IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 20, 2017, accepted October 30, 2017, date of publication November 3, 2017,
date of current version December 5, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2769664

Distributed Cache Placement and User Association in Multicast-Aided Heterogeneous Networks

SHUO HE, HUI TIAN, (Member, IEEE), XINCHEN LYU, GAOFENG NIE, AND SHAOSHUAI FAN
State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China

Corresponding author: Hui Tian (tianhui@bupt.edu.cn)

This work was supported in part by the National Nature Science Foundation of China under Grant 61471060 and in part by the National Science and Technology Major Project under Grant 2017ZX03001003.

ABSTRACT The small cells have been equipped with caching and multicast capabilities to save energy and ease backhaul burden. However, given the increasing diversity of user association in heterogeneous networks, traditional schemes may fail to exploit the energy-saving potential of caching and multicast. In this paper, we propose the model to minimize total power consumption by jointly optimizing the user association and cache deployment. The formulated joint optimization problem is decomposed and reformulated as a NP-complete set partition problem. Motivated by the idea of the tactile networks, the devices can find the candidate user grouping based on its own utility. The utility function is judiciously designed to reduce the searching space without compromising optimality. Then, a heuristic caching algorithm is proposed by rigorously deriving the upper and lower bounds of cache placement. Simulation results show that our proposed scheme outperforms the other existing multicast and caching algorithms in terms of power consumption by up to 28%, while keeping the load among base stations balanced.

INDEX TERMS Heterogeneous networks, multicast transmission, cooperative caching, set partition problem.

I. INTRODUCTION

With the ever-increasing smart mobile devices, the needs for emerging applications and services are exploding, warning a sharp increase in mobile data traffic [1]. Dense deployment of small base stations (SBSs) is a key approach to support the unprecedented growth of mobile data traffic [2]–[4]. However it exerts huge pressure on backhaul links and increases energy consumption on the wireless network side. Enabling the fifth generation of mobile technology (5G), approaches to alleviating the backhaul burden and achieving green communications are necessary but thorny. In addition, efficient approaches to manipulating burst flow generated in the case of disaster recovery environment or live sport matches are also in active demand. Fortunately, the central features of the ever-increasing content demand provide further chances to handle the problems above.

The first feature of the traffic demand is redundant [5], i.e., majority of the requests are generated for only a small number of popular services. The repeated requests for the same contents and the dense deployment of SBSs both will exert huge pressure on backhaul links. To alleviate the burden of backhaul, endowed with caching in the SBSs is a promising approach [5]. By storing possibly reusable contents in advance, the burden of wireless backhaul can be significantly reduced because of the proactive assist of SBSs which cache the required content. However, as pointed in [6], the most popular contents should be given a high priority while caching when BSs are sparsely deployed. While the BSs are densely deployed, the design of caching strategy is challenging.

The second feature of the traffic demand is group-oriented [7]. Plenty of group-oriented applications are emerging, such as one to many files transfer, military group action in battlefield [7] and Samsung's Group Play service [8]. Besides, requests for live sport matches and some new released videos might also be generated simultaneously [9]. However, the simultaneous requests will lead to the network congestion especially on limited wireless capacity. To relieve the ''on air'' congestion, enabling multicasting at Base Stations(BSs) is considered as an effective approach to serve requests for the same content occurring simultaneously or in a time-limited window.

Though the two techniques mentioned above seem to target for different directions, they can be integrated seamlessly

to improve the performance of wireless network. Especially with the ever-growing energy consumption due to the steep increase of service demand, the development of green 5G networks has become a mainstream concern. However, existing works which jointly consider cache placement and multicast scheduling have not considered the influence of cooperative caching promoted by the dense deployment of SBSs. Specially, cooperative caching can enable more user traffic to be offloaded to the low-power small cells to alleviate congestion at high-power MBSs in heterogenous networks [10].

In this paper, we focus on the cooperative multicast scheduling and cooperative caching placement in heterogeneous wireless networks. Due to the dense deployment of SBSs in the next generation networks, increasing the diversity of cached contents by cooperative caching can offload more data traffic. In addition, much lower transmit power is required when the user request is served by an SBS in the vicinity compared with the power consumption required by an MBS [11]. Hence, the potential power-saving gain can be achieved by enabling the cooperative multicast scheduling among neighboring SBSs to exploit all the possible multicast opportunities. The main contributions of our paper are summarized as follows:

- Facilitated with the densely deployed SBSs, we exploit the possible multicast opportunities with cooperative caching placement. Unlike [17]–[21], we take into consideration collaboration at BSs both for the design of multicast scheduling and caching placement. And a joint cooperative caching and multicast scheduling model minimizing the system power consumption is proposed, in which more user requests can be offloaded to the lower-power SBSs rather than served by the higher-power MBS. Thus the system load can be better balanced.
- The developed model is proved NP-hard. To make the problem more tractable, we introduce the definitions of user group and user cluster and then decompose the problem into two subproblems: cooperative multicast scheduling and cooperative caching placement.
- Furthermore, we present a distributive user association algorithm for further utilizing the Set Partition Problem (SPP) to solve the subproblem of cooperative multicast scheduling. Together with the user association, we propose the Multicast and Cooperative Caching (MCC) algorithm to jointly optimize the cache placement and minimize the system power consumption.

The remainder of this paper is organized as follows: Section II gives a brief review of the main researches on the two techniques. In Section III, the formulation of the multicast scheduling and cooperative cache placement problem is presented and the proposed MCC algorithm is introduced in Section IV. In Section V, effectiveness of the proposed approach is provided and followed by the conclusions of the paper in Section VI. The key notations used throughout the paper are summarized in Table 1.

TABLE 1. List of key notations.

Symbol	Physical Meaning
$\mathcal{N}$	set of all the contents
$\mathcal{K}$	set of all mobile users
$\mathcal{K}'_n$	set of users that request content n
$\mathcal{K}_n^m$	set of users that request content n and are served by BS m
$\mathcal{G}_i^n$	a possible user group indexed by i in which all users request content n
$\mathcal{G}^n$	user cluster consisting several user groups in which all users request content n
P_{mk}	power consumption required by BS m to serve user k
t_{mk}	transmission decision for BS m to user k
x_{mn}	caching decision for content n to BS m
$y_{m\mathcal{G}_i^n}$	multicast decision for BS m to user group $\mathcal{G}_i^n$
c_{mi}	power cost required by BS m to serve user group $\mathcal{G}_i^n$

II. RELATED WORK

Multicast and cache have been regarded as two promising approaches to handle the problems brought by the explosive growth of mobile data traffic. In this paper, we aim to combine the advantages of cooperative multicast scheduling and cooperative caching to further exploit the potential power reduction. In the following, we discuss the related work, emphasizing the main differences compared to our work.

A. RELATED WORK

The caching placement approaches have been studied in [12]–[16]. The joint design and optimization of caching and user association policy minimizing the average download delay were explored in [12]. However, user requests were served through multiple unicast transmissions leading to extra power consumption. Similar to [12], Cui *et al.* [13] studied the optimal caching and user association in heterogeneous networks with wireless backhaul, where the multicast transmission were also not considered. A cooperative transmission scheme under random caching at SBSs was proposed in [14] and the caching distribution was optimized to maximize the successful transmission probability. Chae *et al.* [15] identified a tradeoff between the content diversity and the cooperative gain and proposed a probabilistic caching policy for balancing the tradeoff optimally. However, the works mentioned above were all studied based on unicast transmission. In [7], a general solution addressing the design of caching policy for reliable multicast service was presented. Bao *et al.* [16] proposed an optimal client cache size allocation scheme combined with a multicast technique-batching.

In addition to the researches on caching placement strategies with unicast transmission or multicast-aware transmission for-mentioned, multicast scheduling schemes in cache-enabled wireless networks have also been studied in many existing works [17]–[21]. The optimal dynamic

multicast scheduling to jointly minimize the average delay, power and fetching cost was studied in [17], where the caching design was given. Similar to [17], the work in [18] established a content-centric content request queue model and proposed a structure-aware stochastic content multicast scheduling algorithm to jointly optimize the average delay and service costs for elastic services in heterogeneous cellular networks. Zhang *et al.* [19] designed a combination scheme of parallel transmission where the MBS and SBSs parallely transmitted the requested contents, and cooperative transmission where multiple SBSs cooperatively transmitted the requests to the multicast group users. However, this scheme ignored the cooperative scheduling of MBS and SBSs. In addition, each BS cached the most popular contents ignoring the gain brought by the cooperative caching. Similarly, Cui *et al.* [20] presented a multicasting design with a random caching and derived an asymptotically optimal algorithm to maximize the successful transmission probability for single-tier networks. However, the proposed caching design confined the full usage of content diversity. Both the optimization and performance analysis were extended to the multi-tier heterogeneous networks in [21]. However, both in [20] and [21], the main focus were put on the multicast transmission of requests for the cached contents while we also take into consideration the multicast transmission opportunities for the uncached contents.

