Power Consumption Minimization in Cache-Enabled Mobile Networks

Fang Dong, Tianyu Wang, *Member, IEEE*, and Shaowei Wang, *Senior Member, IEEE*

***Abstract*—Caching at base stations (BSs) can enhance the performance of a mobile network, which has gained much attention in the past few years. In this paper, we investigate the energy-saving issue in the cache-enabled mobile network, where we try to minimize the total system power consumption with limited radio resource and storage capacity of the BSs. We decouple the formulated difficult optimization task into a series of tractable subproblems and develop efficient algorithms to solve them with reasonable computation load. The key idea of our proposed scheme is to associate as many as possible users with the BSs that have stored the users' requested files in their caches while considering the spectrum and power budgets of the BSs. Numerical results show our proposed scheme can significantly reduce the system power consumption compared with others.**

***Index Terms*—Approximation algorithm, cache, mobile networks, user association.**

I. Introduction

THE throughput of mobile network has been growing steadily with the introduction of novel network architecture and advanced signal processing techniques. In the past decade, cloud radio access network [1]–[3] and heterogeneous network [4]–[6] are proposed as efficient ways to offer high throughput by densely deploying radio access points such as remote radio heads and small cells in the service area to improve area spectrum efficiency; recently, massive MIMO [4] and mmWave [7] are also introduced to enhance the system capacity of future mobile networks. However, global mobile data traffic has experienced an explosive increase throughout the world with the rapid proliferation of end users. The trend is likely to continue according to Cisco's investigation which claims that global mobile traffic is expected to rise about 40 times in next five years [8]. As a result, mobile networks always face heavy transmission burden, especially for the backhauls connecting the base stations (BSs) with the core network.

Consider content delivery scenario such as video streaming and gaming, though tremendous items requested by users include the same content, all these requests have to be handled by the core network via the backhauls in current mobile communication systems. Massive data are transported via the backhauls repeatedly [9], resulting in tremendous waste of the network resource including the backhauls and the computing resources of the core network. Moreover, the quality of experience of users could be deteriorated due to the unavoidable latency. However, if these repeatedly requests of users are handled at the BSs, the burden of backhauls could be alleviated consequently, as well as the load of the core network. Moreover, the perceived latency could also be reduced in this way. Caching at the BSs has been put forward as a meaningful paradigm to meet this goal in 5G mobile networks and beyond [10], where popular items are put at the edge of mobile networks [11] to reduce the load of backhauls, the transmission latency and the system energy consumption.

The capacity limitation of the backhauls has been the bottleneck of mobile network. Deploying caches at the BSs is deemed as a promising scheme to address this challenge since the cost of disk storage has sharply reduced in the past decades. In [12], a new architecture is presented to cope with the explosive requirements of video content in wireless networks. It indicates that the distributed caches at the BSs can significantly improve the throughput. It is also shown that the capacity of backhauls can be saved up to 22% by exploiting proactive caching scheme [13], and higher throughput gain can be achieved further by increasing the storage capacity at the edge of network. In [14], a new-type framework is proposed to jointly optimize caching and computing, and a distributed algorithm is developed to address the formulated optimization task. In [15], a cost-effective cache deployment is proposed to maximize the system capacity while satisfying the transmission rate requirements. A user prefix caching scheme is proposed to reduce the average playback delay in [16], where each user can cache a part of video clips in advance. In [17], a posterior caching scheme for a cooperative network is discussed, where the cache placement is carried out firstly. In [18], a column generation method is developed to maximize the system throughput by jointly considering caching, routing and channel assignment. The cache placement problem in content distribution network is discussed in [19], where two distributed algorithms are presented to minimize the storage and

Manuscript received March 12, 2019; revised April 23, 2019; accepted April 25, 2019. Date of publication April 30, 2019; date of current version July 16, 2019. This work was supported in part by the National Natural Science Foundation of China under Grants 61671233 and 61801208, in part by the Jiangsu Science Foundation under Grant BK20170650, in part by the Postdoctoral Science Foundation of China under Grants BX201700118 and 2017M621712, in part by the Jiangsu Postdoctoral Science Foundation under Grant 1701118B, and in part by the Open Research Fund of National Mobile Communications Research Laboratory under Grant 2019D02. The review of this paper was coordinated by Prof. A. Jamalipour. *(Fang Dong and Tianyu Wang are co-first authors.) (Corresponding author: Shaowei Wang.)*

F. Dong and S. Wang are with the School of Electronic Science and Engineering, Nanjing University, Nanjing 210023, China (e-mail: mf1623009@smail.nju.edu.cn; wangsw@nju.edu.cn).

T. Wang is with the School of Electronic Science and Engineering, Nanjing University, Nanjing 210023, China, and also with the National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China (e-mail: tianyu.alex.wang@nju.edu.cn).

Digital Object Identifier 10.1109/TVT.2019.2914023

the access cost. Storage and latency in the cache-aided network measured by normalized delivery time are investigated in [20], and it is shown that the traffic load and the transmission rate can benefit from the proposed scheme simultaneously. In [21], caching and energy issues are jointly considered in the small cell networks, and an effective power control method is adopted to cut down the energy consumption.

In brief, caching popular items at the edge of the mobile networks can reduce the duplicated transmission and enhance network performance such as system throughput and energy saving. In this paper, we concentrate on the system power consumption issue facing the cache-enabled mobile network, which is also an important problem and is concerned recently. In [22], an effective system power consumption model is proposed with the analysis of the cache influence on energy efficiency. A user clustering method is proposed to improve the cache hit probability in [23], where the authors try to improve the system energy efficiency. A global energy consumption model is proposed in [24], where the maximum-distance separable code is introduced to maximize energy efficiency. The energy consumption in the content-centric network is studied in [25], where the average response to data transmission is largely determined by energy consumption. In view of the bandwidth issue of BSs, a user association and content caching strategy is developed to maximize the number of content requirements at the local BSs in [26]. In [27], both cache and request routing issues are investigated with the analysis of routing and caching complexity. Device to device link scheduling and power allocation are jointly considered in [28]. In [29], a power allocation method is proposed for a cache-aided network while considering the capacity limitation of backhauls and the traffic demand of users. In [30], cache and user association are jointly investigated to minimize the total delay of users. The channel quality of backhauls is investigated in [31], where cache policy and user association are designed to minimize the system download delay.

However, the aforementioned caching and user association strategies rarely involve the radio resource allocation issue in mobile networks, such as spectrum and power distribution among users. Recall that the radio resource budget of the BSs imposes strict constraints on the admission of users [32], [33] and is also inextricably bound up with caching strategies in mobile networks. As far as the authors have known, the storage capacity of caches and radio resource budgets of BSs have not been jointly taken into account thoroughly in the literature. Under practical scenarios, although a user can find the content item he/she requests from a BS, he/she may not obtain the content files from the BS due to its limited radio resources. That is to say, if many users have been served by a given BS, no radio resources can be assigned to a new coming user for this BS even though the BS has cached the requested files by the coming users. In this paper, we take the spectrum and power budgets of the BSs into consideration and investigate how to associate users with the BSs equipped with capacity limited caches to reduce the system energy consumption. The motivation of our proposal is as follows: Consider a cache-enabled mobile network where each BS can cache limited content files that would be requested by users, if we associate users with the BSs that can provide the requested content files locally, the requests of the users would be met without involving the backhauls between the BSs and the core network. Then the burden of the core network is alleviated, as well as the backhauls. The contributions of this paper are summarized as follows:

- We propose a generalized radio access scheme for cache-enabled mobile networks, where we jointly consider the spectrum and power budgets of BSs and the storage capacity of caches. Our proposed scheme provides feasible solutions to practical mobile networks with cache enhancement.
- We formulate a pragmatic power minimization model that includes the power consumption of BS maintenance, radio transmission and backhauls. Our problem formulation sheds insights into the performance improvement for the cache-enabled networks and can be extended straightforwardly to other scenarios.
- We develop an efficient approximation algorithm to address the formulated user association problem, which yields performance guaranteed solutions with reasonable complexity. Numerical results show significant advantages compared with other methods.

