EURASIP Journal on Wireless Communications and Networking

RESEARCH Open Access

Network utility optimization-based joint user association and content placement in heterogeneous networks

Qianbin Chen[1†], Hong Chen[1], Rong Chai[1*†] and Dongmei Zhao[2]

Abstract

The rapid growth of traffic demands has posed challenges and difficulties on both the radio access networks (RANs) and the backhaul links. While heterogeneous networks (HetNets) are expected to offer diverse radio access capabilities and improve the transmission performance of user equipments (UEs) significantly through integrating various RANs efficiently, the backhaul links may still experience challenges in offering quality of service (QoS) guaranteed services to UEs. To tackle these problems, caching technology, more specifically, caching user contents at the infrastructures of different RANs is proposed as an effective approach. In this paper, we consider the joint user association and cache content placement problem in cache-enabled HetNets. Stressing the tradeoff between user download delay and caching cost, we introduce the concept of utility function which characterizes the joint network performance as the weighted sum of user download delay and the caching cost and formulate the joint user association and cache content placement problem as a network utility optimization problem. As the formulated optimization problem is a nonlinear integer optimization problem which cannot be solved conveniently using traditional optimization tools, we transform the original optimization problem equivalently into three convex subproblems by applying Lagrange partial relaxation and McCormick envelopes, and then propose an iterative algorithm. Within each iteration, for a given set of Lagrange multipliers, the three subproblems are solved respectively by means of the modified Kuhn-Munkres (K-M) algorithm and the locally optimal solutions can be obtained, based on which the Lagrange multipliers can be updated through applying subgradient method. Simulation results demonstrate the effectiveness of the proposed algorithm.

Keywords: Heterogeneous networks, User association, Cache content placement, User download delay, Caching cost

1 Introduction

The rapidly growing requirements for high-speed mobile broadband applications, such as video streaming and online games, have posed great challenges on both radio access networks (RANs) and core networks. By integrating different RANs in an efficient and coordinated manner, heterogeneous networks (HetNets) are expected to improve the transmission performance of user equipments (UEs) significantly [1]. In HetNets, UEs should associate with the infrastructures of the RANs, such as base stations (BSs) of cellular networks or the access points (APs) of wireless local area networks (WLANs) before conducting information transmissions. Different user association or cell selection strategies may result in different network transmission performance and quality of service (QoS) to users due to the heterogeneity of RANs.

While various user association schemes have been proposed for HetNets [2–7] and effectively enhanced the transmission performance of the RANs, the backhaul links may still cause challenges and difficulties in offering QoS guaranteed services to UEs. In particular, the demanding requirements of user services, such as multimedia streaming, web browsing applications, and socially interconnected networks, may cause network congestion and long transmission delay in backhaul links [8]. One promising approach for achieving backhaul offloading and

*Correspondence: chairong@cqupt.edu.cn
†Equal contributors
[1]Key Lab of Mobile Communication Technology, Chongqing University of Posts and Telecommunications, Chongwen Road, Chongqing, People's Republic of China
Full list of author information is available at the end of the article

reducing user download latency is to deploy cache storages at the mobile edge networks, e.g., the BSs or APs of the HetNets [9, 10].

It has been shown in previous research works that the transmission performance of UEs can be enhanced significantly by caching contents at the infrastructures of the RANs [11–16]; hence, designing reasonable cache content placement schemes by taking into account both caching constraints and possible performance enhancement is of particular importance. However, while most of previous research works study user association problem or cache content placement problem independently, these two problems are indeed closely related. For further enhancing the network performance and user QoS, it is highly desirable to jointly design user association and content placement scheme in HetNets, which has been demonstrated in some recent research works [17–19].

Besides, it should be noted that the cost of storing contents at the BSs or APs can be relatively high. More specifically, caching a large number of files at the BSs or APs requires a large memory size, which may be expensive. Furthermore, accessing the requested contents from the caches of BSs or APs also results in content fetching delay, which may vary depending on the accessing BSs or APs.

In this paper, we consider the joint user association and content placement problem in cache-enabled HetNets and define a utility function as the weighted sum of user download delay and caching cost; the joint user association and cache content placement problem is then formulated as a network utility minimization problem. The original optimization problem is then decomposed into three subproblems based on McCormick envelopes and Lagrange partial relaxation, an iterative algorithm is designed to obtain the optimal strategy for joint user association and content placement.

The major contributions of this paper are summarized as follows:

- We study the joint user association and content placement problem of cache-enabled HetNets which consists multiple RANs. To achieve joint resource management and performance enhancement of various access networks, we propose a joint resource management architecture, based on which a joint user association and content placement algorithm is designed.
- The problem of user association or content placement in HetNets has been studied separately in previous works [2–7] and [11–16]. In this paper, we jointly consider the user association and content placement problem in cache-enabled HetNets and design jointly optimal strategies so that the overall performance of the networks can be maximized.
- Stressing the tradeoff between user download delay and caching cost, we characterize the joint network performance as the weighted sum of user download delay and the caching cost via applying the concept of utility function and formulate the joint user association and content placement problem as a network utility minimization problem.
- Since the formulated optimization problem is a nonlinear integer optimization problem which cannot be solved conveniently using traditional optimization tools, we apply Lagrange partial relaxation and McCormick envelopes method and transform the original optimization problem equivalently into three convex subproblems, which can then be solved by a proposed iterative algorithm. Within each iteration, for a given set of Lagrange multipliers, the three subproblems are solved respectively by means of the modified Kuhn-Munkres (K-M) algorithm and the locally optimal solutions can be obtained, based on which the Lagrange multipliers can be updated through applying subgradient method.

The rest of the paper is organized as follows. Section 2 presents an overview of related works. The system model considered in this paper and the proposed joint resource management architecture are described in Section 3. The proposed optimization problem is formulated in Section 4. In Section 5, the solution to the formulated optimization problem is presented. Simulation results are discussed in Section 6. Finally, the conclusions are drawn in Section 7.

2 Related works

In this section, we present an overview of related works, including user association schemes and content placement schemes of HetNets.

2.1 User association schemes of HetNets

In HetNets, UEs with multiple interfaces are allowed to associate with the BSs or APs of different RANs. The design of user association schemes in HetNets is of particular importance which should account for both the wireless channel and interference characteristics between UEs and the infrastructures of RANs as well as the heterogeneity of RANs.

In the past few years, user association or cell selection have been studied for HetNets [2–7]. In [2], the authors address the user association problem in the downlink transmissions of a multi-tier HetNet and propose a unified distributed algorithm, which aims at maximizing the sum utility of long-term rate and minimizing global outage probability at the same time. In [3], the authors examine the latency and reliability issues of fiber-wireless (FiWi)-enhanced LTE-A HetNets and propose a backhaul-aware user association algorithm to achieve inter-cell load balancing and network performance improvement. The authors in [4] investigate the user association and resource allocation problem in HetNets and propose an optimal user association and resource allocation algorithm which minimizes the average packet delay of user traffic across the network.