While the works mentioned above are optimization of caching policy in multicast-aware networks or design of multicast scheduling scheme in cache-enabled networks, studies on joint optimization of multicast scheduling and cache placement were presented in [22]–[25]. Firstly, Poularakis *et al.* [22] jointly studied the caching policy and multicast scheduling scheme in delay tolerant networks by formulating a discrete optimization problem, which was the most relevant to ours. However, collaboration at the BSs were not considered. Therefore, when the contents that requested by users could not be found in the cache of the local SBS, they could only be served by MBS at a more higher level of power consumption instead of being served by the neighboring SBSs. In addition, the work in [23] addressed the performance limits of client caching enabled video-on-demand services in wireless multicast networks with asynchronous requests. And a joint cache allocation and multicast delivery scheme minimizing the average bandwidth consumption was proposed and analyzed. Combined the advantages of multicast content delivery and cooperative content sharing, a compound caching policy was developed in [24]. Similarly, based on the characteristics of multicast and cooperation among BSs, the cooperative caching scheme for multicasting was proposed in [25] to improve the cache hit ratio and efficiency of content delivery. However, they both ignored the interaction between SBSs. Therefore, the potential of reducing power consumption was not fully exploited.

In our paper, in order to combine the advantages of cooperative multicast transmission and cooperative caching placement, we jointly optimize the caching placement with user association to further explore the potential of power reduction by exploiting all the possible multicasting opportunities. Moreover, the proposed scheme can be implemented in a distributed manner to reduce the signaling and computation complexity.

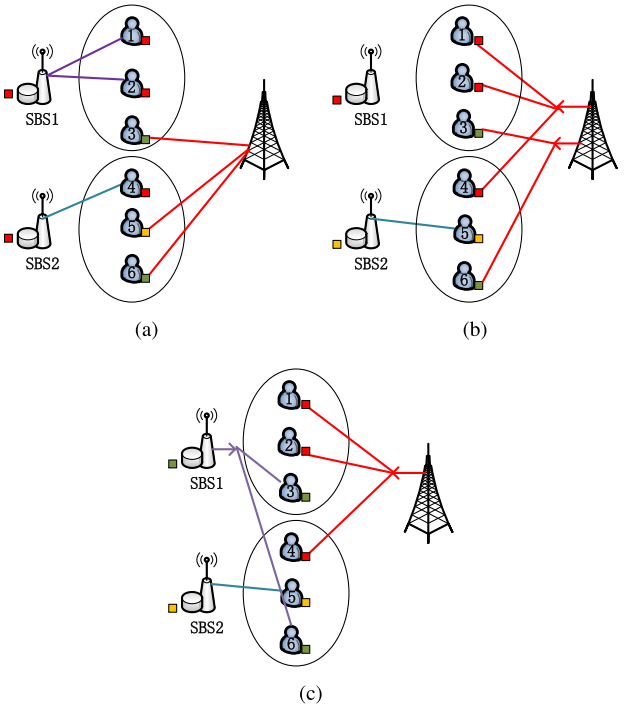


FIGURE 1. A motivating example of the transmission scheme. (a) unicast based caching. (b) multicast based caching. (c) multicast based and cooperative caching.

B. MOTIVATING EXAMPLE

A motivating example of the above modifications is given in Fig. 1. We consider a multicast and cooperative caching system with two SBSs, six users and three contents in the toy example. Schemes shown in Fig. 1(a) and Fig. 1(b) are also utilized in [22]. In the toy example, we assume each SBS can only store one content due to the limited cache capacity.

The unicast based caching scheme is presented in Fig. 1(a). Each user request is served by unicast transmission. The optimal caching decision is to place the most popular content with respect to the local demand in this scheme [22]. Therefore, the most popular content represented by the red diamond is stored in both SBSs. The power consumption in this scheme is $p_u = p_{11} + p_{12} + p_{24} + p_{03} + p_{05} + p_{06}$. Taking p_{12} for example, p_{12} indicates the power cost required by SBS 1 to serve requests generated by user 2. Notably, the index 0 represents MBS. Similar definitions are applied to the other notations.

In Fig. 1(b), the multicast transmission is introduced. Users that request the same content can be formed into a group and then served by a multicast transmission. Therefore, to save power consumption as much as possible, the optimal caching policy changes. SBS 1 caches the content repre-

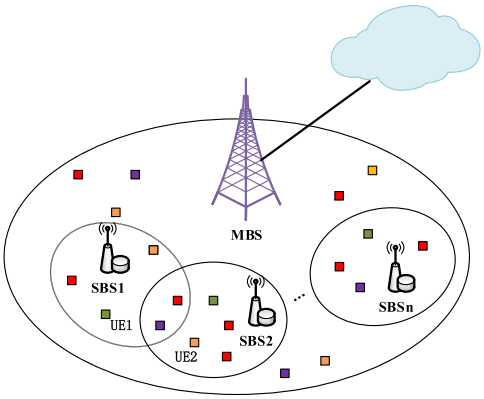


FIGURE 2. Multicast and cooperative cache-enabled heterogeneous wireless networks.

sented by the red diamond according to the local demand and SBS 2 can either store the content represented by the yellow diamond or the green one. User requests for the same content are aggregated and served via MBS multicast transmissions. Hence, the power consumption of the scheme shown in Fig. 1(b) is $p_m = p_{25} + \max\{p_{03}, p_{06}\} + \max\{p_{01}, p_{02}, p_{04}\}$. The last term denotes that the power consumption required by MBS or SBSs to multicast a content to a user group is the highest power required in the group.

However, the multicast and caching scheme illustrated in Fig. 1(b) does not consider the cooperative caching between adjacent BSs. As shown in Fig. 1(c), request for the content represented by the green diamond generated by user 6 is associated to SBS 2 which does not store the content. If user 6 is in the coverage area of SBS 1 at the same time, both requests of user 3 and user 6 can be served by a multicast transmission of SBS 1. Especially when the channel condition between user 6 and MBS is very poor, serving user 6 by SBS 2 is able to reduce the total network power consumption effectively. Hence, the power consumption is $p_c = p_{25} + \max\{p_{13}, p_{16}\} + \max\{p_{01}, p_{02}, p_{04}\}$.

Obviously, we have $p_c \leq p_m \leq p_u$ in most cases. The above example demonstrates the efficiency of the proposed scheme which combines the multicast transmission with cooperative caching and better exploits the available space of content caching and delivering.

III. SYSTEM MODEL

As illustrated in Fig. 2, we consider a heterogeneous network consisting of M base stations (BSs) and K mobile users, where the users can request N contents. Particularly, the BSs are indexed by $\{0, 1, 2, \ldots, M\}$, where the macro BS is denoted by BS 0 and the SBSs are indexed by $m = \{1, 2, \ldots, M\}$. Let $\mathcal{K} = \{1, 2, \ldots, K\}$ denote the mobile users and $\mathcal{N} = \{1, 2, \ldots, N\}$ be the set of contents requested by all the users. Each SBS is able to cache part of the contents, and the macro BS is assumed to have access to all the contents in $\mathcal{N}$.

Due to the ultra dense deployment of small cells, the mobile users in Fig. 2 can be in the coverage of multiple SBSs. Consider a discrete-time communication system, e.g., 5G networks, operating in time slots. In each time slot, users requesting the same content n are assigned with the same color, naturally forming a group $\mathcal{K}'_n$ and being served by multicast transmission. If BS m multicasts a certain content to its scheduling user group, the transmitting power required should satisfy all the pending requests. Users in coverage hole of SBSs can only be served by MBS. Besides, a MBS multicast transmission is also used to satisfy the requests generated by users whose associated SBSs have not cached the requested content.

The popularity of the contents in $\mathcal{N}$ follows the Zipf distribution, and is identical to all users [26], [27]. Particularly, the probability of content n being requested can be given by

$$P(n) = \frac{n^{-\alpha}}{\Sigma_{f=1}^{N} f^{-\alpha}}, \tag{1}$$

where α indicates the skewness of the request distribution [27]. During a time slot, each user is assumed to request at most one content from the content set $\mathcal{N}$ [20].