The reminder of this paper is as follows. We present system model and formulate our optimization task in Section II. In Section III, we describe the algorithms in detail. Numerical results and discussions are presented in Section IV. Finally, we conclude this paper in Section V.

TABLE I
NOTATIONS

Symbol	Semantics
$a_{k,n}$	Index of cache missed or not
$b_{k,n}$	Bandwidth allocated by base station n to user k
b_n^{max}	Total bandwidth of BS n
$\mathcal{F}$	Set of files
F	Number of files
$h_{k,n}$	Power gain between BS n and user k
$I_{k,n}$	Maximum interference from adjacent BSs in unit bandwidth
$\mathcal{K}$	Set of users
K	Number of users
M	Storage capacity of BSs
N	Number of BSs
N_0	PSD of AWGN
$\mathcal{N}$	Set of BSs
$p_{k,n}$	Power of BS n allocated to user k
p_n^{max}	Maximum transmission power of BS n
P_m	Power consumption for maintaining at BS n
$P_{n,bh}$	Power consumption for backhauling at BS n
$P_{n,t}$	Power consumption for transmitting at BS n
$r_{k,n}$	Available transmission rate from BS n to user k
R_k^{min}	Minimal rate requirement of user k
$\rho_{k,n}$	User association index between user k and BS n

II. System Model and Problem Formulation

Frequently used notations are listed in Table I. Suppose a cache-enabled mobile network illustrated in Fig. 1, where N BSs are used to serve K users. The BSs can connect to the core

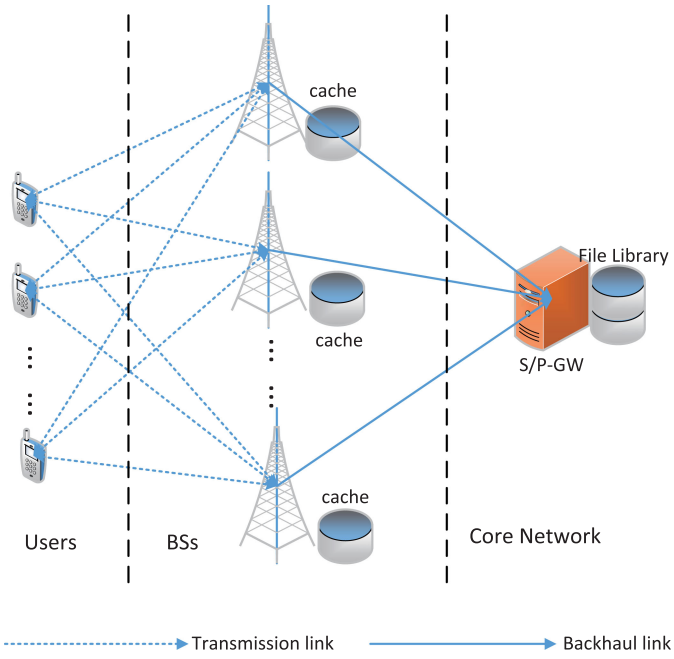

Fig. 1. System model.

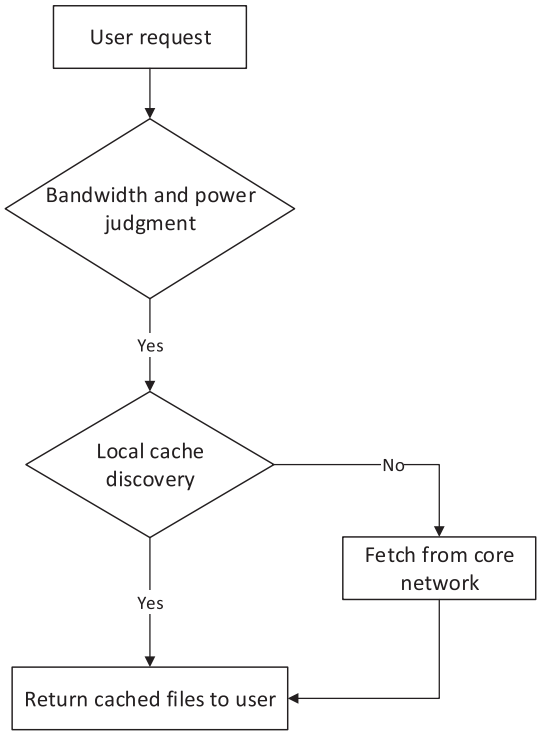

Fig. 2. Service of a user request by our strategy.

network through backhauls and have certain storage space. S/P-GW represents the service gateway and the packet data network gateway. The users and the BSs are randomly located in the service area. The set of users is expressed by $\mathcal{K} = \{1, 2, \ldots, K\}$ and the set of BSs represents as $\mathcal{N} = \{1, 2, \ldots, N\}$. There are $\mathcal{F} = \{1, 2, \ldots, F\}$ files with equal size that have a certain possibility of being requested by users. Recall that it is justifiable to assume each file has equal size since each file can be split into several small blocks. Each BS has limited storage capacity so only a certain number of files are possible to be cached at the BSs. We denote the limited storage space as M, and $M < F$. Each BS independently stores M files from library $\mathcal{F}$ and a file can be cached by multiple BSs. All requested files can always be found in the core network. So a user associated with a BS can obtain a file from the core network via the backhaul connecting the BS to the core network, or just from the cache of the BS on condition that the requested content is stored at the BS.

The maximum transmission power and total available bandwidth of BS n are $p_n^{\max}$ and $b_n^{\max}$, respectively. Let $b_{k,n}$ and $p_{k,n}$ be the bandwidth and the power that BS n allocated to user k, respectively. The interference from the adjacent BSs in unit bandwidth can be expressed by $I_{k,n}$:

$$I_{k,n} = \sum_{n^* \in \mathcal{N}, n^* \neq n} \frac{p_{n^*}^{\max} h_{k,n^*}}{b_{n^*}^{\max}}. \tag{1}$$

Therefore, the achievable transmission rate of user k served by BS n can be written as:

$$r_{k,n} = b_{k,n} \log_2 \left[1 + \frac{p_{k,n} h_{k,n}}{b_{k,n}(N_0 + I_{k,n})} \right]. \tag{2}$$

A user can get the requested files from BS n on the condition that the BS has enough radio resource to support the required transmission rate of the user. Fig. 2 shows the behavior of our strategy of dealing with user request. From another perspective, the serving BS has to ask for the requested content files from the core network if they cannot be found in the cache of the BS. Recall that the access cost that a user obtains his/her requested files from the cache at the BS is significantly different from that of the core network from the viewpoint of power consumption.