The authors in [5] study user association and resource allocation problem in multi-antenna HetNets and propose a jointly optimal scheme aiming at maximizing the network utility which is defined as a function of user data rate and associate probability. In [6], the authors consider the joint optimization problem of user association, subchannel allocation, and power allocation for downlink transmission in multi-cell multi-association orthogonal frequency division multiple access (OFDMA) HetNets with the objective of maximizing the weighted sum-rate and propose an alternating optimization method to solve the joint optimization problem. The authors in [7] deal with the problem of user association in HetNets and propose a novel collaborative filtering (CF)-based wireless network recommendation system, which involves social interactions among UEs. A satisfaction game is formulated to deal with this problem, and a utility function is defined to measure a UE's satisfaction.

2.2 Content placement schemes of HetNets

Content placement is fairly critical in reaping the benefit brought by caching. In cache-enabled HetNets, a UE can fetch contents from multiple infrastructures of RANs as the coverage of several BSs and APs overlaps and hence different cache content placement strategies can affect user QoS as well as network transmission performance.

In recent years, content deployment strategies with different optimization objectives in HetNets have been studied [11–16]. The authors in [11, 12] aim to maximize the cache hit probability, defined as the probability that a file requested by the typical user is delivered successfully to the user, to achieve the optimal cache content placement strategies. In [11], the authors address the content placement problem in cache-enabled multi-tier HetNet and the optimal tier-level placement policies are yielded which achieve the maximal hit probability over content placement probabilities. The authors in [12] study the effect of retransmissions on the optimal cache placement policy for both static and mobile user scenarios in cache-enabled small cell networks and determine the optimal caching probability of the files that maximizes the hit probability. In [13], the authors consider a caching system consisting of a video retailer (VR) and a number of network service providers (NSPs) aiming to achieve the maximal network profit. The caching strategy is obtained by modeling the system within the framework of a Stackelberg game and establishing the profit models for both the VR and the NSPs.

The authors in [14] consider the joint content placement and service scheduling problem in femtocell caching networks. To maximize the traffic volume served from the cache, the authors formulate the femto BSs' decision-making process as an Markov decision process (MDP) and develop an efficient online randomized algorithm to achieve the optimal content placement and service scheduling strategy. In [15], the authors investigate the content placement problem in a HetNet where a tier of multi-antenna macro BSs (MBSs) is overlaid with a tier of helpers with caches and propose an optimal content placement strategy which maximizes the successful offloading probability of the MBSs. The authors in [16] study the cache content placement problem in caching enabled small cell networks to minimize the average backhaul load subject to the cache capacity constraints. The optimization problem is formulated and solved to obtain the optimal cache content placement strategy.

Most of the previous works focus on either user association scheme or content placement scheme design in HetNets. Since user association strategy over the network may affect the content placement design significantly, it is highly desirable to jointly design user association and content placement strategy in HetNets in order to further enhance the network performance and user QoS. This has been demonstrated in some recent research works [17–19]. The authors in [17] study the problem of joint user association and caching placement in multicast-aided HetNets and propose a joint cooperative caching and multicast scheduling scheme to minimize the system power consumption. The formulated joint optimization problem is then solved by a distributed algorithm to reduce the signaling and computation complexity. In [18], the authors consider the caching and user association problem in a HetNet with wireless backhaul and formulate the problem as a mixed discrete-continuous optimization to minimize the total time that the HetNet must be active in order to satisfy the average requests for given bandwidth and cache resources. They show that the optimal caching is to store the most popular files at each pico BS and the optimal user association strategy is of a threshold form. The authors in [19] consider the joint design of caching and user association policy in a cache-enabled HetNet. Taking into account the characteristics of wireless channels and backhaul links, the authors propose an average download delay minimization-based joint caching and user association strategy.

Meanwhile, we should note that the cost of caching contents at the BSs or APs can be high, which may affect the user association and content placement decisions. More specifically, the more contents stored in the caches of the RANs, the more cost is required to cache the files, which may weaken the benefits achieved from caching severely. Furthermore, fetching contents from caches of different BSs or APs may result in different content fetching delay. Therefore, different from previous works, the joint user association and cache content placement problem for cache-enabled HetNets is studied in this paper by emphasizing the tradeoff between user download delay and caching cost. We define the network utility of the HetNet as a weighted sum of file downloading delay and caching cost and find the optimum user association and content placement solution to optimize the utility.

3 System model and proposed joint resource management architecture

In this section, we describe the system model considered in this paper and propose a joint resource management architecture.

3.1 System model

We consider the downlink transmission in a HetNet consisting of M_1 cellular BSs, M_2 WLAN APs, and N UEs, where UEs in the network may access a BS or an AP for information interactions. For convenience, we introduce the concept of general AP (GAP) which can be either a BS or an AP and indexed by i, where $i = 1 : M_1$ is a BS and $i = M_1 + 1 : M$ is an AP, where $M = M_1 + M_2$.

We assume that each GAP is connected to the core network through a wired backhaul link; thus, UEs with particular content requirement will be able to access the content server and fetch the content through interacting with the associated GAP and the backhaul link. On the other hand, by applying caching technology, we assume that each GAP is equipped with a cache of finite memory storing some of the popular contents which might be requested by the UEs. Denote the maximum cache capacity of GAP i as $S_i^{\max}$, $1 \leq i \leq M$. Let K denote the number of content files and L_k denote the size of file k, $1 \leq k \leq K$. In the case that one UE's required content is stored at its associated GAP, the UE will fetch the content directly from the GAP without interacting with the remote content server. Let $\alpha_{jk} \in \{0, 1\}$ denote the association variable between UE j and content k, we set $\alpha_{jk} = 1$ if UE j requests file k; and $\alpha_{jk} = 0$, otherwise, where $1 \leq j \leq N$. Let $x_{ij} \in \{0, 1\}$ denote the binary association variable between GAP i and UE j, i.e., $x_{ij} = 1$, if UE j is associated with GAP i; otherwise, $x_{ij} = 0$. Denote the binary content placement variable at the GAPs as $\delta_{ik} \in \{0, 1\}$, i.e., $\delta_{ik} = 1$, if GAP i caches file k; otherwise, $\delta_{ik} = 0$. The system model considered in this paper is described in Fig. 1.