By referring to the SINR criteria [28], the minimum transmit power of BS m to deliver a content to user k, denoted by P_{mk}, can be given by

$$P_{mk} = P_0 - g_k - g_m + l_{mk} + \psi_k, \tag{2}$$

where P_0 represents the receiver sensitivity of the device; g_k and g_m denote the antenna gain of user k and BS m, respectively; l_{mk} is the path loss between BS m and user k; ψ_k is the shadow component. Typically, the transmission power of the MBS is much higher than that of the SBSs due to the large path loss l_{mk} to provide full coverage of the network [18].

IV. PROBLEM FORMULATION AND TRANSFORMATION

Green communication is one of the major challenges in 5G networks. In this paper, we aim to minimize the power consumption of network operators, while ensuring the quality of service of all the users. The problem is formulated to jointly optimize multicast scheduling and caching policy, given by

$$\begin{aligned}
\mathbf{P}: \min & \sum_{n\in\mathcal{N}} \sum_{m\in\mathcal{M}\cup\{0\}} \max_{k\in\mathcal{K}'_n} P_{mk} \cdot t_{mk}, \\
\text{s.t. C1:} & (1 - x_{mn}) \cdot t_{mk} = 0, \forall n \in \mathcal{N}, k \in \mathcal{K}'_n, m \in \mathcal{A}_k, \\
\text{C2:} & \sum_{m\in\mathcal{M}\cup\{0\}} t_{mk} = 1, \forall n \in \mathcal{N}, k \in \mathcal{K}'_n, \\
\text{C3:} & \sum_{n\in N} x_{mn} \leq S_m, \forall n \in \mathcal{N}, m \in \mathcal{M}, \\
\text{C4:} & t_{mk} \in \{0, 1\}, \forall k \in \mathcal{K}'_n, m \in \mathcal{M}, \\
\text{C5:} & x_{mn} \in \{0, 1\}, \forall n \in \mathcal{N}, m \in \mathcal{M},
\end{aligned} \tag{3}$$

where $x_{mn} \in \{0, 1\}$ denotes whether content n is stored in SBS m ($x_{mn} = 1$) or not ($x_{mn} = 0$). The transmission decision is represented by t_{mk}, where $t_{mk} = 1$ indicates that BS m serves user k, otherwise, $t_{mk} = 0$. Let $\mathcal{A}_k$ be the BS cluster which can serve user k.

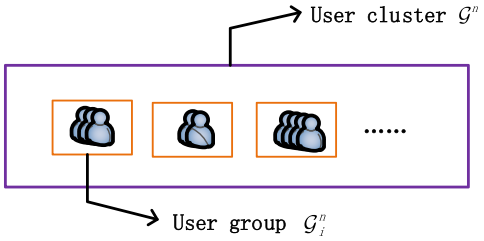


FIGURE 3. The diagrammatic sketch of user group and user cluster.

Constraint $C1$ states that SBS m needs to cache content n to serve the users requesting this content. Constraint $C2$ guarantees that user k can be served by at most one BS at the same time. Constraint $C3$ ensures that the size of stored contents in BS m will not exceed the capacity of its cache. Both x_{mn} and t_{mk} are binary variables, and therefore, constraint $C1$ can be rewritten as

$$\text{C6: } x_{mn} \geq t_{mk}, \forall n \in \mathcal{N}, k \in \mathcal{K}'_n, m \in \mathcal{A}_k. \tag{4}$$

Each BS m will multicast content n to the users in group $\mathcal{K}_n^m = \{k | t_{mk} = 1, \forall k \in \mathcal{K}'_n\}$. In order to satisfy the requirements of all the users in $\mathcal{K}_n^m$, the transmission power should be set as the maximum of the destined users, i.e., $\max_{k\in\mathcal{K}'_n} P_{mk} \cdot t_{mk}$ as shown in the objective of **P**.

Solving the integer programming problem **P** is non-trivial. First, the integer programming problem is NP-hard in general [29]. It will be even challenging to solve **P** when considering its complicate constraints on the optimization variables. Moreover, the transmission decisions are heavily coupled with each other, due to the non-linear objective function. This limits the efficiency of existing integer programming solver [30].

To improve the tractability of problem **P**, we introduce user cluster and user group in Definition 1, and then remove the coupling brought by the maximum notation in the objective.

Definition 1: The user group is a set of users. In this paper, any subset i of $\mathcal{K}'_n$ can be called a user group, denoted by $\mathcal{G}_i^n$. User cluster $\mathcal{G}^n$ is defined as the set of all the user groups, i.e., $\mathcal{G}^n = \{\mathcal{G}_1^n, \mathcal{G}_2^n, \ldots, \mathcal{G}_i^n, \ldots\}$. Namely, we have

$$\mathcal{G}^n = \{\mathcal{G}_i^n : \mathcal{G}_i^n \subseteq \mathcal{K}'_n\}, \tag{5}$$

where $\mathcal{G}^n$ is the collection of all subsets of $\mathcal{K}'_n$. The relationship between user group and user cluster is dictated in Fig. 3.

The main objective of **P** is to optimize caching placement and assign each user with a serving BS. By introducing the definition of user grouping, **P** can be reformulated as selecting the optimal set of user groups in $\mathcal{G}_n$ and the corresponding BS for multicasting the content to each selected user group $\mathcal{G}_i^n$, given by

$$\begin{aligned}
\mathbf{P1}: \min & \sum_{n\in\mathcal{N}} \sum_{m\in\mathcal{M}\cup\{0\}} c_{mi} \cdot y_{m\mathcal{G}_i^n} \\
\text{s.t. } & \text{C3, C5,} \\
& \text{C7: } x_{mn} \geq y_{m\mathcal{G}_i^n}, \forall n \in \mathcal{N}, \mathcal{G}_i^n \in \mathcal{G}^n, \\
& \text{C8: } y_{m\mathcal{G}_i^n} = \{0, 1\}, \forall m \in \mathcal{M}, \mathcal{G}_i^n \in \mathcal{G}^n, \\
& \text{C9: } \cup_{m\in\mathcal{M}} y_{m\mathcal{G}_i^n} \cdot \mathcal{G}_i^n = \mathcal{K}'_n, \\
& \text{C10: } y_{m\mathcal{G}_i^n} \cdot \mathcal{G}_i^n \cap y_{m\mathcal{G}_j^n} \cdot \mathcal{G}_j^n = \emptyset, \forall \mathcal{G}_i^n, \mathcal{G}_j^n \in \mathcal{G}^n, n \in \mathcal{N},
\end{aligned} \tag{6}$$

where $y_{m\mathcal{G}_i^n}$ is a binary variable indicating whether BS m serves user group $\mathcal{G}_i^n$ ($y_{m\mathcal{G}_i^n} = 1$) or not ($y_{m\mathcal{G}_i^n} = 0$). And we have

$$y_{m\mathcal{G}_i^n} = 1 \Rightarrow t_{mk} = 1, \forall k \in \mathcal{G}_i^n. \tag{7}$$

where $c_{mi} = \max_{k\in\mathcal{G}_n^i} P_{mk}$ denotes the cost of assigning the user group $\mathcal{G}_i^n$ to the BS m. In **P1**, the strong coupling among variables in the objective has been removed.

However, **P1** is still non-trivial due to the severe coupling between the cache placement and transmission decisions. Particularly, C9 indicates that the solution should contain all the users requesting content n and C10 implies that each user group in the solutions should be disjoint, i.e., each user can only be served by a BS at a time. These two constraints introduce severe coupling between the cache placement and transmission decisions.

V. THE DESIGN OF MCC ALGORITHM

In this section, we decomposed **P1** into two subproblems, i.e., (a) multicast scheduling given the cache placement, and (b) optimizing cache placement given the multicast scheduling of subproblem (a). We solve these two subproblems iteratively to obtain the solution for **P1**. In subproblem (a), the transmission decisions are decoupled between different requested contents of the users, and can be decomposed into N subproblems, given by

$$\begin{aligned}
\mathbf{P2}: \min & \sum_{m\in\mathcal{M}\cup\{0\}} c_{mi} \cdot y_{m\mathcal{G}_i^n} \\
& \text{s.t. C8, C9 and C10.}
\end{aligned} \tag{8}$$

In the following, we focus on the decomposed subproblem **P2** and propose the Multicast and Cooperative Caching (MCC) algorithm to jointly optimize the transmission scheduling and cache placement iteratively.