Let $a_{k,n}$ denote whether the file required by user k can be found in the storage space of BS n or not:

$$a_{k,n} = \begin{cases} 1 & \text{File requested by user } k \text{ not cached at BS } n; \\ 0 & \text{Otherwise.} \end{cases} \tag{3}$$

Binary decision variable $\rho_{k,n}$ denotes if user k is served by BS n:

$$\rho_{k,n} = \begin{cases} 1 & \text{User } k \text{ served by BS } n; \\ 0 & \text{Otherwise.} \end{cases} \tag{4}$$

We investigate the power consumption in the cache-assisted mobile network. To this end, we should model the overall system power consumption, including the maintenance power of the BSs, the transmission power of radio links and the backhauls power consumption. Let P_m represent the maintenance power consumed by the active BSs, which is a constant and can be reduced only by switching BSs off. $P_{n,t}$ denotes the power consumption at BS n for data transmission. $P_{n,bh}$ occurs when the cache cannot be found locally, which represents the energy for BS n to fetch contents from the core network via the backhauls. According to [34], the power consumption of the backhauls for

BS n can be expressed as:

$$P_{n,bh} = \frac{P_{bh}^{*} R_{n,bh}}{C_{bh}^{\max}}, \tag{5}$$

where P_{bh}^{*} represents the power consumption of the backhauls when the maximum transmission rate $C_{bh}^{\max}$ is provided, $R_{n,bh}$ is the given backhaul requirement of a base station, i.e., the sum data transmission via the backhaul.

There are two kinds of access cost related to power consumption in considered network: When the requested file by the user can be found in the storage space of the BS serving the user, only are the maintenance P_m and the transmission $P_{n,t}$ needed; when the requirement cannot be found in the storage unit of the BS, it should be obtained from the core network, thus increasing the backhaul power consumption $P_{n,bh}$. Consequently, the total power cost at BS n is expressed as follows:

$$P_n = P_m + P_{n,t} + P_{n,bh}. \tag{6}$$

Moreover, the total power consumption P is expressed as:

$$P = \sum_{n\in\mathcal{N}} P_m + \sum_{k\in\mathcal{K}} \sum_{n\in\mathcal{N}} \left(\rho_{k,n} P_{n,t} + \rho_{k,n} a_{k,n} P_{n,bh}\right). \tag{7}$$

We focus on minimizing the total power consumption regarding radio access network under the condition of the bandwidth and power budgets of infrastructures, meanwhile, users's rate requirements are also considered. The optimization task can be formulated as:

$$\begin{aligned}
\min_{\rho_{k,n}, b_{k,n}, p_{k,n}} \quad & P \\
\text{s.t.} \quad & C_1 : \sum_{k\in\mathcal{K}} b_{k,n} \le b_n^{\max}, \forall n \in \mathcal{N}, \\
& C_2 : \sum_{k\in\mathcal{K}} p_{k,n} \le p_n^{\max}, \forall n \in \mathcal{N}, \\
& C_3 : \rho_{k,n} r_{k,n} \ge R_k^{\min}, \forall k \in \mathcal{K}, \\
& C_4 : b_{k,n} \ge 0, p_{k,n} \ge 0, \forall k \in \mathcal{K}, \forall n \in \mathcal{N}, \\
& C_5 : \rho_{k,n} \in \{0,1\}, \forall k \in \mathcal{K}, n \in \mathcal{N}, \\
& C_6 : \sum_{n=1}^{N} \rho_{k,n} \le 1, \forall k \in \mathcal{K}, n \in \mathcal{N}.
\end{aligned} \tag{8}$$

The inequality of C_1 is the bandwidth restrict and C_2 is the power budget of each BS. C_3 denotes users' rate requirements. C_4 is intuitive. C_5 and C_6 indicate that a user can only be serviced by unique BS. Since $\rho_{k,n}$ is binary variable, (8) defines a mixed integer programming problem that is difficult to solve even for medium-sized scale case.

III. Our Proposed Algorithms

Our proposed algorithm consist of three nested procedures. Firstly, we introduce a radio resource allocation algorithm, which efficiently allocates the bandwidth and power resources of a BS to meet the rate requirements of a set of users served by this BS. Secondly, a performance-guaranteed approximation algorithm is developed to tackle the user association problem for given sets of users and BSs, where the radio resource allocation algorithm is called repeatedly. Thirdly, a two-round iterative procedure is presented to minimize the total system power consumption, where the user association algorithm is called repeatedly.

A. Bandwidth and Power Allocation for One BS

For a given set of user $\mathcal{K}_n$ served by BS n, we attempt to find the minimum power consumed by BS n under the condition of satisfying the user's rate demand $R_k^{\min}$. Assume all obtainable bandwidth of BS n which denoted as $b_n^{\max}$ are used to serve $\mathcal{K}_n$. The optimization problem is expressed as follows:

$$\begin{aligned}
\min_{b_{k,n}, p_{k,n}} \quad & \sum_{k\in\mathcal{K}_n} p_{k,n} \\
\text{s.t.} \quad & C_1 : \sum_{k\in\mathcal{K}_n} b_{k,n} = b_n^{\max}, \\
& C_2 : r_{k,n} = R_k^{\min}, \forall k \in \mathcal{K}_n, \\
& C_3 : b_{k,n} \ge 0, p_{k,n} \ge 0, \forall k \in \mathcal{K}_n.
\end{aligned} \tag{9}$$

Use (2) to replace $r_{k,n}$ in (9), we have

$$p_{k,n} = \frac{b_{k,n}}{H_{k,n}} \left(2^{\frac{R_k^{\min}}{b_{k,n}}} - 1\right), \tag{10}$$

where $H_{k,n} = \frac{h_{k,n}}{N_0 + I_{k,n}}$. Thus, the formulation (9) can be converted into a convex form:

$$\begin{aligned}
\min_{b_{k,n}} \quad & \sum_{k\in\mathcal{K}_n} \frac{b_{k,n}}{H_{k,n}} \left(2^{\frac{R_k^{\min}}{b_{k,n}}} - 1\right) \\
\text{s.t.} \quad & C_1 : \sum_{k\in\mathcal{K}_n} b_{k,n} = b_n^{\max}, \\
& C_2 : b_{k,n} \ge 0, \forall k \in \mathcal{K}_n.
\end{aligned} \tag{11}$$

We can get the lagrangian expression of (11) as

$$\begin{aligned}
L = & \sum_{k\in\mathcal{K}_n} \frac{b_{k,n}}{H_{k,n}} \left(2^{\frac{R_k^{\min}}{b_{k,n}}} - 1\right) \\
& + \lambda \left(\sum_{k\in\mathcal{K}_n} b_{k,n} - b_n^{\max}\right) - \sum_{k\in\mathcal{K}_n} \mu_{k,n} b_{k,n}.
\end{aligned} \tag{12}$$

