The corresponding partial group sparse beamforming (P-GSBF) algorithm is very similar to the F-GSBF algorithm except that Step 1 is replaced by solving $\mathcal{P}_{\text{P-GSBF}}$.

For the F-GSBF algorithm and the P-GSBF algorithm, the number of SOCP subproblems grows linearly with BU and $(BU - \|\mathbf{L}\|_0)$, respectively.

V. SIMULATION RESULTS

In this section, we evaluate the performance of the proposed algorithms. We assume that BSs and mobile users are uniformly and independently distributed in a square region of 1000 meters on each side. Each BS is equipped with 2 antennas. The distance between the *i*th user and the *j*th BS is d_{ij}. The channel vector from the *j*th BS to the *i*th user is modeled as

$$\mathbf{h}_{ij} = \sqrt{10^{-PL_{ij}/10}\varphi\zeta_{ij}}\mathbf{g}_{ij}, \quad (10)$$

where PL_{ij} is the path loss at distance d_{ij}, φ is the BS transmit antenna power gain, ζ_{ij} is the log-normal shadowing coefficient and $\mathbf{g}_{ij}$ is the small-scale fading coefficient. We adopt the 3GPP Long Term Evolution (LTE) standard to depict path loss, i.e., $PL^{\text{dB}}_{ij} = 148.1 + 37.6 \log_{10}\left(d^{\text{km}}_{ij}\right)$. The parameters are shown in Table I.

First we compare the performance of different algorithms. Consider a network of 3 BSs and 3 mobile users, with a fixed cache association $\mathbf{L} = [1\,0\,0;\ 0\,0\,1;\ 0\,1\,0]$. The weighting coefficient is set to be $\lambda = 0.5$, which means that the backhaul cost and the transmit power cost are of equal weight. In Fig. 2, the results are averaged over 1000 independent realizations. We provide two cases as benchmarks. Benchmark I is the case where the data assignment is obtained directly using $\mathbf{N} = \mathbf{1} - \mathbf{L}$ without any design, after which the transmit power minimization beamforming is performed. Benchmark II is the exhaustive search, acting as a global optimal value. Compared to Benchmark I, both the F-GSBF algorithm and the P-GSBF algorithm can reduce the cost, which means the data assignment design will bring benefits. Moreover, with only linear complexity, the two proposed algorithms are close to Benchmark II. Overall, the result demonstrates the effectiveness of the proposed low-complexity algorithms.

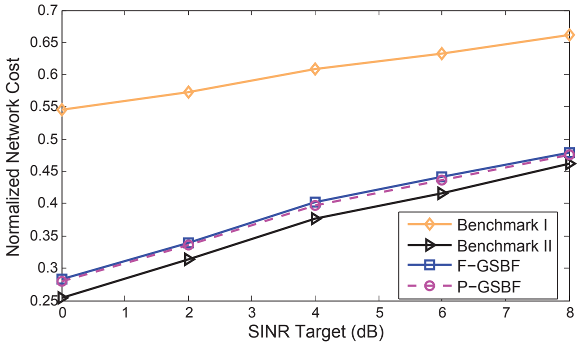

Figure 2. Average normalized network cost of different algorithms for a given cache association.

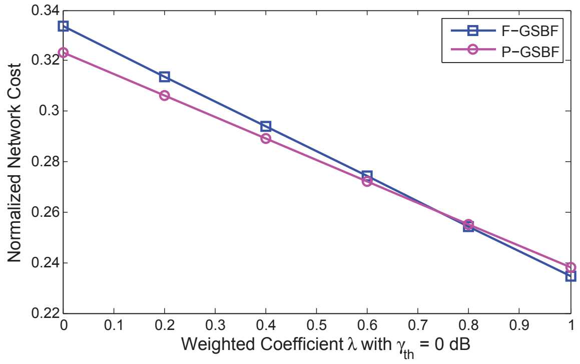

Figure 3. Average normalized network cost versus λ.

Fig. 3 demonstrates the influence of λ on the network cost, using the same set of simulation parameters as for Fig. 2. When λ increases, it means that the backhaul cost will play a less important role. In contrast, when λ approaches zero, the backhaul cost will dominate the total network cost. As shown in the figure, the P-GSBF algorithm outperforms the F-GSBF algorithm when we emphasize the backhaul cost. On the other hand, the F-GSBF algorithm is advantageous when the transmit power cost is the main concern.

We next consider a more general scenario where the number of 1's in $\mathbf{L}$, i.e., $\|\mathbf{L}\|_0$, is fixed but their positions are random. A larger $\|\mathbf{L}\|_0$ means that more requested files can be found at the local cache. Fig. 4 illustrates the impact of $\|\mathbf{L}\|_0$ on the network cost, where the results are averaged over 30 realizations of $\mathbf{L}$ and channels, and the weighting coefficient is set to be $\lambda = 0.2$. It can be observed that when $\|\mathbf{L}\|_0$ is small, the performances of the two proposed algorithms are close to each other. When $\|\mathbf{L}\|_0$ becomes larger, the total network cost lowers down and the P-GSBF algorithm outperforms the F-GSBF algorithm.

VI. Conclusions

In this paper, we have proposed a new system architecture, which takes advantage of distributed caching to reduce the backhaul and transmit power cost in cellular networks. To exploit the benefit of caching, we have formulated a joint optimization problem of data assignment and beamforming. As the design problem is an MINLP, two efficient algorithms of low complexity have been proposed by utilizing the group sparsity structure of the beamformer. Simulation results have showed a significant reduction in the network cost by introducing caches. Meanwhile, the proposed low-complexity algorithms can achieve comparable performance to the exhaustive search. Thus, the proposed methodology is promising to improve the performance of cooperative wireless caching networks.

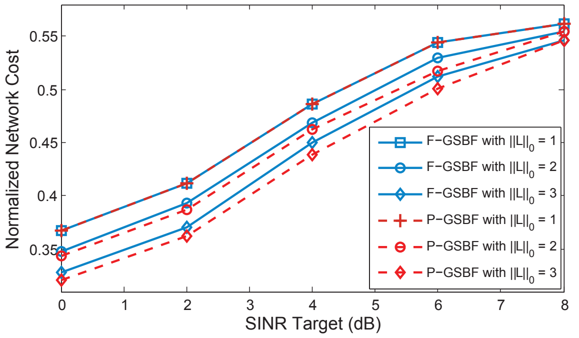

Figure 4. Average normalized network cost of different caching association.

References

[1] 4G Americas, "Supporting wireless video growth and trends," White Paper, Apr. 2013.

[2] 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, Apr. 2013.

[3] 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.

[4] D. Gesbert, S. Hanly, H. Huang, S. Shamai Shitz, O. Simeone, and W. Yu, "Multi-cell MIMO cooperative networks: A new look at interference," *IEEE J. Sel. Areas Commun.*, vol. 28, no. 9, pp. 1380–1408, Dec. 2010.

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

[6] 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.

[7] 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.

[8] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon, "I tube, you tube, everybody tubes: Analyzing the world's largest user generated content video system," in *Proc. ACM IMC*, New York, NY, Oct. 2007.

[9] M. Tawarmalani and N. V. Sahinidis, "Global optimization of mixed-integer nonlinear programs: A theoretical and computational study," *Math. Program.*, vol. 99, no. 3, pp. 563–591, Apr. 2004.

[10] A. Gershman, N. Sidiropoulos, S. Shahbazpanahi, M. Bengtsson, and B. Ottersten, "Convex optimization-based beamforming," *IEEE Signal Process. Mag.*, vol. 27, no. 3, pp. 62–75, May 2010.

[11] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[12] 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.