A. SET PARTITION PROBLEM (SPP)

Given all possible user groups, **P2** can be regarded as a set partition problem (SPP) [31] with the following structure

$$\begin{aligned}
\mathbf{SPP}: \min & \sum_{j=1}^{n} c_j x_j \\
\text{s.t. } & \sum_{j=1}^{n} a_{ij} x_j = 1, \forall i = 1, 2, \ldots, m, \\
& x_j = \{0, 1\}, \quad \forall i = 1, 2, \ldots, m,
\end{aligned} \tag{9}$$

where $I = \{1, 2, \ldots, m\}$ and $J = \{1, 2, \ldots, n\}$ denote the set of row and column index, respectively. Let $P_j = \{i \in I | a_{ij} = 1\}$ be the set of row index with value of 1 in the column j. Solution J' is a partition to (9), when satisfying

$$\cup_{j\in J'} P_j = I, \tag{10}$$

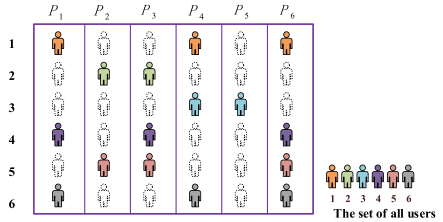


FIGURE 4. A sketch map of the user partition, which is the coefficient matrix for users (rows) and user groups (columns). The columns 1, 2 and 5 or the columns 3 and 4 form a possible partition.

$$P_j \cap P_h = \emptyset, \quad \forall j, h \in J', j \neq h. \tag{11}$$

Fig. 4 illustrates the relation between set partition in SPP and user partition in **P2**. In Fig. 4, the set of all users can be regarded as the set of row index in SPP, i.e., $I = \{1, 2, 3, 4, 5, 6\}$. P_j stands for a possible user group j, e.g., $a_{22} = 1$ and $a_{52} = 1$, hence, $P_2 = \{2, 5\}$, namely user 2 and user 5 belong to user group 2. Obviously, user groups 3 and 4 contain all users in I and users in each group are different. Therefore, user group 3 and 4 can be considered as a user partition of I.

The objective of SPP is to find the best partition with minimal cost among all the possible partitions. Similarly, given all the possible user groups, **P2** aims to determine the best user partition to minimize the total power consumption required by each associated BS among all the possible user partitions.

Hence, the problem **P2** should be solved in two phases: user association and transmission determining. Due to the complexity of obtaining all the subsets of $\mathcal{K}'_n$ and the limit of the coverage of BSs, firstly, we design an efficient algorithm to partition users into different possible groups. Then, we can chain these groups to feasible clusters, i.e., feasible partitions.

B. THE DESIGN OF THE ALGORITHM

In this section, we reduce the searching space of user grouping by judiciously designing a utility function to specify a set of candidate optimal user groups. Then, a genetic algorithm (GA) is adopted to obtain a near-optimal solution to **P2**. Discovering the upper and lower bounds of **P2**, we propose a heuristic algorithm to place the contents for each BS.

Note that the complexity of the proposed algorithm, summarized in Algorithm 3, mainly depends on finding all the possible user groups. By exploiting the proposed utility, we can reduce the complexity and distribute the computation to each mobile devices. This will enhance the practicality and efficiency of the proposed algorithm.

1) USER ASSOCIATION

As the expanding of network size, the number of possible user groups is tremendous. Due to the limit of coverage areas of each BS, substantial redundancy of user groups can be reduced.

Each BS can only serve users in its own coverage area. Therefore, for any BS $m \in \mathcal{M}'_n$, users in its coverage area can be formulated as a possible group. Specially, for the set of users only in the coverage area of MBS and users in the coverage area of the SBSs that have not cached the requested content, they should be classified into the same group. This group is recorded as $\mathcal{G}_2^n$ and can only be served by MBS.

Remarkably, the transmission decision for users in overlapping is non-trivial. In this paper, we define a utility function to describe the gain of a transmission assignment for users in overlapping areas shown in (12).

$$u_{ik} = max\{c(\mathcal{G}_i^n), c(\mathcal{G}_i^n \cup k)\} - max\{c(\mathcal{G}_i^n)\}. \tag{12}$$

u_{ik} stands for the cost increase of assigning user k to group i. $c(\mathcal{G}_i^n)$ denotes the power consumption required by the associated BS of group $\mathcal{G}_i^n$ to serve all the users requesting content n in group $\mathcal{G}_i^n$ by a multicast transmission.

Lemma 1: User k should be associated to the group $\mathcal{G}_i^n$ with the minimum value of u_{ik} without loss of optimality. And there is $u_{ik} \geq 0$.

Proof: See Appendix A. ■

Note that the utility function can significantly reduce the candidate user groups, without comprising the optimality of **P2**. Moreover, the user grouping computation can be distributed to each mobile device. This can be achieved by enabling devices to associate with proper BSs and reporting the grouping results to the MBS to find the best user group.

The main process of user association is presented in Algorithm 1.

Algorithm 1 The UC Algorithm

Require:
x_{mn}, $\mathcal{M}'$, $\mathcal{K}'_n$;
Ensure:
$\mathcal{G}^n = \{\mathcal{G}_1^n, \mathcal{G}_2^n, \ldots, \mathcal{G}_i^n, \ldots\}$, $c^n = \{c_1^n, c_2^n, \ldots, c_i^n, \ldots\}$;
1: Initialize $\mathcal{G}_1^n = \mathcal{K}'_n$, $c_1^n = \max_{k \in c_1^n} P_{mk}$, c_2^n, $c_2^n = \max_{k \in \mathcal{G}_2^n} P_{mk}$, i=3;
2: **for all** SBSs $m \in \mathcal{M}'$ which satisfy that $x_{mn} = 1$ **do**
3: $\mathcal{G}_i^n = \{k | k \in \mathcal{K}'_n \cap k \in B_m\}$,$c_i^n = \max_{k \in c_i^n} P_{mk}$;
4: i++;
5: **end for**
6: **for all** users k in overlapping areas **do**
7: **for all** SBSs m that user k belongs to **do**
8: calculate the utiliy u_{mk} and select the minimum one m';
9: $\mathcal{G}_i^n = \mathcal{G}_{m'+2}^n \cup k$, $c_i^n = \max_{k \in \mathcal{G}_i^n} P_{mk}$;
10: i++;
11: **end for**
12: **end for**
13: **for all** clusters exist in $\mathcal{G}^n$ **do**
14: add all possible combinations of the groups $\{\mathcal{G}_3^n, \mathcal{G}_4^n, \ldots, \mathcal{G}_i^n\}$ to $\mathcal{G}_2^n$ to form new groups and calculate c_i^n;
15: **end for**

As outlined in Algorithm 1, we consider the set of users only in the coverage area of each BS as possible user groups.

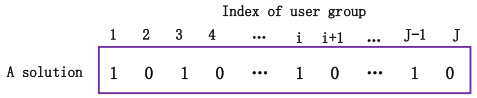


FIGURE 5. Code representation.

Then for users in overlapping areas, we utilize the utility function to determine their serving BS. This process can reduce the amount of user groups which can not reach the objective function in problem **P2**. The objective of line $13-15$ is to select the set of users to be served by a multicast transmission of MBS rather than served by SBSs. The above steps can remove users with poor channel conditions from the user group which will be served by MBS in the multicast-aware caching algorithm in [22].

2) TRANSMISSION DETERMINING

In this section, we propose the Transmission Determining (TD) heuristic algorithm based on the Genetic Algorithm (GA) to solve problem **P2**. The GA provides an efficient approach to solve the NP-complete problem [32].

- Coding Representation. Considering the binary encoding scheme, each user group has a identifier $j, j = 1, 2, \ldots, J$, where J is the total number of user groups obtained by Algorithm 1. A solution i can be represented as $\mathcal{S}_i = \{\mathcal{G}_1^{ni}, \ldots, \mathcal{G}_j^{ni}, \ldots, \mathcal{G}_J^{ni}\}$, where $\mathcal{G}_j^{ni}$ is a $0-1$ binary variable and denotes whether user group $\mathcal{G}_j^n$ is selected in solution S_i, i.e., $\mathcal{G}_j^{ni} = 1$ indicates that users in group $\mathcal{G}_j^n$ will be selected as a whole to be served by a certain BS and 0 otherwise. Specially, when $\mathcal{G}_j^n = 1\,or\,0$, we have $y_{m\mathcal{G}_j^n} = 1\,or\,0$, where m is the associated BS of group $\mathcal{G}_j^n$. Fig. 5 shows a coding example.
- Fitness and Selection. Given a possible solution S_i, the fitness of the solution is given as follows:

$$f_i^n = \sum_{j=1}^{J} c_j^n \cdot \mathcal{G}_j^{ni}. \tag{13}$$

The tournament selection method [33] is used in which d solutions are selected randomly and the one with the highest fitness remains for further genetic processing. The process repeats until the mating pool is filled.