In (12), λ and $\mu_{k,n}$ are the lagrange multipliers. Denote $b_{k,n}^{*}$ and $\lambda^{*}, \mu_{k,n}^{*}$ be primal and dual optimum points with zero duality gap. $b_{k,n}^{*}$, λ^{*} and $\mu_{k,n}^{*}$ can be expressed as follows:

$$\lambda^{*} = -\frac{1}{H_{k,n}} \left[\left(1 - \frac{R_k^{\min} \ln 2}{b_{k,n}^{*}}\right) 2^{\frac{R_k^{\min}}{b_{k,n}^{*}}} - 1\right], \tag{13}$$

$$\sum_{k\in\mathcal{K}_n} b_{k,n}^{*} = b_n^{\max}, \tag{14}$$

$$\mu_{k,n}^{*} = 0, b_{k,n}^{*} > 0. \tag{15}$$

We can obtain $b_{k,n}^{*}$ and λ^{*} by using dichotomy. The allocation algorithm for bandwidth and power is described in Table II (Algorithm 1). ϵ and Γ are a tolerance and an appropriate positive parameter, respectively. Assume $P_n(\mathcal{K}_n) = \sum_{k\in\mathcal{K}_n} p_{k,n}^{*}$ as the

TABLE II
BANDWIDTH AND POWER ALLOCATION

Algorithm 1

Initialization: $Cnt = 0, \lambda^{(Cnt)} = 0, \lambda_{min} = 0, \lambda_{max} = \Gamma$;
repeat
  $Cnt = Cnt + 1$;
  $\lambda^{(Cnt)} = (\lambda_{min} + \lambda_{max})/2$;
  for $k \in \mathcal{K}_n$
    find out $b_{k,n}$ that satisfies (11);
    $b_{k,n}$=max$\{0, b_{k,n}\}$;
  endfor
  if $\sum_{k\in\mathcal{K}_n} b_{k,n} > b_n^{max}$
    $\lambda_{min} = \lambda^{Cnt}$;
  else
    $\lambda_{max} = \lambda^{Cnt}$;
  endif
until $|\lambda^{(Cnt)} - \lambda^{(Cnt-1)}| \le \epsilon$;
for $k \in \mathcal{K}_n$
    $b^*_{k,n} = b_{k,n}$;
    Calculate $p^*_{k,n}$ using (10);
endfor
return $b^*_{k,n}, p^*_{k,n}, \sum_{k\in\mathcal{K}_n} p^*_{k,n}$

optimum value of (11). If $P_n(\mathcal{K}_n)$ is less than the power budget of BS n, we can claim that BS n can serve $\mathcal{K}_n$ users with given rate requirements.

B. User Association

We propose an approximation algorithm to solve the user association problem for given sets of users and BSs. Consider K users with rate requirements and N BSs with limited bandwidth and power budgets, we need to decide which BS each user should be associated with. It is an NP-hard optimization problem since the possible association combinations of K users and N BSs are N^K. The required power $p_n(\{k\})$ for BS n serving user k can be calculated as:

$$p_n(\{k\}) = \frac{b_n^{\max}}{H_{k,n}} \left(2^{\frac{R_k^{\min}}{b_n^{\max}}} - 1 \right). \tag{16}$$

Initialize $\mathcal{K}_n = \varnothing$, $\mathcal{K}_{temp} = \mathcal{K}$ and $\mathcal{N}_{temp} = \mathcal{N}$. Define $\mathcal{K}_{n'}$ as the set of users who have been served. The remaining users and candidate BSs are denoted by $\mathcal{K}_{temp}$ and $\mathcal{N}_{temp}$, respectively. We calculate $p_n(\{k\})$ for $k \in \mathcal{K}_{temp}, n \in \mathcal{N}_{temp}$ in each loop and work out the index (k', n') in the light of the lowest required power $p_{n'}(\{k'\})$, then we calculate $p_{n'}(\mathcal{K}_{n'} \cup \{k'\})$. If $p_{n'}(\mathcal{K}_{n'} \cup \{k'\}) \le p_n^{\max}$, it means that the user k' can be associated with BS n', thus we add k' into K_{temp}, $K_{temp} = K_{temp}\backslash\{k'\}$; else, BS n' would not meet the rate demands of the remaining users because user k' consumes the least power given the same amount of bandwidth. Consequently, BS n' should be removed from $\mathcal{N}_{temp}$. The procedure terminates on condition that all users have been served or no BSs can serve any user. Our proposed user approximation process is elaborated in Table III (Algorithm 2).

Lemma 1: Given two sets of users $\mathcal{K}_1$, $\mathcal{K}_2$, where $|\mathcal{K}_1| = |\mathcal{K}_2|$. For BS n, if $p_n(\{k_1\}) \ge p_n(\{k_2\})$, $\forall k_1 \in \mathcal{K}_1$, $k_2 \in \mathcal{K}_2$, we have $p_n(\mathcal{K}_1) \ge p_n(\mathcal{K}_2)$.

Proof: Detailed proof is shown in Appendix. ■

TABLE III
USER ASSOCIATION PROCEDURE

Algorithm 2

Initialization: $\mathcal{K}_n = \varnothing, \forall n \in \mathcal{N}$; $\mathcal{K}_{temp} = \mathcal{K}$; $\mathcal{N}_{temp} = \mathcal{N}$;
Call **Algorithm 1** to calculate $p_n(\{k\}), k \in \mathcal{K}, n \in \mathcal{N}$;
repeat
  $(k', n') = \arg\min_{(k,n):k\in\mathcal{K}_{temp}, n\in\mathcal{N}_{temp}} p_n(\{k\})$;
  if $p_{n'}(\mathcal{K}_{n'} \cup \{k'\}) \le p_{n'}^{max}$
    $\mathcal{K}_{n'} \leftarrow \mathcal{K}_{n'} \cup \{k'\}$;
    $\mathcal{K}_{temp} \leftarrow \mathcal{K}_{temp}\backslash\{k'\}$;
  else
    $\mathcal{N}_{temp} \leftarrow \mathcal{N}_{temp}\backslash\{n'\}$;
  endif
until $\mathcal{K}_{temp} = \varnothing$ or $\mathcal{N}_{temp} = \varnothing$
return $\mathcal{K}_n$

Corollary 1: Given BS n and two sets of users $\mathcal{K}_1$, $\mathcal{K}_2$. If $p_n(\{k_1\}) \ge p_n(\{k_2\})$, $\forall k_1 \in \mathcal{K}_1$, $k_2 \in \mathcal{K}_2$, and $p_n(\mathcal{K}_1) < p_n(\mathcal{K}_2)$, then $|\mathcal{K}_1| < |\mathcal{K}_2|$.

The proof of corollary is intuitive. If $|\mathcal{K}_1| \ge |\mathcal{K}_2|$, a subset $\mathcal{K}' \subseteq \mathcal{K}_1$ can always be found to meet the condition $p_n(\mathcal{K}_1) \ge p_n(\mathcal{K}') \ge p_n(\mathcal{K}_2)$ according to Lemma 1, where $|\mathcal{K}'| = |\mathcal{K}_2|$.

Theorem 1: Algorithm 2 is a $\frac{1}{2}$-approximation algorithm for user association problem.

Proof: Let $\mathcal{K}^*$ be the set of satisfied users of the optimum solution. For each BS $n \in \mathcal{N}$, denote $\mathcal{K}_n^*$ as the set of users associated with BS n in the optimal solution. $\mathcal{K}'$ represents the set of satisfied users selected by Algorithm 2 and $\mathcal{K}'_n$ is the set of users associated with BS $n \in \mathcal{N}$.