To allow resource sharing among multiple UEs accessing one GAP, we assume that the bandwidth resource of one GAP is divided into a number of subchannels with equal bandwidth and each UE can only be allocated one

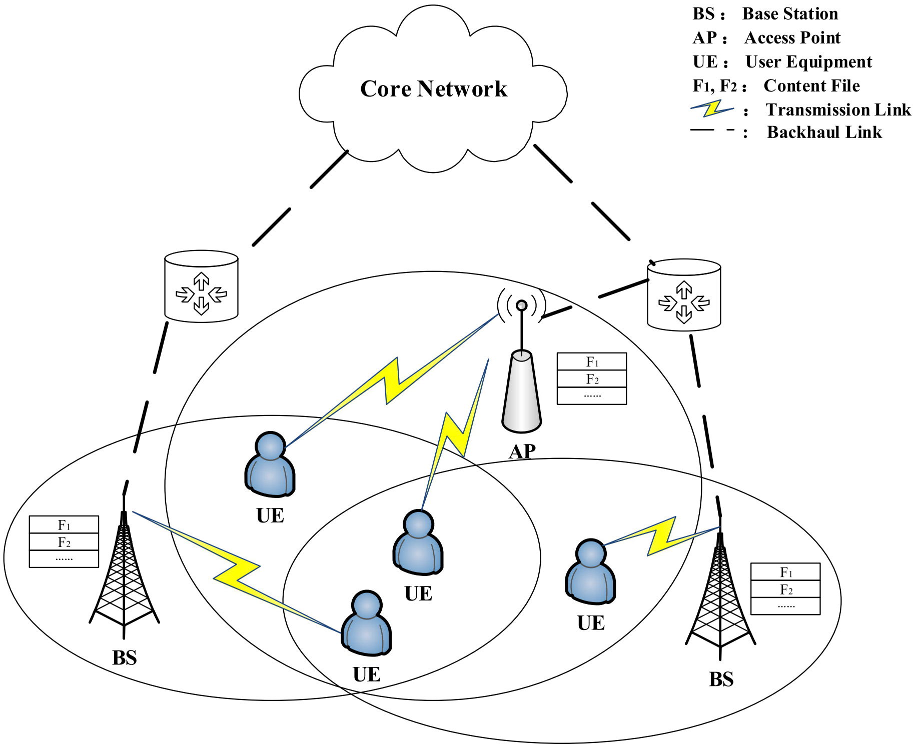

Fig. 1 System model. The system model considered in this paper

subchannel. Let $W_i^{\max}$ denote the available bandwidth of GAP i and W_i denote the bandwidth of each subchannel of GAP i, the maximum number of users associated to GAP i can be calculated as $A_i = \left\lfloor \frac{W_i^{\max}}{W_i} \right\rfloor$.

In this paper, we assume the received signal of UE j from GAP i denoted by y_{ij} can be expressed as $y_{ij} = \sqrt{p_i g_{ij}} s_{ij} + z_{ij}$, where s_{ij} represents the transmitted data symbol of GAP i when transmitting to UE j and p_i denotes the transmission power of GAP i, g_{ij} represents the channel gain of the link from GAP i to UE j, $g_{ij} = l_{ij} g_{ij}^0$, where l_{ij} denotes the path loss of the link between GAP i and UE j and g_{ij}^0 denotes the slow fading channel coefficient of the link from GAP i to UE j, which is modeled as a Rayleigh distributed random variable, z_{ij} is the additive white Gaussian noise (AWGN) with zero mean and variance σ_{ij}^2.

3.2 Proposed joint resource management architecture

The heterogeneity of various RANs and the distributed nature of the networks pose challenges to efficient resource management in future HetNets. In order to tackle this problem, we first propose a joint resource management architecture, which includes three types of functional entities, i.e., global resource management entity (GRME), local resource management entity (LRME), and user resource management entity (URME), and is shown in Fig. 2. Their main functions are summarized as follows.

URME is a functional module embedded in each UE. It is responsible for collecting and storing user status information, including channel state information (CSI) and service requirements, and sending the collected information to the associated LRME, which is a functional module deployed in each RAN. LRME is responsible for monitoring and storing network status information, collecting user status information from the associated URMEs, and then sending the network and user status information to the GRME. In practice, an LRME can be integrated to a GAP. GRME is a functional module deployed on top of all the RANs. It receives network and user state information from different LRMEs and conducts the proposed user association and content placement algorithm and sends the strategies to the LRMEs. Upon receiving association and resource allocation strategies from the GRME, the LRMEs conduct the corresponding operations and forward the strategies to the associated URMEs.

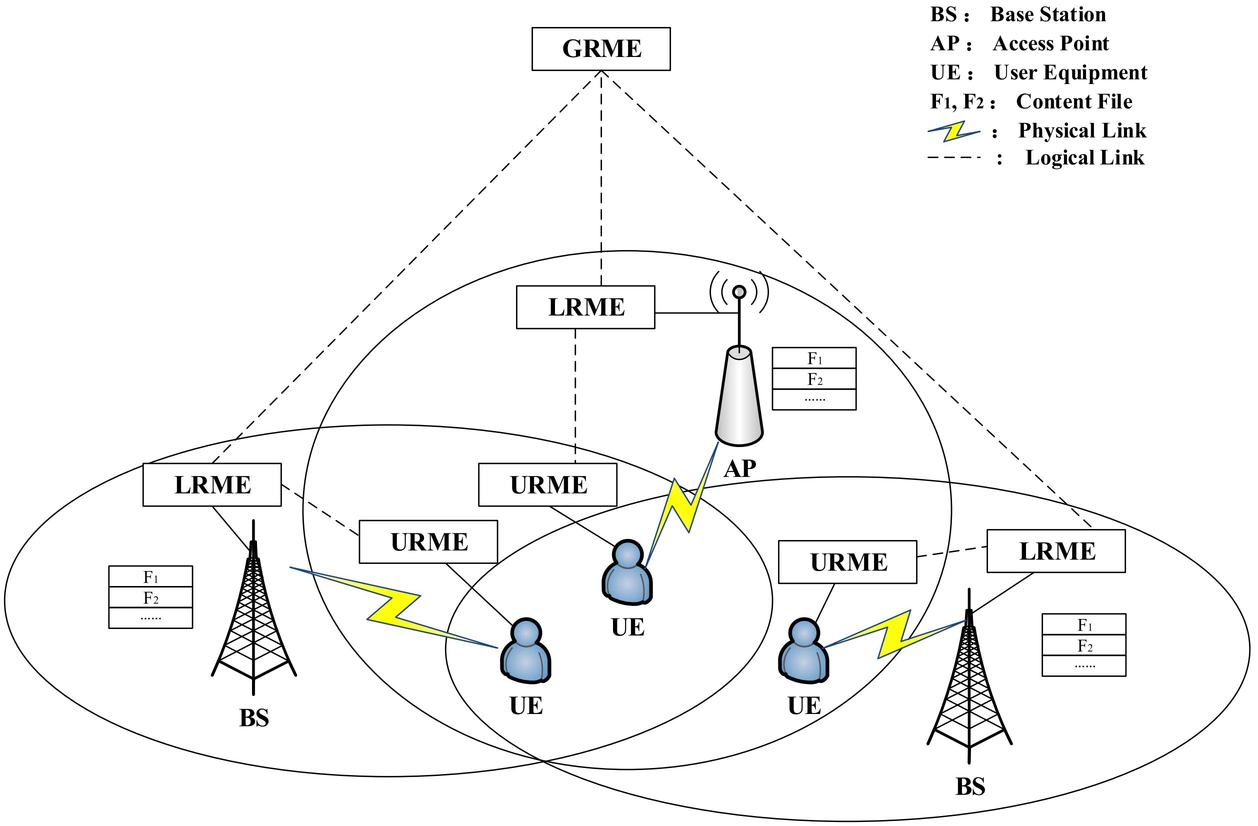

Fig. 2 Proposed joint resource management architecture. The proposed joint resource management architecture based on which the joint resource management scheme can be conducted

It should be mentioned that the information interactions between GRME, LRMEs, and URMEs can be performed over a common control channel. In this paper, we assume that efficient information interactions between GRME, LRMEs, and URMEs can be achieved, and design joint user association and content placement algorithm accordingly.