- Crossover and Mutation. To explore more promising and new search space, many crossover techniques exist. In this paper, we adopt the uniform crossover operator [34].

After crossover, mutation is applied to each child which is an operator that will change the value of a certain gene. The GA introduces mutation to improve the local search ability and maintains the population diversity.

However, the child generated by the crossover and mutation may be infeasible due to the strong constraints in problem **P2**. Some users may be assigned to different BSs at the same time or served by no BSs. The improvement algorithm in [34] is used to modify the feasibility of the filial generation.

In summarize, the TD heuristic algorithm based on the genetic algorithm is shown in Algorithm 2.

Algorithm 2 The TD Algorithm

```
Set the number of iteration t := 0;
Initialize S(t);
Evaluate S(t);
for each generation do
    Select {S_1, S_2} from S(t);
    g_child = crossover(S_1, S_2);
    g_child = mutate(S_1, S_2);
    Repair the feasibility of g_child;
    if g_child ∈ S(t) then
        delete g_child and go to 5;
    else
        S(t+1) ← S(t) ∪ g_child;
    end if
    Evaluate S(t+1);
    t ← t+1;
end for
Find min_{S∈S(t)} S;
Set S* ← S;
return S*, f(S*).
```

3) THE MULTICAST AND COOPERATIVE CACHING ALGORITHM

In the above sections, the problem **P2** is solved with the given cache placement. Next, we are going to determine the cache placement based on the proposed heuristic algorithm.

Firstly, we analyze the results of problem **P2**. Let $\{t_{mk}^*\}$ be the set of sub-optimal solution obtained by Algorithm 2. P^* is denoted as the corresponding power consumption, which can be expressed as

$$P^* = \sum_{m \in \mathcal{M} \cup \{0\}} \max_{k \in \mathcal{K}_n'} P_{mk} \cdot t_{mk}^*. \tag{14}$$

Assume that the capacity of each cache is infinite, the power consumption of the corresponding solution in problem **P2** is denoted by $\underline{P}$.

Lemma 2: The power consumption of multicast and cooperative cache can be bounded by

$$\underline{P} \le P^* \le \overline{P} = \max_{k \in \mathcal{K}_n'} P_{0k} \cdot t_{0k}.$$

Proof: See Appendix B. ■

By incorporating Lemma 2 and the property in (4), we propose the MCC algorithm.

Firstly, we assume the SBSs store all the content and compute the value of $\underline{P}$ and the corresponding transmission decision. Next, based on the obtained results of t_{mk}, we set $x_{mn} = \max\{t_{mk}, \forall k \in \mathcal{K}_n'\}$ preliminarily. However, the limit of cache capacity in some SBSs may be broken. Then, we set $x_{mn} = 0$ in the SBSs which break the cache capacity successively and record the gap between the power consumption obtained after setting $x_{mn} = 0$ and $\underline{P}$. We try to store the content which makes the power consumption closer to $\underline{P}$.

Algorithm 3 The MCC Algorithm

Require:
S_m;
Ensure:
P^*, x_{mn};
1: Initialize $x_{mn} = 1, \forall m \in \mathcal{M}, n \in \mathcal{N}$, current cache capacity S'_m, $\overline{P}$, $\Phi = \emptyset$ (the set of $x_{mn} = 0, \forall m \in \mathcal{M}, n \in \mathcal{N}$), $\Omega = \emptyset$ (the set of $x_{mn} = 1, \forall m \in \mathcal{M}, n \in \mathcal{N}$);
2: Calculate $\underline{P}$ and t_{mk} by Algorithm 2;
3: **repeat**
4: Based on the temporary results of t_{mk}, determine the cache policy by $x_{mn} = \max\{t_{mk}, \forall k \in \mathcal{K}'_n\}$ and record x_{mn} in the corresponding set of Φ or Ω, $i = 1$.
5: Calculate the size S'_m of Ω for each SBS m.
6: **if** $S'_m > S_m$ **then**
7: Set the value of x_{mn} in Ω to zero and then obtain the corresponding $P' = f(\mathcal{S}^*)$, record $M(n) = P' - \underline{P}$;
8: Set M in ascending order, Choose the first S_m elements in order and set the corresponding $x_{mn} = 1$, otherwise $x_{mn} = 0$;
9: **end if**
10: Base on the temporary results of line $4 - 9$, optimize t_{mk} and P^* using Algorithm 2;
11: $i++$;
12: **until** $i > MaxIteration$
13: **return** P^*, x_{mn}.

VI. SIMULATION RESULTS

In this section, we provide numerical simulations to evaluate the performance of the proposed MCC algorithm. We consider a scenario with a MBS and 20 uniformly distributed SBSs of which the coverage areas may be overlapped. All the SBSs are assumed to have equal cache size. The antenna gain of BSs and users are $g_n = 2.14dBi$ and $g_u = 2.14dBi$, respectively. The distance-dependent pathloss of MBS and SBSs are modeled as $l_{0k} = 128.1 + 37.6log10(d_{0k})$ and $l_{mk} = 140.7 + 36.7log10(d_{mk})$ respectively, where d_{mk} denotes the distance between BS m and user k in kilometers [35].

To mitigate the interference in the considered system, we take into consideration two cases. Firstly, multicast transmission can reduce a lot of redundant transmission compared with unicast transmission. Especially when the system is lightly loaded, each BS only uses a small portion of frequency subchannels for multicasting [36]. Therefore, the BSs are allocated orthogonal subchannels for multicasting. When the system is heavy loaded in multicasting service, the subchannels for multicast transmission will not be fully orthogonalized. Then an adaptive clustering framework for mitigating the inter-cell interference can be used in which a directed interference graph is designed to capture the dominant interference [37]. Together with this interference management strategy, our proposed algorithm can be conducted in the simulation.

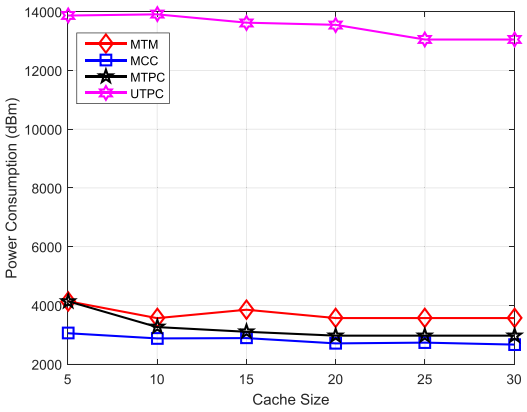


FIGURE 6. Performance comparison of the four algorithms with different cache size, where $K = 100$ and $\alpha = 1.5$.

And in each simulation trial, each user generates content requests independently to a database of $N = 500$ contents. Let $\mathbf{A}(t) \triangleq (A_{m,n}(t))$ be the request arrival process [18], where $A_{m,n}(t)$ denotes the number of requests for content n generated by users in BS m during time slot t. And the arrival process of user request is modeled by the Independent Reference Model (IRM) [17], [18]. In IRM, the probability that next request is independent of the earlier requests. Specially, the arrival of user requests is not always simultaneous, but we only focus on the multicast scheduling for the requests arriving in the same time slot. In addition, the content popularity follows the Zipf distribution. And in the simulation, to get the average value of each point, the analysis is repeated for 100 times in which both the locations of SBSs and users remain unchanged.

To demonstrate the power-saving performance of the proposed algorithm, we consider the following four strategies for comparison:

- Unicast Transmission and Popularity-based Caching (UTPC): The scheme is currently used in many caching system. Each SBS caches the locally most popular contents independently from the other SBSs. And each user is served by a unicast transmission of its associated SBS only when it stores the requested content, otherwise the request can only be satisfied by MBS.
- Multicast Transmission and Popularity-based Caching (MTPC): In this scheme, each SBS stores the locally most popular contents independently and multicast transmission is adopted. A MBS multicast will occur when at least one user cannot find the requested content n in the SBS cache, i.e., when a request for the content n is generated only within the coverage area of MBS or when the request generated by a user associated to an SBS which has not stored it, all the requests for the content n will be served by the MBS multicast transmission [22].
- Multicast Transmission of MBS (MTM): All the user requests are served by MBS multicast transmission.
- Multicast and Cooperative Caching (MCC): Our proposed algorithm.