As the process of user association shown in Algorithm 2, BS n provides services to user k consumes lower power compared with any user in $\mathcal{K}_n^*\backslash\mathcal{K}'$. From this, we can draw

$$p_n(\{k_2\}) \ge p_n(\{k_1\}), \forall k_1 \in \mathcal{K}'_n, k_2 \in \mathcal{K}_n^*\backslash\mathcal{K}'. \tag{17}$$

Additionally, we obtain

$$p_n(\mathcal{K}'_n \cup \{k_2\}) > p_n^{\max}, \tag{18}$$

for each user $k_2 \in \mathcal{K}_n^*\backslash\mathcal{K}'$; or user k_2 would be allocated to BS n. For the users in $\mathcal{K}_n^*$ we have

$$p_n^{\max} \ge p_n(\mathcal{K}_n^*\backslash\mathcal{K}'). \tag{19}$$

Combining (18) and (19), we obtain

$$p_n(\mathcal{K}'_n \cup \{k_2\}) > p_n(\mathcal{K}_n^*\backslash\mathcal{K}'). \tag{20}$$

According to (17) and (20) and the corollary of Lemma 1, we obtain

$$|\mathcal{K}_n^*\backslash\mathcal{K}'| < |\mathcal{K}'_n \cup \{k_2\}| = |\mathcal{K}'_n| + 1. \tag{21}$$

That is,

$$|\mathcal{K}'_n| \ge |\mathcal{K}_n^*\backslash\mathcal{K}'|. \tag{22}$$

Then,

$$\begin{aligned} 2 \cdot |\mathcal{K}'| &= |\mathcal{K}'| + \sum_{n\in\mathcal{N}} |\mathcal{K}'_n| \\ &\ge |\mathcal{K}^* \cap \mathcal{K}'| + \sum_{n\in\mathcal{N}'} |\mathcal{K}_n^*\backslash\mathcal{K}'| \\ &= |\mathcal{K}^* \cap \mathcal{K}'| + |\mathcal{K}^*\backslash\mathcal{K}'|. \end{aligned} \tag{23}$$

TABLE IV
POWER CONSUMPTION MINIMIZATION PROCEDURE

Algorithm 3

Initialization: $\mathcal{K}_n = \varnothing, \forall n \in \mathcal{N}$; $\mathcal{K}_{cache} = \mathcal{K}$; $\mathcal{N}_{cache} = \mathcal{N}$;
Calculate $a_{k,n}$ using (3);
First round for cached users ($a_{k,n} = 0$)
repeat
  Call **Algorithm 2** with $\mathcal{K}_{temp} := \mathcal{K}_{cache}$, $\mathcal{N}_{temp} := \mathcal{N}_{cache}$;
until $\mathcal{K}_{cache} = \varnothing$ or $\mathcal{N}_{cache} = \varnothing$
Second round for remaining users
repeat
  Call **Algorithm 2** with $\mathcal{K}_{temp} := \mathcal{K}_{left}$, $\mathcal{N}_{temp} := \mathcal{N}_{left}$;
until $\mathcal{K}_{left} = \varnothing$ or $\mathcal{N}_{left} = \varnothing$

Attention that $\mathcal{K}^* \cap \mathcal{K}'$ represents the selected users in $\mathcal{K}^*$ and $\mathcal{K}^* \backslash \mathcal{K}'$ represents the unselected users in $\mathcal{K}^*$. Therefore $(\mathcal{K}^* \cap \mathcal{K}') \cup (\mathcal{K}^* \backslash \mathcal{K}') = \mathcal{K}^*$, i.e. $|\mathcal{K}^* \cap \mathcal{K}'| + |\mathcal{K}^* \backslash \mathcal{K}'| = |\mathcal{K}^*|$, so we obtain

$$|\mathcal{K}'| \geq \frac{1}{2}|\mathcal{K}^*|. \tag{24}$$

■

C. Power Consumption Minimization

If as many as possible users can get their requested files from the caches, the total power consumption can be decreased as much as possible as indicated in (7) as aforementioned. Our proposed power consumption minimization algorithm is shown in Table IV (Algorithm 3).

Let $\mathcal{K}_{cache}$ be the cached users (which means the number of users which obtain their requested files from the caches of their associated BS) and $\mathcal{N}_{cache}$ is candidate BSs in the first round. Initialize $\mathcal{K}_n = \varnothing, \forall n \in \mathcal{N}; \mathcal{K}_{cache} = \mathcal{K}; \mathcal{N}_{cache} = \mathcal{N}$. Let $\mathcal{K}_{n'}$ assemble the collection of users serviced by BS n' yielded by user association procedure currently. As can be found in Table IV, in the first round, we associate as many as possible users with the BSs that store the files demanded by these users, where we need to solve the user association problem with radio resource budgets. Denote the users not associated with any BS after the first round as $\mathcal{K}_{left}$ and $\mathcal{N}_{left} = \mathcal{N}$. We also employ user association procedure to find out a feasible BS association for these users. Obviously, the users associated with BSs in the second round have to visit the core network via backhauls.

IV. NUMERICAL RESULTS

We estimate the effectiveness of our proposed algorithms with a series of numerical experiments. Suppose that BSs are randomly distributed in the range of 2×2 km^2. The available bandwidth of BS is randomly selected from 20 MHz to 100 MHz and the maximum transmission power of a BS is 1 W. The maintenance power of BS is set to 3.1 W. The maximum backhaul transmission rate $C_{bh}^{\max}$ is 1Gbps with the power dissipation of $P_{bh}^* = 50$ W. The path loss model is based on 3GPP standard and counted as $140.7 + 36.7 \log_{10}(D)$, where D (in km) represents distance between BSs and users. The value of standard deviation of the logarithmic normal shadow and the noise power spectral density are 10 dB and -184 dBm/Hz, respectively. The rate requirements of users range from 2 Mbps to 20 Mbps.

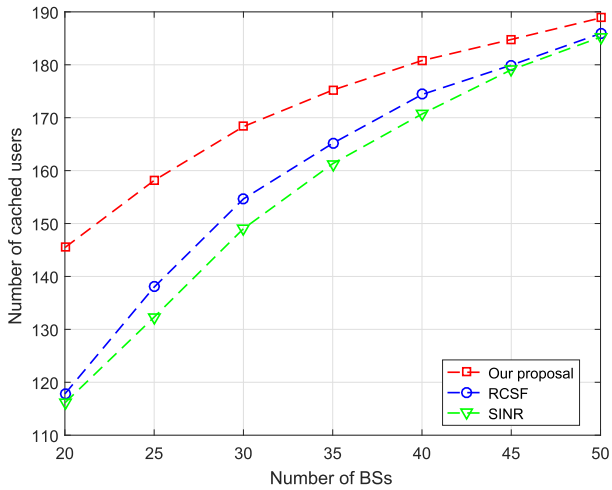

Fig. 3. Number of cached users as a function of the number of BSs.

Our proposed method is compared with other representative methods proposed in the literature. The signal to interference plus noise ratio (SINR) scheme means a user is always connected to the BS that provides the largest SINR. The heuristic algorithm proposed in [35] named rarely cached served first (RCSF), gives the priority to the users whose required files are seldom stored in the local BSs. We evaluate the performances of these algorithms from the following aspects: the number of BSs N, users K, files F and the cache size M. All simulation results are obtained by an average of 500 Monte Carlo tracks.