4 Utility function optimization formulation

In this section, we formulate the joint user association and cache content placement problem in HetNet as a utility optimization problem.

4.1 Objective function

We first consider the download delay for UE j when acquiring file k through GAP i. The main components of the download delay are the wireless transmission delay from the GAPs, the content fetching delay from the caches, and the backhaul delay through the core network, the formulations of which are described in following subsections.

4.1.1 Wireless transmission delay

Whether UEs acquire requested files from the caches of GAPs or remote content server, there is a transmission delay of wireless link between GAPs and UEs which mainly depends on the wireless communication states. When UE j with the request of file k is associated with GAP i, the transmission delay of wireless link can be expressed as

$$D_{ijk}^{\mathrm{ac}} = \frac{L_k}{R_{ij}^{\mathrm{t}}} \tag{1}$$

where R_{ij}^{t} denotes the data rate of the link between GAP i and UE j, which can be calculated as

$$R_{ij}^{\mathrm{t}} = W_i \log_2\left(1 + \frac{p_i g_{ij}}{\sigma_{ij}^2}\right) \tag{2}$$

where σ_{ij}^2 denotes the noise power of the link between UE j and GAP i.

4.1.2 Content fetching delay

To access the contents stored at local caches, the content fetching delay, i.e., the time duration required for searching the contents in the cache should be considered. In general, content fetching delay may vary depending on the types of storage devices, the underlying file searching mechanisms, and the amount of caching contents. Referring to [20], we model the content fetching delay as an exponentially distributed random variable and denote D_{ijk}^{ca} as the average content fetching delay required when UE j fetching file k from GAP i.

4.1.3 Backhaul transmission delay

When the contents users request cannot be fetched from local caches, the users should download the requested contents from the core network through backhaul links, thus resulting in backhaul transmission delay.

In this paper, we mainly consider the backhaul transmission delay of the links between GAPs and the gateways of the core network. Denoting the number of hops between GAP i associating user j which requesting file k and the gateway of the core network as N_i, we can express the backhaul delay denoted by D_{ijk}^{co} as

$$D_{ijk}^{\mathrm{co}} = \sum_{l=1}^{N_i} D_{ijkl}^{\mathrm{co}} \tag{3}$$

where D_{ijkl}^{co} denotes the backhaul delay of the lth hop link between GAP i and the gateway of the core network when the GAP i associating user j which requesting file k. D_{ijkl}^{co} is composed of the transmission delay and propagation delay over the link at the lth hop as well as the processing delay at the lth hop node, thus can be expressed as

$$D_{ijkl}^{\mathrm{co}} = D_{ijkl}^{\mathrm{t,co}} + D_{ijkl}^{\mathrm{s,co}} + D_{ijkl}^{\mathrm{p,co}} \tag{4}$$

where $D_{ijkl}^{\mathrm{t,co}}$ and $D_{ijkl}^{\mathrm{s,co}}$ denote the transmission delay and the propagation delay over the lth hop link between GAP i and the gateway respectively and $D_{ijkl}^{\mathrm{p,co}}$ denotes the average processing delay at the lth hop node of the backhaul link between GAP i and the gateway when UE j with the request of file k accessing GAP i. $D_{ijkl}^{\mathrm{t,co}}$ in (4) can be calculated as

$$D_{ijkl}^{\mathrm{t,co}} = \frac{L_k}{R_{il}^{\mathrm{co}}} \tag{5}$$

where R_{il}^{co} denotes the available transmission rate over the lth hop link between GAP i and the gateway; $D_{ijkl}^{\mathrm{s,co}}$ in (4) can be expressed as

$$D_{ijkl}^{\mathrm{s,co}} = \frac{H_{il}}{v} \tag{6}$$

where H_{il} denotes the distance between the lth hop node and the $(l+1)$th hop node of the backhaul link between GAP i and the gateway and v is the propagation speed of the wired link.

While being transmitted over wired backhaul link, user data packets are processed at each transmission node, resulting in processing delay. Referring to [21], it can be shown that the average processing delay at the lth hop between GAP i and the gateway when transmitting file k for UE j can be expressed as

$$D_{ijkl}^{\mathrm{p,co}} = \kappa_{il}(a_{il} + L_k \varpi_{il}) \tag{7}$$

where κ_{il}, a_{il}, and ϖ_{il} are constants, representing the processing capability of the lth hop node in the backhaul link of GAP i.

Consequently, the download delay for UE *j* when acquiring file *k* through GAP i, denoted by D_{ijk}, can be written as

$$D_{ijk} = D^{\rm ac}_{ijk} + \delta_{ik} D^{\rm ca}_{ijk} + (1-\delta_{ik}) D^{\rm co}_{ijk} \tag{8}$$

Hence, the total user download delay when accessing the requested contents, denoted by D, can be expressed as

$$D = \sum_{i=1}^{M}\sum_{j=1}^{N}\sum_{k=1}^{K} \alpha_{jk} x_{ij} D_{ijk} \tag{9}$$

In addition to the user download delay, we also concern about the caching cost in the cache-enabled HetNet. In general, the memory cost of caching user contents at the infrastructures of RANs is jointly determined by the volume of contents and the price of storage devices. In this paper, we assume that the memory cost of the GAPs is proportional to the size of contents and the unit price of content caching and express the total memory cost of the GAPs denoted by C as

$$C = \sum_{i=1}^{M}\sum_{k=1}^{K} \delta_{ik} \rho_i L_k \tag{10}$$

where ρ_i denotes the unit price of caching contents at GAP i.

Stressing the tradeoff between user download delay and the caching cost, we define the network utility, denoted by U as follows

$$U = D + \lambda C \tag{11}$$

where λ is a weighting factor.

4.2 Optimization constraints

To minimize the network utility in terms of joint user association and cache content placement strategies, a number of optimization constraints have to be considered. In this subsection, we discuss the optimization constraints.

4.2.1 Transmission rate constraint

In HetNets, users with various service requirements may pose different QoS constraints on serving GAPs. In this paper, stressing service sensitiveness to transmission rate, we assume that users may have different data rate requirements. Let R_j denote the transmission rate of UE *j*, $1 \le j \le N$, we can express R_j as

$$R_j = \sum_{i=1}^{M} x_{ij} R^{\rm t}_{ij}. \tag{12}$$

Let $R_j^{\min}$ denote the minimal transmission rate requirement of UE *j*, we can express the transmission rate constraint of UE *j* as

$$R_j \ge R_j^{\min}. \tag{13}$$

4.2.2 Maximum cache capacity constraint

On account of hardware conditions, the cache capacity of GAPs should subject to a maximum cache capacity constraint. Denoting S_i as the size of the content files which are cached in GAP i, $1 \le i \le M$, we can express S_i as

$$S_i = \sum_{k=1}^{K} \delta_{ik} L_k. \tag{14}$$

Denote the maximum cache capacity of GAP i as $S_i^{\max}$, we obtain the cache capacity constraint:

$$S_i \le S_i^{\max}. \tag{15}$$

4.2.3 User association constraints

According to the assumptions on user association, i.e., each user can only access one GAP and the number of users that access to each GAP should less than the maximum allowable number of users, we can obtain the association variable constraints

$$\sum_{i=1}^{M} x_{ij} = 1, \tag{16}$$

$$\sum_{j=1}^{N} x_{ij} \le A_i. \tag{17}$$

4.3 Optimization problem

By combining the optimization objective function and the constraints, the network utility minimization-based joint user association and cache content placement problem can be formulated as

$$\begin{aligned} &\min_{x_{ij},\delta_{ik}} U \qquad\qquad (18)\\ &\text{s.t. } \mathrm{C1}: \sum_{i=1}^{M} x_{ij} = 1\\ &\quad\;\; \mathrm{C2}: \sum_{j=1}^{N} x_{ij} \le A_i\\ &\quad\;\; \mathrm{C3}: S_i \le S_i^{\max}\\ &\quad\;\; \mathrm{C4}: R_j \ge R_j^{\min}\\ &\quad\;\; \mathrm{C5}: x_{ij} \in \{0,1\}\\ &\quad\;\; \mathrm{C6}: \delta_{ik} \in \{0,1\}. \end{aligned}$$

Through solving above optimization problem, we can obtain the optimal joint user association and cache content placement strategies.

5 Solution of the optimization problem

The optimization problem formulated in (18) is a nonlinear integer optimization which cannot be solved conveniently using traditional optimization tools. In this section, we apply Lagrange partial relaxation and

McCormick envelopes to obtain the Lagrangian dual problem. Benefited from the McCormick envelopes, the coupling among optimization variables in the optimization problem is removed; hence, we will be able to equivalently transform the optimization problem into of three subproblems. To jointly calculate the optimization variables and the Lagrange multipliers contained in the three subproblems, we propose an iterative algorithm in which given a set of Lagrange multipliers, the three subproblems are solved respectively by means of the modified K-M algorithm and the locally optimal optimization variables can be obtained, based on which the Lagrange multipliers can be updated through applying subgradient method. The detail process will be described in this section.

5.1 Equivalent transformation of original optimization problem

The tight coupling of the user association variables and the content placement variables in the objective function in (18) causes the difficulties in solving the problem. In this subsection, by introducing new variables and applying Lagrange partial relaxation and McCormick envelopes, the original optimization problem is equivalently transformed into three subproblems.

5.1.1 Introduction of new variable

To decouple the user association variables x_{ij} and the content placement variables δ_{ik} in the objective function in (18), we introduce a new variable, $\varphi_{ijk} = x_{ij}\delta_{ik}$, and the optimization problem can be rewritten as follows [19]

$$\begin{aligned}\min_{x_{ij},\delta_{ik},\varphi_{ijk}} \ & \sum_{i=1}^{M}\sum_{j=1}^{N}\sum_{k=1}^{K}\Big(\alpha_{jk}x_{ij}D_{ijk}^{\mathrm{ac}} + \alpha_{jk}\varphi_{ijk}D_{ijk}^{\mathrm{ca}} + \alpha_{jk}x_{ij}D_{ijk}^{\mathrm{co}} \\ & \qquad -\alpha_{jk}\varphi_{ijk}D_{ijk}^{\mathrm{co}} + \lambda\rho_i\delta_{ik}L_k\Big) \\ \text{s.t.} \quad & \mathrm{C1}-\mathrm{C6} \ \text{ in (18)} \\ & \mathrm{C7}: \ \varphi_{ijk} = x_{ij}\delta_{ik}, \forall i,j,k. \end{aligned} \tag{19}$$

where C7 is a non-convex constraint. By applying McCormick envelopes [22], the constraint C7 can be equivalently expressed as

$$\begin{aligned} &\mathrm{C8}: \ \varphi_{ijk} \geq 0 \\ &\mathrm{C9}: \ \varphi_{ijk} \geq \delta_{ik} + x_{ij} - 1 \\ &\mathrm{C10}: \ \varphi_{ijk} \leq x_{ij} \\ &\mathrm{C11}: \ \varphi_{ijk} \leq \delta_{ik}. \end{aligned}$$

Hence, we can further rewrite the optimization problem formulated in (19) as:

$$\begin{aligned}\min_{x_{ij},\delta_{ik},\varphi_{ijk}} \ & \sum_{i=1}^{M}\sum_{j=1}^{N}\sum_{k=1}^{K}\Big(\alpha_{jk}x_{ij}D_{ijk}^{\mathrm{ac}} + \alpha_{jk}\varphi_{ijk}D_{ijk}^{\mathrm{ca}} + \alpha_{jk}x_{ij}D_{ijk}^{\mathrm{co}} \\ & \qquad -\alpha_{jk}\varphi_{ijk}D_{ijk}^{\mathrm{co}} + \lambda\rho_i\delta_{ik}L_k\Big) \\ \text{s.t.} \quad & \mathrm{C1}-\mathrm{C6}, \ \ \mathrm{C8}-\mathrm{C11}. \end{aligned} \tag{20}$$

5.1.2 Lagrangian dual problem formulation

In this section, the Lagrange partial relaxation method [23] is introduced to solve the optimization problem defined in (20). Relaxing the constraints C9–C11, we can obtain the Lagrange function of (20)

$$\begin{aligned} L\left(\mu_{ijk},\upsilon_{ijk},\theta_{ijk},x_{ij},\delta_{ik},\varphi_{ijk}\right) = & \sum_{i=1}^{M}\sum_{j=1}^{N}\sum_{k=1}^{K}\Big[\alpha_{jk}x_{ij}D_{ijk}^{\mathrm{ac}} \\ & + \alpha_{jk}\varphi_{ijk}D_{ijk}^{\mathrm{ca}} + \alpha_{jk}x_{ij}D_{ijk}^{\mathrm{co}} \\ & - \alpha_{jk}\varphi_{ijk}D_{ijk}^{\mathrm{co}} + \lambda\rho_i\delta_{ik}L_k \\ & + \mu_{ijk}\left(\delta_{ik} + x_{ij} - 1 - \varphi_{ijk}\right) \\ & + \upsilon_{ijk}\left(\varphi_{ijk} - x_{ij}\right) \\ & + \theta_{ijk}\left(\varphi_{ijk} - \delta_{ik}\right)\Big] \end{aligned} \tag{21}$$

where μ_{ijk}, υ_{ijk}, and θ_{ijk} are the respective dual Lagrange multipliers for C9, C10, and C11, which must meet the following constrains:

$$\begin{aligned} &\mathrm{C12}: \ \mu_{ijk} \geq 0, \\ &\mathrm{C13}: \ \upsilon_{ijk} \geq 0, \\ &\mathrm{C14}: \ \theta_{ijk} \geq 0. \end{aligned}$$

Hence, we can formulate the Lagrange dual problem as follows:

$$\begin{aligned} \max_{\mu_{ijk},\upsilon_{ijk},\theta_{ijk}} \ \min_{x_{ij},\delta_{ik},\varphi_{ijk}} \ & L(\mu_{ijk},\upsilon_{ijk},\theta_{ijk},x_{ij},\delta_{ik},\varphi_{ijk}) \\ \text{s.t.} \quad & \mathrm{C1}-\mathrm{C6}, \ \mathrm{C8}, \ \mathrm{C12}-\mathrm{C14}. \end{aligned} \tag{22}$$