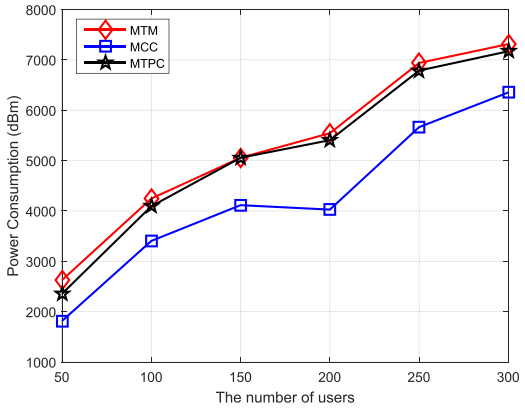


FIGURE 7. Performance comparison of the four algorithms with different user number, where $S_m = 10$ and $\alpha = 1.5$.

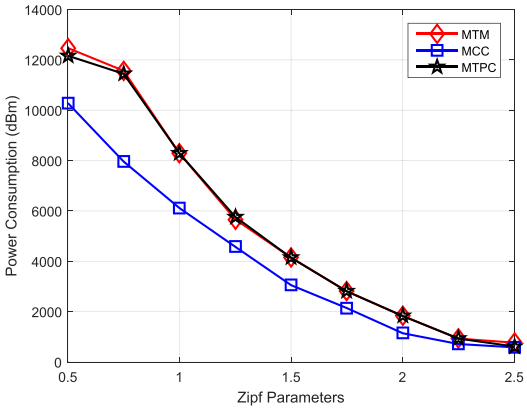


FIGURE 8. Performance comparison of MTM, MCC and MTPC algorithms with different Zipf parameters, where $S_m = 5$ and the user number is 100.

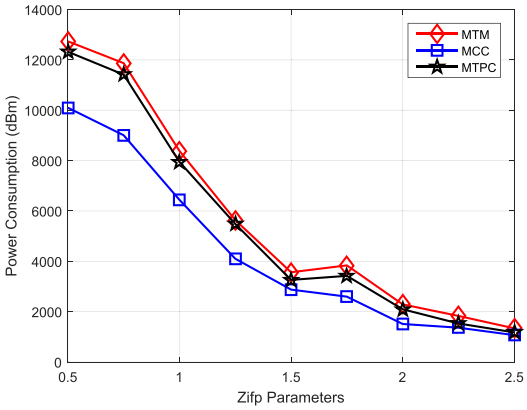


FIGURE 9. Performance comparison of MTM, MCC and MTPC algorithms with different Zipf parameters, where $S_m = 10$ and the user number is 100.

Fig. 6 illustrates the power consumption versus the cache size for the aforementioned four policies. We can observe that better performance can be achieved by the proposed MCC policy. With the increase of cache size, the power consumption obtained by MCC decreases slightly and that of the MTM and UTPC remain almost steady. In addition, for the MTPC, the performance is greatly improved due to the increase of cache capacity. That is because, the lager cache capacity implies the more transmission opportunity for SBSs. Hence, users can be served at a lower power consumption of SBSs, especially for those with poor channel conditions to MBS. Our proposed algorithm is not very sensitive to the cache size because the cooperative caching enlarges the capacity of each cache to some extent. For UTPC, the gain obtained by the increase of cache size can be negligible due to the large amount of user requests which are required to be served by unicast transmission.

Fig. 7 shows the power consumption of the three algorithms as the number of users increasing from 50 to 300. We can see that the power consumption keeps increasing with the growing number of users in the MTM, MTC and MTPC algorithms. There is almost no difference between the curves of the MTM and MTPC due to the limited cache capacity and a substantial amount of user requests. With limited cache capacity, more MBS multicast transmissions tend to be adopted to serve users which is similar with the MTM. Our proposed algorithm MCC outperforms the other two.

Fig. 8, Fig. 9 and Fig. 10 compare the power consumption among the different multicast and caching policy with different Zipf parameters, where the cache size is 5, 10 and 20 respectively. It can be seen that, in Fig. 8 the performance of MTM and MTPC is almost the same. The reason is that, with the limited cache capacity, the chances for SBSs transmission is slim. Substantial amount of user requests are served by MBS, i.e., the MTM policy. In contrast, there is little influence on the proposed algorithm. In Fig. 10, the performance gap between the MTM and MTPC becomes wider than that in Fig. 9. The MCC algorithm significantly outperforms the other two even with limited cache capacity. On the other hand, the power consumption falls steadily as the value of Zipf parameter increases. In general, α implies the skewness of the content popularity distribution. Larger α indicates that substantial amount of user requests are generated centering on only a few number of contents. Therefore, a multicast transmission can satisfy more user requests due to their concentrative property.

To show the efficiency of the MCC, MTPC and UTPC in balancing the load, we calculated the load balancing based on the Jain's fairness factor which ranges from $\frac{1}{M+1}$ (worst case) to 1 (best case) [38].

Fig. 11 compares the performance of the three algorithms in load balancing with different cache size. Obviously, the MTPC scheme shows a seriously unbalanced load. The reason is that large proportions of users will be associated with the MBS since users requesting the same content may be scattered in the coverage areas of different SBSs and not all the SBSs store the requested content and the requests can only be served by a multicasting transmission of MBS. By contrast, our proposed scheme achieves a better performance in load balancing. The cooperation of neighboring

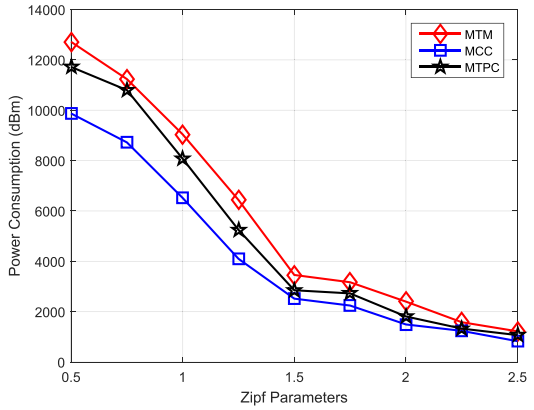


FIGURE 10. Performance comparison of MTM, MCC and MTPC algorithms with different Zipf parameters, where $S_m = 20$ and the user number is 100.

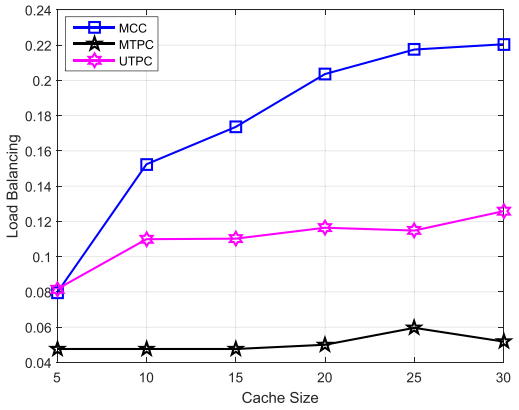


FIGURE 11. Load balancing comparison with different cache size, where $\alpha = 1.5$ and the user number is 100.

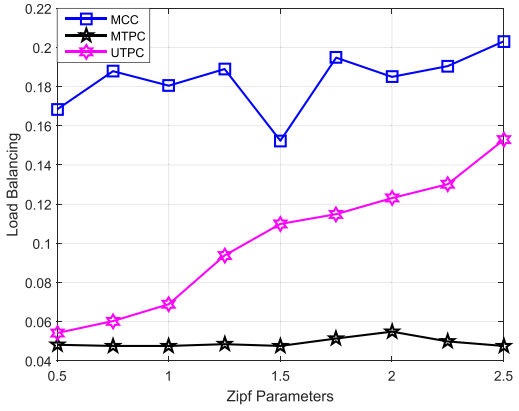


FIGURE 12. Load balancing comparison with different Zipf parameters α, where the number of users is 100 and the cache size is 10.

SBSs relieves the MBS load pressure and transfers congested users to a lightly loaded SBS which will improve the overall network's power consumption. Moreover, as the cache size increases, the curve of MCC goes up because more traffic can be offloaded to the SBSs which store the requested contents in the MCC scheme.