Figure 3 demonstrates the number of users which find their required contents from the caches of their corresponding BS (cached users) as a function of the number of BSs, where $K = 200, F = 50, M = 10$. As can be seen from Fig. 3, the method we proposed can serve much more users for a given number of BSs as compared to other two methods. Particularly, when the number of BSs is not large enough, which means wireless resources are scarce, our proposed user association algorithm shows significantly performance improvement. The similar performance gain can also be found in Fig. 4, where we can see total power consumption decreases with the increase of the BSs. It can be explained as follows: more BSs are deployed, and more files requested by users can be found in the caches at the BSs. Consequently, the power consumption can be reduced for a given set of users. Again, our proposed algorithm yields the lowest power consumption.

Figure 5 and Figure 6 show the curves of the cached users and the total power consumption as a function of the number of users, where $N = 20, F = 50, M = 10$. We can see from Fig. 5 and Fig. 6 that the cached users and the total power consumption increase when the scale of users becomes larger. When the size of users is relatively small, the performance gap for our algorithm is slight as compared with others. It is because that the radio resource of BS with spectrum and power budgets is not the key factor in this case. In other words, a user is more possible to be served by the BS that caches his/her requested files

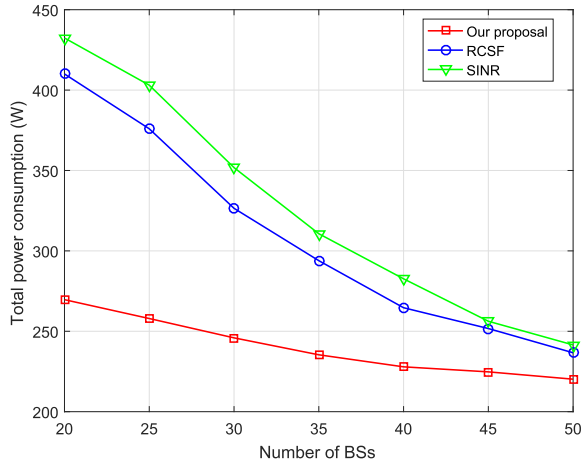

Fig. 4. Power consumption as a function of the number of BSs.

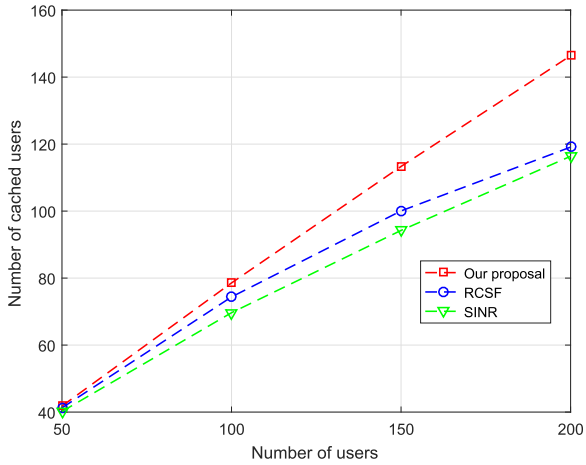

Fig. 5. Number of cached users as a function of the number of users.

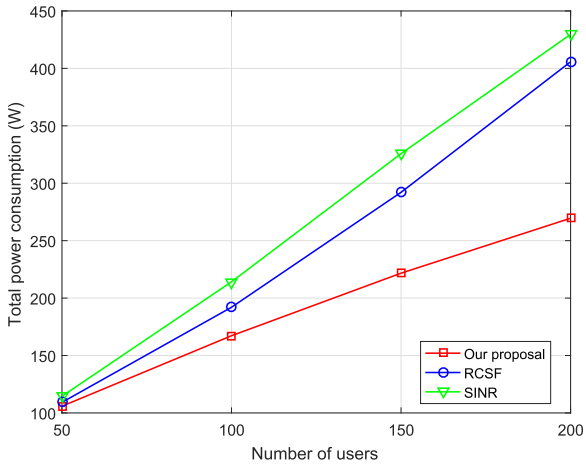

Fig. 6. Power consumption as a function of the number of users.

and also has enough radio resource (bandwidth and power) to provide the user with the required rate. However, as the users's scale becomes larger, radio resource budgets of BSs may impede users associating with the BSs even though these BSs cache the requested files. Our proposed algorithm outperforms the other two policies as can be seen from Fig. 5 and Fig. 6.

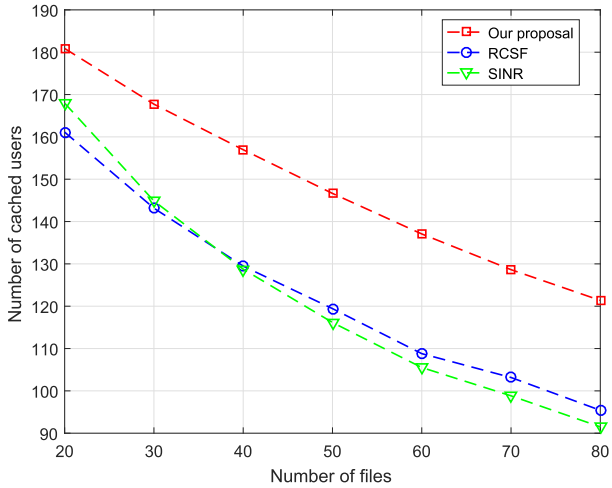

Fig. 7. Number of cached users as a function of the number of files.

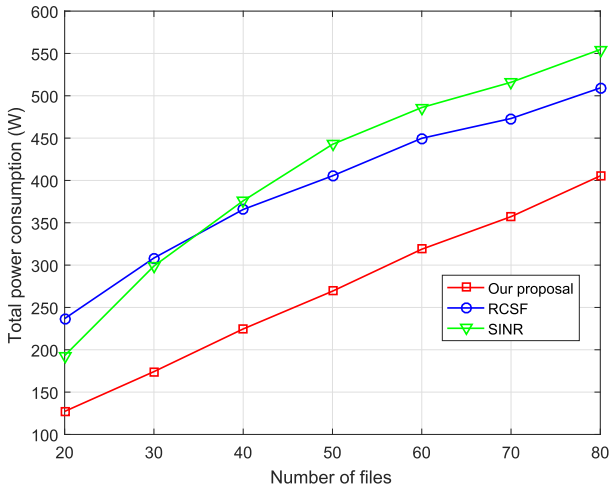

Fig. 8. Power consumption as a function of the number of files.

Figure 7 and Figure 8 illustrate the impact of the number of files on the number of cached users and the total power consumption, where $K = 200, N = 20, M = 10$. We can see from Fig. 7 that the cached users reduce with the diversity of files increase. It can be explained: The hit rate will become smaller if the number of files increases given cache size and BSs, so that more users have to visit the core network to get the requested files. Therefore, the total power consumption has a certain growth as shown in Fig. 8. Our algorithm guarantees that more users served by the local BSs and greatly reduces the power consumption as can be seen from Fig. 7 and Fig. 8.

With the increase of cache size, the number of cached users and the system power consumption can be seen in Fig. 9 and Fig. 10, respectively, where $K = 200, N = 20, F = 50$. Fig. 9 indicates that cached users increase when the cache size becomes large for constant number of files. This is because more files can be stored at the BSs with large cache size so that the hit rate becomes large, and more users can be served by the local BSs. Thus, the reason for the reduction in system power consumption can be explained for the same reason as can be seen from Fig. 10.