For a given set of Lagrange multipliers, the Lagrange function can be expressed as

$$L(\mu_{ijk},\upsilon_{ijk},\theta_{ijk},x_{ij},\delta_{ik},\varphi_{ijk}) = f(x_{ij}) + g(\delta_{ik}) + h(\varphi_{ijk}) - \mu_{ijk} \tag{23}$$

where

$$f(x_{ij}) = \left(\alpha_{jk}D_{ijk}^{\mathrm{ac}} + \alpha_{jk}D_{ijk}^{\mathrm{co}} + \mu_{ijk} - \upsilon_{ijk}\right)x_{ij}, \tag{24}$$

$$g(\delta_{ik}) = (\lambda\rho_i L_k + \mu_{ijk} - \theta_{ijk})\delta_{ik}, \tag{25}$$

$$h(\varphi_{ijk}) = \left(\alpha_{jk}D_{ijk}^{\mathrm{ca}} - \alpha_{jk}D_{ijk}^{\mathrm{co}} - \mu_{ijk} + \upsilon_{ijk} + \theta_{ijk}\right)\varphi_{ijk}. \tag{26}$$

5.1.3 Subproblem formulation

Since there is no variable coupling in $f(x_{ij}), g(\delta_{ik}), h(\varphi_{ijk})$, and the corresponding optimization constraints, we can decompose the optimization problem of minimizing $L(\mu_{ijk}, \upsilon_{ijk}, \theta_{ijk}, x_{ij}, \delta_{ik}, \varphi_{ijk})$ subject to constraints into three subproblems, i.e., subproblem 1 (SP1): user association subproblem; SP2: content placement subproblem; SP3: joint optimization subproblem. The three subproblems can be formulated as

$$\begin{aligned} \mathrm{SP1}: &\min_{x_{ij}} \sum_{i=1}^{M}\sum_{j=1}^{N}\sum_{k=1}^{K}\left(\alpha_{jk}D_{ijk}^{\mathrm{ac}} + \alpha_{jk}D_{ijk}^{\mathrm{co}} + \mu_{ijk} - \upsilon_{ijk}\right)x_{ij} \\ &\text{s.t.}\quad \mathrm{C1, C2, C4, C5.} \end{aligned} \tag{27}$$

$$\begin{aligned} \mathrm{SP2}: &\min_{\delta_{ik}} \sum_{i=1}^{M}\sum_{j=1}^{N}\sum_{k=1}^{K}(\lambda\rho_i L_k + \mu_{ijk} - \theta_{ijk})\delta_{ik} \\ &\text{s.t.}\quad \mathrm{C3,\ C6.} \end{aligned} \tag{28}$$

$$\begin{aligned} \mathrm{SP3}: &\min_{\varphi_{ijk}} \sum_{i=1}^{M}\sum_{j=1}^{N}\sum_{k=1}^{K}\left(\alpha_{jk}D_{ijk}^{\mathrm{ca}} - \alpha_{jk}D_{ijk}^{\mathrm{co}} - \mu_{ijk} + \upsilon_{ijk} + \theta_{ijk}\right)\varphi_{ijk} \\ &\text{s.t.}\quad \mathrm{C8.} \end{aligned} \tag{29}$$

5.2 Proposed iterative method

To obtain the optimal solution of the subproblems SP1, SP2, and SP3, the Lagrange multipliers should be jointly solved together with the optimization variables. To this end, we propose an iterative method, which compute the optimal solution and the Lagrange multipliers of the subproblems iteratively.

5.2.1 K-M algorithm-based optimal solutions to the subproblems

It can be shown that for a given set of Lagrange multipliers, the subproblems SP1, SP2, and SP3 are constrained integer optimization problem, which is equivalent to the optimal matching problem in the bipartite graph theory, thus can be solved by applying the classical algorithms such as modified K-M algorithm [24]. In this subsection, we assume that the Lagrange multipliers are given and seek for the locally optimal solution to the subproblems based on the modified K-M algorithm.

The following are some definitions and a theorem related to modified K-M algorithm.

Complete bipartite graph: Given a graph $G = (V; E)$ where V denotes a set of vertices and E denotes a set of edges connecting pairs of vertices. If the set V can be divided into two disjoint and non-empty sets, X and Y, i.e., $V = X \cup Y$ and $X \cap Y = \Phi$, where Φ denotes the empty set, every edge in E connects one vertex in X to another vertex in Y and no edge connects two vertices of the same set, we call G a complete bipartite graph.

Weighted complete bipartite graph: A complete bipartite graph $G = (V; E)$ is a weighted complete bipartite graph if any edge $e_{x,y} \in E$ connecting $x \in X$ and $y \in Y$ is assigned a non-negative weight $w(x, y)$.

Maximum matching: A matching H of graph $G = (V; E)$ is defined as a subset $H \subseteq E$ which meets the condition that for $\forall e_{x,y}, e_{x',y'} \in H$, $e_{x,y}$ and $e_{x',y'}$ are not adjacent in G. The size of a matching H, denoted by $|H|$ is defined as the number of edges contained in H. A matching H is called a maximum matching if for any other matching H' of G, the condition $\left|H'\right| \leq |H|$ holds.

Optimal matching: The maximum matching H of graph $G = (V; E)$ is called an optimal matching if it achieves the maximum sum weight, i.e.,

$$\sum_{e_{x,y}\in H} w(x, y) \geq \sum_{e_{x,y}\in H'} w(x, y). \tag{30}$$

Feasible vertex labeling: A real valued function l is called a feasible vertex labeling if for any $x \in X$ and $y \in Y$, $l(x) + l(y) \leq w(x, y)$ holds.

Equality subgraph: If l is a feasible vertex labeling, let G_l denote a subgraph of G, if the condition $l(x) + l(y) = w(x, y)$ holds, then G_l is called the equality subgraph of G with respect to l.