Fig. 12 shows the load balancing of the three algorithms with different Zipf parameters. As can be seen in Fig. 12, our proposed model outperforms the other two schemes significantly. As the zipf parameter increases, the MTPC scheme holds a relatively stable value, which is due to that not all the SBSs store the requested contents in most cases, a large proportion of users can only be served by the multicast transmission of MBS which will lead to overload of MBS. Our proposed algorithm can provide users with more candidate base stations while minimizing the system power consumption. Part of the MBS's traffic load can be shared by SBSs. Therefore, a better load balancing can be achieved by MCC. In addition, with the increasing of zipf parameters, the performance of UTPC is obviously improved. The reason is that a larger value of zipf parameter means the content requests are more centralized and more easily found in local caches. Hence, more requests can be served by a unicast transmission locally rather than served by MBS.

VII. CONCLUSION

In this paper, the design of green network is investigated by taking into consideration the multicast and cooperative caching techniques. We formulate the multicast and cooperative caching problem to minimize the network power consumption. Based on the formulation, we decouple the multicast scheduling for each content given the cache placement. To make the problem more tractable, we decompose it into two phases: user association and transmission determining. And then we transform the problem into the Set Partition Problem. Finally, a distributed multicast and cooperative caching (MCC) algorithm is proposed. Simulation results show that the proposed approach performs better in terms of load balancing and is able to reduce the power consumption up to 28% compared with the existing multicast and caching scheme. Moreover, the algorithm outperforms better than the benchmark algorithms even with limited cache capacity, large user number and different content popularity distributions.

APPENDIX A
PROOF OF LEMMA 1

Obviously, $u_{ik} \geq 0$ because when the power consumption required by the associated BS of group i to serve user k is smaller than $c(\mathcal{G}_i^n)$, we have $c(\mathcal{G}_i^n \cup k) = c(\mathcal{G}_i^n)$, otherwise, $c(\mathcal{G}_i^n \cup k) \geq c(\mathcal{G}_i^n)$. Therefore, there is $u_{ik} \geq 0$.

Assume $u_{jk} > u_{ik}$ but user k is partitioned into $\mathcal{G}_j^n \cup k$ and group $\mathcal{G}_j^n \cup k$ is selected in the optimal solution. Regard the other groups in the optimal solution as a big group which is denoted by $\overline{\mathcal{G}_j^n}$. Then the total power consumption of the optimal solution is $c(\mathcal{G}_j^n \cup k) + c(\overline{\mathcal{G}_j^n})$. Obviously, $\mathcal{G}_j^n$, $\mathcal{G}_i^{n'} = \mathcal{G}_i^n + k$ and $\mathcal{G}^{n'} = \overline{\mathcal{G}_j^n} - \mathcal{G}_i^n$ is also a feasible solution. Its power consumption is $c(\mathcal{G}_j^n) + c(\mathcal{G}_i^{n'}) + c(\mathcal{G}^{n'})$.

We have $u_{jk} > u_{ik}$, i.e., $(max\{c(\mathcal{G}_j^n), c(\mathcal{G}_j^n \cup k)\} - max\{c(\mathcal{G}_j^n)\}) > (max\{c(\mathcal{G}_i^n), c(\mathcal{G}_i^n \cup k)\} - max\{c(\mathcal{G}_i^n)\})$. Therefore, $(max\{c(\mathcal{G}_j^n), c(\mathcal{G}_j^n \cup k)\} + max\{c(\mathcal{G}_i^n)\}) > (max\{c(\mathcal{G}_i^n), c(\mathcal{G}_i^n \cup k)\} + max\{c(\mathcal{G}_j^n)\})$. Based on the above inequality, there is $c(\mathcal{G}_j^n \cup k) + c(\overline{\mathcal{G}_j^n}) = c(\mathcal{G}_j^n \cup k) + c(\mathcal{G}_i^n) + c(\mathcal{G}^{n'}) > c(\mathcal{G}_j^n) + c(\mathcal{G}_i^n \cup k) + c(\mathcal{G}^{n'}) = c(\mathcal{G}_j^n) + c(\mathcal{G}_i^{n'}) + c(\mathcal{G}^{n'})$.

Thus, the optimal solution is $\mathcal{G}_j^n$, $\mathcal{G}_i^{n'}$ and $\mathcal{G}^{n'}$, namely, user k should belong to group $\mathcal{G}_i^n$ which contradicts with the assumption that the optimal solution is reached when user k belongs to group $\mathcal{G}_j^n$.

APPENDIX B
PROOF OF LEMMA 2

Assume $\{\hat{t_{mk}}\}$ and $\hat{P}$ are the set of optimal transmission decision and the corresponding power consumption, respectively. And $\hat{P}$ satisfies $\hat{P} > \overline{P}$. However, when user requests are all served by the multicast transmission of MBS, the power consumption is $\max_{k\in\mathcal{K}_n'} P_{0k} \cdot t_{0k}$. This contradicts the assumption. Therefore, the power consumption of the problem is upper-bounded by $\overline{P}$.

As for the lower bound $\underline{P}$, denote the full cache placement as $\mathbf{x}^f$. Turn certain elements 1 from matrix $\mathbf{x}^f$ into 0 and denote the new matrix as $\mathbf{x}'$ which can be regarded as a new cache placement strategy. Without loss of generality, we assume that $\underline{P} = P_m^f + P_0^f$, which means that users request content n are served partly by the multicast transmission of SBS m and partly by MBS. And there is $P_m^f + P_0^f \leq \overline{P}$. However, if content n is not stored in SBS m in the cache placement $\mathbf{x}'$, the sub-optimal solution of Algorithm 2 is $\max_{k\in\mathcal{K}_n'} P_{0k} \cdot t_{0k}$, i.e., $\overline{P}$. Therefore, the lower bound is $\underline{P}$.

REFERENCES

[1] X. Chen, L. Jiao, W. Li, and X. Fu, "Efficient multi-user computation offloading for mobile-edge cloud computing," *IEEE/ACM Trans. Netw.*, vol. 24, no. 5, pp. 2795–2808, Oct. 2016.

[2] S. Andreev *et al.*, "Exploring synergy between communications, caching, and computing in 5G-grade deployments," *IEEE Commun. Mag.*, vol. 54, no. 8, pp. 60–69, Aug. 2016.

[3] S. Samarakoon, M. Bennis, W. Saad, M. Debbah, and M. Latva-Aho, "Ultra dense small cell networks: Turning density into energy efficiency," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 5, pp. 1267–1280, May 2016.

[4] K. Poularakis and L. Tassiulas, "Code, cache and deliver on the move: A novel caching paradigm in hyper-dense small-cell networks," *IEEE Trans. Mobile Comput.*, vol. 16, no. 3, pp. 675–687, Mar. 2017.

[5] X. Wang, M. Chen, T. Taleb, A. Ksentini, and V. C. M. Leung, "Cache in the air: Exploiting content caching and delivery techniques for 5G systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 131–139, Feb. 2014.

[6] A. F. Molisch, G. Caire, D. Ott, J. R. Foerster, D. Bethanabhotla, and M. Ji, "Caching eliminates the wireless bottleneck in video aware wireless networks," *Adv. Electr. Eng.*, vol. 2014, Nov. 2014, Art. no. 261390.

[7] L. Chen and G. Feng, "Caching policy for reliable multicast in Ad Hoc networks," in *Proc. Int. Conf. Commun., Circuits Syst. (ICCCAS)*, vol. 1. Nov. 2013, pp. 104–108.

[8] J. Park, T. Kim, W. Lee, D. Byun, and Y. Bae, "Cache aided matrix network coding based multicast scheme over wireless networks," in *Proc. IEEE Int. Conf. Consum. Electron. (ICCE)*, Jan. 2015, pp. 287–288.

[9] Y. Cao, T. Jiang, X. Chen, and J. Zhang, "Social-aware video multicast based on device-to-device communications," *IEEE Trans. Mobile Comput.*, vol. 15, no. 6, pp. 1528–1539, Jun. 2016.

[10] N. Wang, E. Hossain, and V. K. Bhargava, "Backhauling 5G small cells: A radio resource management perspective," *IEEE Wireless Commun.*, vol. 22, no. 5, pp. 41–49, Oct. 2015.

[11] Y.-H. Chiang and W. Liao, "ENCORE: An energy-aware multicell cooperation in heterogeneous networks with content caching," in *Proc. 35th Annu. IEEE Int. Conf. Comput. Commun.*, Apr. 2016, pp. 1–9.

[12] Y. Wang, X. Tao, X. Zhang, and G. Mao, "Joint caching placement and user association for minimizing user download delay," *IEEE Access*, vol. 4, pp. 8625–8633, 2016.