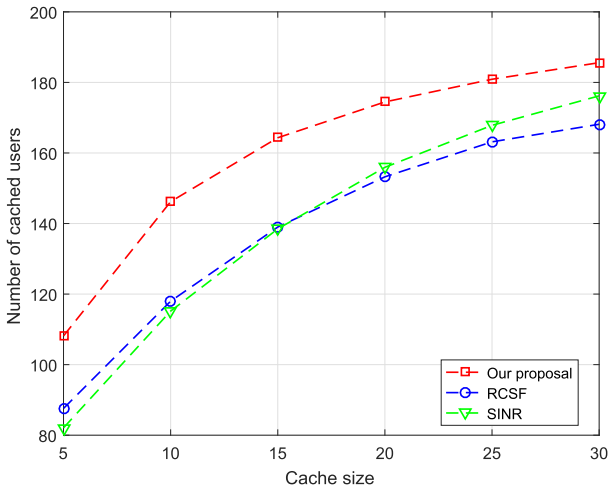

Fig. 9. Number of cached users as a function of the cache size.

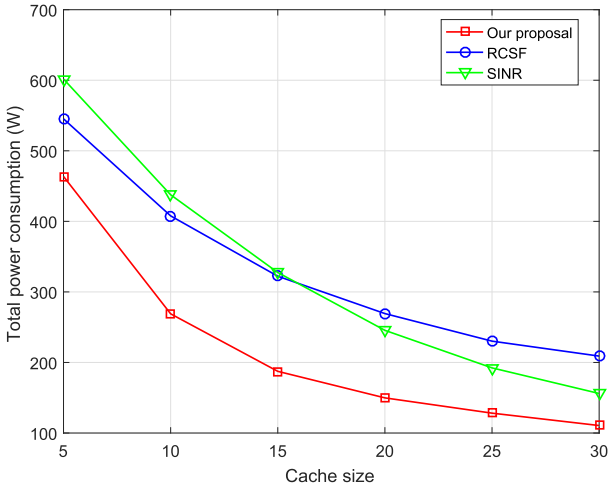

Fig. 10. Power consumption as a function of the cache size.

V. Conclusion

In this paper, we investigated the power consumption minimization problem in the cache-enabled mobile networks, where the radio resource limitations of the base stations and the storage capacity of the caches are considered jointly. We developed a block nested procedure to address the formulated intractable optimization problem efficiently. Numerical experiments show that our proposed scheme can notably reduce the system power consumption. For future work, investigations show that the popularity of files is quite different in practical scenarios [36], [37], it would produce more promising schemes if we jointly take the placement of files and the radio resource allocation at base stations into consideration.

Appendix
Proof of Lemma 1

Fact 1: Given $B > 0, b_1 > 1, b_2 > 1$, assume that $b_1 - 1 \geq B \cdot (b_2 - 1)$, then we have

$$b_1^n - 1 \geq B \cdot (b_2^n - 1), \forall n \geq 1, n \in \mathbf{R}.$$

Denote $p_{k,n}^*$ and $b_{k,n}^*$ be the optimum power and bandwidth allotment for $p_n(\mathcal{K}_1)$, respectively. Based on (16) and $p_n(\{k_1\}) \geq p_n(\{k_2\}), \forall k_1 \in \mathcal{K}_1, k_2 \in \mathcal{K}_2$, we can get

$$\frac{1}{H_{k_1,n}} \cdot \left(2^{R_{k_1}^{\min}/b_n^{\max}} - 1\right) \geq \frac{1}{H_{k_2,n}} \cdot \left(2^{R_{k_2}^{\min}/b_n^{\max}} - 1\right).$$

For each $k_1 \in \mathcal{K}_1$, $k_2 \in \mathcal{K}_2$, we obtain

$$\begin{aligned} p_{k_1,n}^* &= \frac{b_{k_1,n}^*}{H_{k_1,n}} \cdot \left(2^{\frac{R_{k_1}^{\min}}{b_n^{\max}} \cdot \frac{b_n^{\max}}{b_{k_1,n}^*}} - 1\right) \\ &\geq \frac{b_{k_1,n}^*}{H_{k_2,n}} \cdot \left(2^{\frac{R_{k_2}^{\min}}{b_n^{\max}} \cdot \frac{b_n^{\max}}{b_{k_1,n}^*}} - 1\right) \\ &= \frac{b_{k_1,n}^*}{H_{k_2,n}} \cdot \left(2^{R_{k_2}^{\min}/b_{k_1,n}^*} - 1\right), \end{aligned} \tag{25}$$

where the inequality in (25) is always satisfied according to Fact 1. We assign as much as bandwidth to the users in $\mathcal{K}_2$ as the users in $\mathcal{K}_1$. That is $b_{k_1,n}^*$. Thus, we have

$$\begin{aligned} p_n(\mathcal{K}_1) &= \sum_{k_1 \in \mathcal{K}_1} p_{k_1,n}^* \\ &\geq \sum_{k_2 \in \mathcal{K}_2} \frac{b_{k_1,n}^*}{H_{k_2,n}} \cdot \left(2^{R_{k_2}^{\min}/b_{k_1,n}^*} - 1\right) \\ &\geq p_n(\mathcal{K}_2). \end{aligned} \tag{26}$$

Acknowledgment

The authors would like to thank the editors and the anonymous reviewers whose invaluable comments helped substantially improve the presentation of this paper.

References

[1] W. Zhao and S. Wang, "Traffic density-based RRH selection for power saving in C-RAN," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 12, pp. 3157–3167, Dec. 2016.
[2] Q. Shen, Z. Ma, and S. Wang, "Deploying C-RAN in cellular radio networks: An efficient way to meet future traffic demands," *IEEE Trans. Veh. Tech.*, vol. 67, no. 8, pp. 7887–7891, Aug. 2018.
[3] S. Xu and S. Wang, "Baseband unit pool planning for cloud radio access networks: An approximation algorithm," *IEEE Commun. Lett.*, vol. 21, no. 2, pp. 358–361, Feb. 2017.
[4] M. Feng, S. Mao, and T. Jiang, "Joint frame design, resource allocation and user association for massive MIMO heterogeneous networks with wireless backhaul," *IEEE Trans. Wireless Commun.*, vol. 17, no. 3, pp. 1937–1950, Mar. 2018.
[5] S. Wang, W. Zhao, and C. Wang, "Budgeted cell planning for cellular networks with small cells," *IEEE Trans. Veh. Tech.*, vol. 64, no. 10, pp. 4797–4806, Oct. 2015.
[6] W. Zhao, S. Wang, C. Wang, and X. Wu, "Approximation algorithms for cell planning in heterogeneous networks," *IEEE Trans. Veh. Tech.*, vol. 66, no. 2, pp. 1561–1572, Feb. 2017.
[7] M. Feng, S. Mao, and T. Jiang, "Dynamic base station sleep control and RF chain activation for energy-efficient millimeter-wave cellular systems," *IEEE Trans. Veh. Tech.*, vol. 67, no. 10, pp. 9911–9921, Oct. 2018.
[8] "Cisco visual networking index: Global mobile data traffic forecast update," Cisco Syst., Inc., San Jose, CA, USA, White Paper, Feb. 2016.
[9] K. Hamidouche, W. Saad, M. Debbah, J. B. Song, and C. S. Hong, "The 5G cellular backhaul management dilemma: To cache or to serve," *IEEE Trans. Wireless Commun.*, vol. 16, no. 8, pp. 4866–4879, Aug. 2017.
[10] H. Wang, G. Ding, F. Gao, J. Chen, J. Wang, and L. Wang, "Power control in UAV-supported ultra dense networks: Communications, caching, and energy transfer," *IEEE Commun. Mag.*, vol. 56, no. 6, pp. 28–34, Jun. 2018.