Theorem *If l is a feasible vertex labeling of G and H is an optimal matching of X to Y with $H \subseteq G_l$, then H is an optimal assignment from X to Y. Thus, the problem of finding an optimal matching in a complete bipartite graph reduces to the problem of finding a feasible vertex labeling of which the equality subgraph contains an optimal assignment from X to Y.*

According to above definitions and theorem, to solve the optimization problem SP1, we can construct a weighted bipartite graph $G^0 = (V_1, V_2; E)$, where the set of vertices V_1 represents the collection of subchannels of GAPs, $V_1 = \left[\mathrm{SC}_{1,1}, \mathrm{SC}_{1,2}, \cdots, \mathrm{SC}_{1,A_1}, \cdots, \mathrm{SC}_{M,1}, \cdots, \mathrm{SC}_{M,A_M}\right]$, $\mathrm{SC}_{i,q}$ denotes the qth subchannel of the ith GAP, $1 \leq i \leq M$, $1 \leq q \leq A_i$ and the set of vertices V_2 represents the collection of UEs, $V_2 = [\mathrm{UE}_1, \mathrm{UE}_2, \ldots, \mathrm{UE}_N]$, the weight of the edge connecting subchannel of GAP i, and UE j is defined as

$$w(\mathrm{SC}_{i,q}, \mathrm{UE}_j) = \alpha_{jk}D_{ijk}^{\mathrm{ac}} + \alpha_{jk}D_{ijk}^{\mathrm{co}} + \mu_{ijk} - \upsilon_{ijk}. \tag{31}$$

The steps for solving the optimal user association subproblem can be summarized as follows:

1. Define an initial feasible vertex labeling $l(u)$.
2. Given $l(u)$, obtain G_l from G^0 and determine a maximum matching H of G^0.
3. If H is an optimal matching of G^0, the optimization problem is solved and the optimal user association strategy can be obtained correspondingly.
4. If H is not an optimal matching of G^0, a vertex $x \in V_1$ having not being allocated is selected in G_l, set $S = \{x\}$ and $T = \Phi$.
5. Let $N_{G_l}(S)$ denote the collection of vertices which connect with S in G_l. If $N_{G_l}(S) \neq T$, go to step 4. Otherwise, $N_{G_l}(S) = T$. Find

$$\Delta = \min_{u,v}\{l(u)|l(u)+l(v)-w(u,v), u \in S, v \in V_2 - T\} \tag{32}$$

and define a new labeling $l^{'}(u)$ by

$$l^{'}(u) = \begin{cases} l(u) - \Delta, & u \in S \\ l(u) + \Delta, & u \in T \\ l(u), & \text{others.} \end{cases} \tag{33}$$

6. Replace $l(u)$ by $l^{'}(u)$, go to step 2.

Through conducting above process iteratively, an optimal matching of G^0 can be obtained corresponding to the optimal user association strategy. Similar methods can be applied to solve the optimization problem SP2 and SP3 and obtain the optimal content placement variable δ_{ik}^* and the optimal joint variable φ_{ijk}^*. Both subproblems are constrained integer optimization problem, which can be transformed as an optimal matching problem in the bipartite graph and solved based on the modified K-M algorithm.

5.2.2 *Updating Lagrange multipliers*

Upon obtaining the locally optimal solution of x_{ij}^*, δ_{ik}^*, and φ_{ijk}^*, the Lagrange multipliers can be updated using subgradient method, i.e.,

$$\mu_{ijk}(t+1) = \left[\mu_{ijk}(t) - \epsilon_1(\varphi_{ijk}(t) + 1 - \delta_{ik}(t) - x_{ij}(t))\right]^+ \tag{34}$$

$$\upsilon_{ijk}(t+1) = \left[\upsilon_{ijk}(t) - \epsilon_2(x_{ij}(t) - \varphi_{ijk}(t))\right]^+ \tag{35}$$

$$\theta_{ijk}(t+1) = \left[\theta_{ijk}(t) - \epsilon_3(\delta_{ik}(t) - \varphi_{ijk}(t))\right]^+ \tag{36}$$

where $[z]^+ = \max\{0, z\}$, and ϵ_1, ϵ_2, and ϵ_3 are the step sizes with respect of $\mu_{ijk}(t), \upsilon_{ijk}(t)$, and $\theta_{ijk}(t)$, respectively.

Conducting the above process iteratively, we will be able to achieve the convergence of the algorithm [25]. Once the convergence condition meets, we can obtain the globally optimal user association and content placement strategy.

The proposed method is summarized in Algorithm 1.

Algorithm 1 Proposed Iterative Algorithm for Solving the Original Optimization Problem

```
Set the maximum number of iterations $T^{\max}$ and the maximum tolerance $\omega$
Initialize Lagrange multipliers $\mu_{ijk}$, $\upsilon_{ijk}$ and $\theta_{ijk}$
repeat
    Solve SP1 to obtain the locally optimal solution of $x_{ij}$
    Solve SP2 to obtain the locally optimal solution of $\delta_{ik}$
    Solve SP3 to obtain the locally optimal solution of $\varphi_{ijk}$
    Update the Lagrange multipliers:
    $\mu_{ijk}(t+1) = [\mu_{ijk}(t) - \epsilon_1(\varphi_{ijk}(t) + 1 - \delta_{ik}(t) - x_{ij}(t))]^+$
    $\upsilon_{ijk}(t+1) = [\upsilon_{ijk}(t) - \epsilon_2(x_{ij}(t) - \varphi_{ijk}(t))]^+$
    $\theta_{ijk}(t+1) = [\theta_{ijk}(t) - \epsilon_3(\delta_{ik}(t) - \varphi_{ijk}(t))]^+$
    if $\sum_{i=1}^{M}\sum_{j=1}^{N}\sum_{k=1}^{K}[|\mu_{ijk}(t+1) - \mu_{ijk}(t)| + |\upsilon_{ijk}(t+1) - \upsilon_{ijk}(t)| + |\theta_{ijk}(t+1) - \theta_{ijk}(t)|] \leq \omega$ then
        Convergence = true
        return $x_{ij}^* = x_{ij}$, $\delta_{ik}^* = \delta_{ik}$, $\varphi_{ijk}^* = \varphi_{ijk}$
    else
        $t = t + 1$
    end if
until Convergence = true or $t = T^{\max}$
```

6 Simulation results

In this section, we examine the performance of the proposed joint user association and cache content placement algorithm and compare the algorithm with previously proposed algorithms via simulation. In the simulation, we consider a HetNet scenario consisting of two BSs, two APs, and a number of UEs. The size of the simulation region is 100 × 100, and the coordinates of the BSs and APs are respectively (50, 50), (55, 55), (60, 60), and (65, 65). The number of UEs is chosen as 5 and 6 in the simulation, and the UEs are randomly located in the simulation area. Other parameters used in the simulation are summarized in Table 1. The simulation results are averaged over 1000 independent adaptation processes where each adaptation process involves different positions of UEs.

Figure 3 shows the network utility versus the number of iterations obtained from the proposed algorithm. The number of UEs is chosen as 5, and various subchannel bandwidth is considered in the simulation. It can be observed that the algorithm converges within a small number of iterations which demonstrates the effectiveness of the proposed algorithm. Comparing the results obtained for different subchannel bandwidth, we can see that the network utility decreases as the subchannel bandwidth increases.

Table 1 System parameters. The system parameters considered in the simulation

Parameters	Value
Maximum cache capacity of BS	20 Mbits
Maximum cache capacity of AP	10 Mbits
File sizes	9, 8, 10, 9, 8, and 10 Mbits
Content fetching delay at BS	0.2 s
Content fetching delay at AP	0.3 s
Maximum numbers of active users of BS and AP	3
Minimum data rate requirements of users	1, 2, 0.5, 0.5, 1, and 2 Mbps
Small scale fading distribution	Rayleigh fading with zero mean and unit variance
Channel path loss model	$128.1 + 27\log(d)$ dB, d denotes the distance

Figure 4 shows the network utility versus the subchannel bandwidth for different discount factor, i.e., λ. The number of UEs is chosen as 5 and 6 in examining the results. For comparison, we plot the network utility obtained from our proposed algorithm and the algorithm proposed in [18]. It can be seen from the figure that for fixed λ, the network utility decreases with the increase of subchannel bandwidth and the proposed algorithm outperforms the one proposed in [18]. This is because the algorithm proposed in [18] fails to consider the memory cost and the content fetching delay, thus may result in larger network utility. Comparing the network utility obtained for different discount factor and different number of UEs, we can see that the network utility increases with the increase of discount factor and the number of UEs. This result can be expected from the definition of network utility.