[13] Y. Cui, F. Lai, S. Hanly, and P. Whiting, "Optimal caching and user association in cache-enabled heterogeneous wireless networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2016, pp. 1–6.

[14] W. Wen, Y. Cui, F.-C. Zheng, S. Jin, and Y. Jiang, "Random caching based cooperative transmission in heterogeneous wireless networks," in *Proc. IEEE Int. Conf. Commun.*, May 2017, pp. 1–6.

[15] S. H. Chae, T. Q. Quek, and W. Choi, "Content placement for wireless cooperative caching helpers: A tradeoff between cooperative gain and content diversity gain," *IEEE Trans. Wireless Commun.*, vol. 16, no. 10, pp. 6795–6807, Oct. 2017.

[16] Y. Bao, X. Wang, S. Zhou, and Z. Niu, "An energy-efficient client pre-caching scheme with wireless multicast for video-on-demand services," in *Proc. 18th Asia–Pacific Conf. Commun. (APCC)*, Oct. 2012, pp. 566–571.

[17] B. Zhou, Y. Cui, and M. Tao, "Optimal dynamic multicast scheduling for cache-enabled content-centric wireless networks," *IEEE Trans. Commun.*, vol. 65, no. 7, pp. 2956–2970, Jul. 2017.

[18] B. Zhou, Y. Cui, and M. Tao, "Stochastic content-centric multicast scheduling for cache-enabled heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 6284–6297, Sep. 2016.

[19] X. Zhang, H. Gao, and T. Lv, "Multicast beamforming for scalable videos in cache-enabled heterogeneous networks," in *Proc. IEEE Wireless Commun. Netw. Conf.*, Mar. 2017, pp. 1–6.

[20] Y. Cui, D. Jiang, and Y. Wu, "Analysis and optimization of caching and multicasting in large-scale cache-enabled wireless networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 7, pp. 5101–5112, Jul. 2016.

[21] Y. Cui and J. Dongdong, "Analysis and optimization of caching and multicasting in large-scale cache-enabled heterogeneous wireless networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 1, pp. 250–264, Jan. 2017.

[22] 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.

[23] H. Feng, Z. Chen, and H. Liu, "Design and optimization for VoD services with adaptive multicast and client caching," *IEEE Commun. Lett.*, vol. 21, no. 7, pp. 1621–1624, Jul. 2017.

[24] J. Liao, K.-K. Wong, Y. Zhang, Z. Zheng, and K. Yang, "Coding, multicast and cooperation for cache-enabled heterogeneous small cell networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 10, pp. 6838–6853, Oct. 2017.

[25] X. Huang, Z. Zhao, and H. Zhang, "Latency analysis of cooperative caching with multicast for 5g wireless networks," in *Proc. IEEE/ACM 9th Int. Conf. Utility Cloud Comput. (UCC)*, Dec. 2016, pp. 316–320.

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

[27] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and Zipf-like distributions: Evidence and implications," in *Proc. 18th Annu. Joint Conf. IEEE Comput. Commun. Soc. (INFOCOM)*, vol. 1. Mar. 1999, pp. 126–134.

[28] G. Koutitas *et al.*, "Greening the airwaves with collaborating mobile network operators," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp. 794–806, Jan. 2016.

[29] C. H. Papadimitriou, "On the complexity of integer programming," *J. ACM*, vol. 28, no. 4, pp. 765–768, 1981.

[30] R. N. Tomastik and P. B. Luh, "The facet ascending algorithm for integer programming problems," in *Proc. 32nd IEEE Conf. Decision Control*, Dec. 1993, pp. 2880–2884.

[31] D. Levine, "A parallel genetic algorithm for the set partitioning problem," Office Sci. Tech. Inf. (OSTI), Washington, DC, USA, Tech. Rep., 1994. [Online]. Available: https://www.osti.gov/scitech/biblio/10161119

[32] Q. Zhang and Y.-W. Leung, "An orthogonal genetic algorithm for multimedia multicast routing," *IEEE Trans. Evol. Comput.*, vol. 3, no. 1, pp. 53–62, Apr. 1999.

[33] D. E. Goldberg and K. Deb, "A comparative analysis of selection schemes used in genetic algorithms," *Found. Genetic Algorithms*, vol. 1, no. 1, pp. 69–93, 1991.

[34] P. C. Chu and J. E. Beasley, "Constraint handling in genetic algorithms: The set partitioning problem," *J. Heuristics*, vol. 4, no. 4, pp. 323–357, 1998.

[35] J. Ikuno, M. Wrulich, and M. Rupp, *Evolved Universal Terrestrial Radio Access (E-UTRA); Further Advancements for E-UTRA Physical Layer Aspects*, document 3GPP TR 36.814 V9. 0.0, 2010.

[36] *Introduction of the Multimedia Broadcast/Multicast Service (MBMS) in the Radio Access Network (RAN)*, document 3GPP TS 25.346 V14.0.0, 2017.

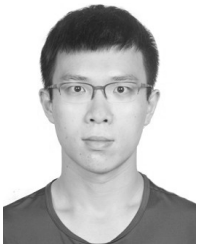
[37] X. Lyu, H. Tian, W. Ni, R. P. Liu, and P. Zhang, "Adaptive centralized clustering framework for software-defined ultra-dense wireless networks," *IEEE Trans. Veh. Technol.*, vol. 66, no. 9, pp. 8553–8557, Sep. 2017.

[38] J. B. Abderrazak, A. Zemzem, and H. Besbes, "QoS-driven user association for load balancing and interference management in HetNets," in *Proc. 6th Int. Conf. Netw. Future*, 2015, pp. 1–3.

SHUO HE received the B.S. degree in communications engineering from the Beijing University of Posts and Telecommunications, Beijing, China, in 2015, where she is currently pursuing the Ph.D. degree in information and communications engineering. Her research interests include wireless caching, popularity prediction, and mobile edge computing.

HUI TIAN (M'03) received the M.S. degree in microelectronics in 1992 and the Ph.D. degree in circuits and systems from the Beijing University of Posts and Telecommunications (BUPT) in 2003. She is a Professor with BUPT and the Director of the State Key Laboratory of Networking and Switching Technology. She is a Committee Member of the Beijing Key Laboratory of Wireless Communication Testing Technology, a Core Member of the Innovation Group, National Natural Science Foundation of China, a member of the China Institute of Communications and an Expert with the Unified Tolling and Electronic Toll Collection Working Group, China National Technical Committee on ITS Standardization. Her research interests include LTE and 5G system design, MAC protocols, resource scheduling, cross-layer design, cooperative relaying in cellular systems, and ad hoc and sensor networks.

She was a co-recipient of the National Award for Technological Invention, the Science and Technology Award of China Communications and the Ten major Scientific and Technological Progresses Award of China's colleges and universities for her contribution in the field of wireless communication. She has been a TPC Member of the IEEE conferences (GlobalCom, WCNC, WPMC, PIMRC, VTC, and ICCC) and a reviewer of the IEEE TVT, the IEEE CL, the IET Communications, the *Transactions on Emerging Telecommunications Technologies*, the *Journal on Wireless Communications and Networking (EURASIP)*, the *Journal of Networks*, the *Journal of Electrical Engineering (Majlesi)*, the *Journal of China University of Posts and Telecommunications*, the *Chinese Journal of Electronics*, the *Journal of Electronics and Information Technology*, and the *Chinese Journal of Aeronautics*. She was a Lead Guest Editor of the *Journal on Wireless Communications and Networking (EURASIP)*.



XINCHEN LYU received the B.S. degree in communications engineering from the Beijing University of Posts and Telecommunications, Beijing, China, in 2014, where he is currently pursuing the Ph.D. degree. His research interests include mobile cloud computing and radio management.

GAOFENG NIE received the B.S. and Ph.D. degrees from the Beijing University of Posts and Telecommunications (BUPT) in 2010 and 2016, respectively. He is currently a Lecturer with BUPT. His research interests are radio resource management in ultra-dense networks, and key technologies in 5G wireless networks.

SHAOSHUAI FAN received the M.S. degree in electrical engineering from Jilin University in 2009, and the Ph.D. degree in electrical engineering from the Beijing University of Posts and Telecommunications (BUPT) in 2015. He was a Post-Doctoral Researcher with BUPT from 2015 to 2017. He is currently a Lecturer with BUPT, China. His research interests include wireless networking and resource management for 5G networks.