[11] Y. Sun, M. Peng, S. Mao, and S. Yan, "Hierarchical radio resource allocation for network slicing in fog radio access networks," *IEEE Trans. Veh. Tech.*, vol. 68, no. 4, pp. 3866–3881, Apr. 2019.
[12] 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.
[13] E. Bastug, M. Bennis, and M. Debbah, "Living on the edge: The role of proactive caching in 5G wireless networks," *IEEE Commun. Mag.*, vol. 52, no. 8, pp. 82–89, Aug. 2014.
[14] Y. Zhou, F. Yu, J. Chen, and Y. Kuo, "Resource allocation for information-centric virtualized heterogeneous networks with in-network caching and mobile edge computing," *IEEE Trans. Veh. Tech.*, vol. 66, no. 12, pp. 11339–11351, Dec. 2017.
[15] S. Zhang, N. Zhang, P. Yang, and X. Shen, "Cost-effective cache deployment in mobile heterogeneous networks," *IEEE Trans. Veh. Tech.*, vol. 66, no. 12, pp. 11264–11276, Dec. 2017.
[16] J. P. Hong and W. Choi, "User prefix caching for average playback delay reduction in wireless video streaming," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp. 377–388, Jan. 2016.
[17] X. Zhao, P. Yuan, H. Li, and S. Tang, "Collaborative edge caching in context-aware device-to-device networks," *IEEE Trans. Veh. Tech.*, vol. 67, no. 10, pp. 9583–9596, Oct. 2018.
[18] A. Khreishah, J. Chakareski, and A. Gharaibeh, "Joint caching, routing, and channel assignment for collaborative small-cell cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 8, pp. 2275–2284, Aug. 2016.
[19] V. Pacifici and G. Dan, "Distributed caching algorithms for interconnected operator CDNs," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 2, pp. 380–391, Feb. 2017.
[20] F. Xu, M. Tao, and K. Liu, "Fundamental tradeoff between storage and latency in cache-aided wireless interference networks," *IEEE Trans. Inf. Theory*, vol. 63, no. 11, pp. 7464–7491, Jun. 2017.
[21] F. Guo, H. Zhang, X. Li, H. Ji, and V. Leung, "Joint optimization of caching and association in energy-harvesting-powered small-cell networks," *IEEE Trans. Veh. Tech.*, vol. 67, no. 7, pp. 6469–6480, Jul. 2018.
[22] D. Liu and C. Yang, "Energy efficiency of downlink networks with caching at base stations," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 907–922, Apr. 2016.
[23] S. E. Hajri and M. Assaad, "Energy efficiency in cache-enabled small cell networks with adaptive user clustering," *IEEE Trans. Wireless Commun.*, vol. 17, no. 2, pp. 955–968, Feb. 2018.
[24] F. Gabry, V. Bioglio, and I. Land, "On energy-efficient edge caching in heterogeneous networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 12, pp. 3288–3298, Dec. 2016.
[25] J. Li, B. Liu, and H. Wu, "Energy-efficient in-network caching for content-centric networking," *IEEE Commun. Lett.*, vol. 17, no. 4, pp. 797–800, Apr. 2013.
[26] K. Poularakis, G. Iosifidis, and L. Tassiulas, "Approximation algorithms for mobile data caching in small cell networks," *IEEE Trans. Commun.*, vol. 62, no. 10, pp. 3665–3677, Oct. 2014.
[27] M. Dehghan *et al.*, "On the complexity of optimal routing and content caching in heterogeneous networks," in *Proc. IEEE INFOCOM*, Hong Kong, China, Apr. 2015.
[28] L. Zhang, M. Xiao, G. Wu, and S. Li, "Efficient scheduling and power allocation for D2D-assisted wireless caching networks," *IEEE Trans. Commun.*, vol. 64, no. 6, pp. 2438–2452, Jun. 2016.
[29] F. Cheng, Y. Yu, Z. Zhao, N. Zhao, Y. Chen, and H. Lin, "Power allocation for cache-aided small-cell networks with limited backhaul," *IEEE Access*, vol. 5, pp. 1272–1283, Mar. 2017.
[30] Y. Cui, F. Lai, S. Hanly, and P. Whiting, "Optimal caching and user association in cache-enabled heterogeneous wireless networks," in *Proc. IEEE GLOBECOM*, Washington, DC, USA, Dec. 2016.
[31] 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, Dec. 2016.
[32] X. Lin and S. Wang, "Joint user association and base station switching on/off for green heterogeneous cellular networks," in *Proc. IEEE Int. Conf. Commun.*, Paris, France, May 2017.
[33] S. Wang and Y. Sun, "Enhancing performance of heterogeneous cloud radio access networks with efficient user association," in *Proc. IEEE Int. Conf. Commun.*, Paris, France, May 2017.
[34] D. Liu and C. Yang, "Energy efficiency of downlink networks with caching at base stations," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 907–922, Apr. 2016.
[35] Y. J. Yu, W. Tsai, and A. Pang, "Backhaul traffic minimization under cache-enabled comp transmissions over 5G cellular systems," in *Proc. IEEE GLOBECOM*, Washington, DC, USA, Dec. 2016.
[36] F. Dong, T. Wang, and S. Wang, "Graph-theoretic approach for cache placement and delay optimization in cache-enabled mobile networks," in *Proc. IEEE Wireless Commun. Signal Process.*, Hangzhou, China, Oct. 2018.
[37] M. Leconte, G. Paschos, L. Gkatzikis, M. Draief, S. Vassilaras, and S. Chouvardas, "Placing dynamic content in caches with small population," in *Proc. IEEE INFOCOM*, San Francisco, CA, USA, Apr. 2016.

Fang Dong received the B.S. degree from Anhui University, Hefei, China, in 2016. She is currently pursuing the M.Sc. degree with the School of Electronic Science and Engineering, Nanjing University, Nanjing, China. Her research interests include resource allocation in cache-enabled mobile networks.

Tianyu Wang (S'11–M'16) received the B.S., M.S., and Ph.D. degrees from Peking University, China. He is currently an Associate Researcher with the School of Electronic Science and Engineering, Nanjing University, Nanjing, China. He has authored or coauthored more than 40 IEEE journal and conference papers, and received the Best Paper Award from the IEEE ICC15, IEEE GLOBECOM14, and ICST ChinaCom12. His current research interest includes network slicing, load balancing, and machine learning in wireless networks.

Shaowei Wang (S'06–M'07–SM'13) received the B.S., M.S., and Ph.D. degrees from Wuhan University, Wuhan, China. He is currently a Full Professor with the School of Electronic Science and Engineering, Nanjing University, Nanjing, China. His research interests mainly include telecommunication systems, operations research, and machine learning. He is on the Editorial Board of IEEE COMMUNICATIONS MAGAZINE, IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, and *Springer Journal of Wireless Networks*. He serves/served on the technical or executive committee of reputable conferences including IEEE INFOCOM, IEEE ICC, IEEE GLOBECOM, and IEEE WCNC.