Figure 5 depicts the network utility versus the subchannel bandwidth for different transmission power of GAPs. To plot the curves, we set the number of users as 6 in examining the results. It can be seen from the figure that for fixed transmission power of GAPs, the network utility decreases with the increase of subchannel bandwidth. For given subchannel bandwidth, we examine the performance of the proposed scheme and the scheme proposed in [18] for different transmission power of GAPs. Comparing the network utility obtained for different transmission power of GAPs, we can see that the network utility decreases with the increase of transmission power of GAPs. It is because the transmission delay of the wireless link decreases with the increase of transmission power of GAPs, thus resulting in smaller network utility. Comparing the curves obtained from our proposed scheme and the scheme proposed in [18], we can see that the proposed scheme offers smaller network utility than the scheme proposed in [18].

Figure 6 shows the network utility versus the subchannel bandwidth for different noise power, i.e., σ^2. To plot

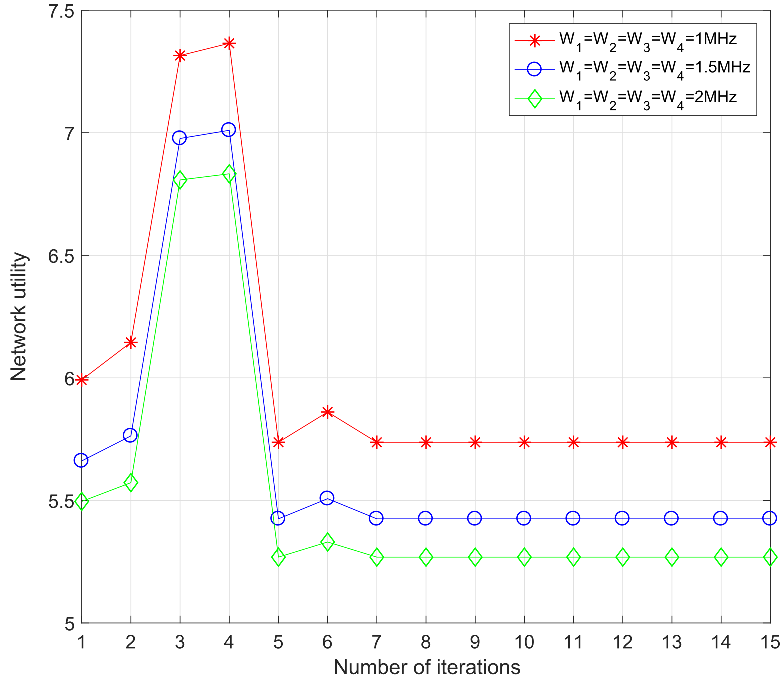

Fig. 3 Network utility versus the number of iterations. The network utility of the system versus the number of iterations obtained from the proposed algorithm

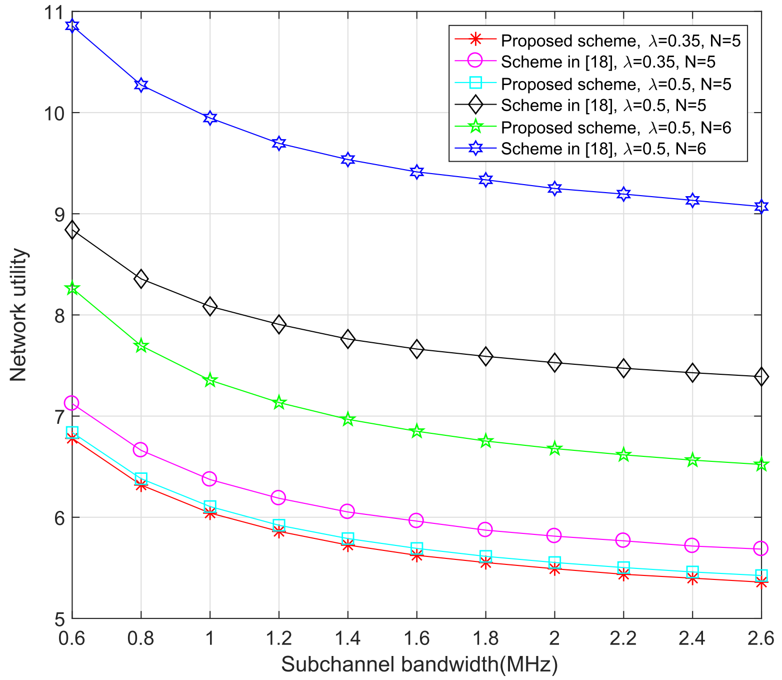

Fig. 4 Network utility versus subchannel bandwidth (different discount factor, different number of users). The network utility of the system versus the subchannel bandwidth for different discount factor and different number of users

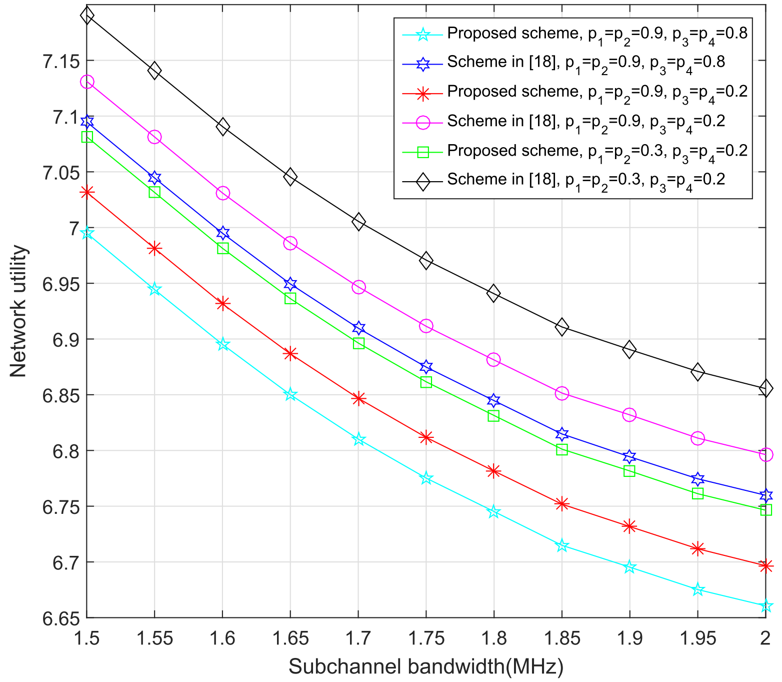

Fig. 5 Network utility versus subchannel bandwidth (different transmission power of GAPs). The network utility of the system versus the subchannel bandwidth for different transmission power of GAPs

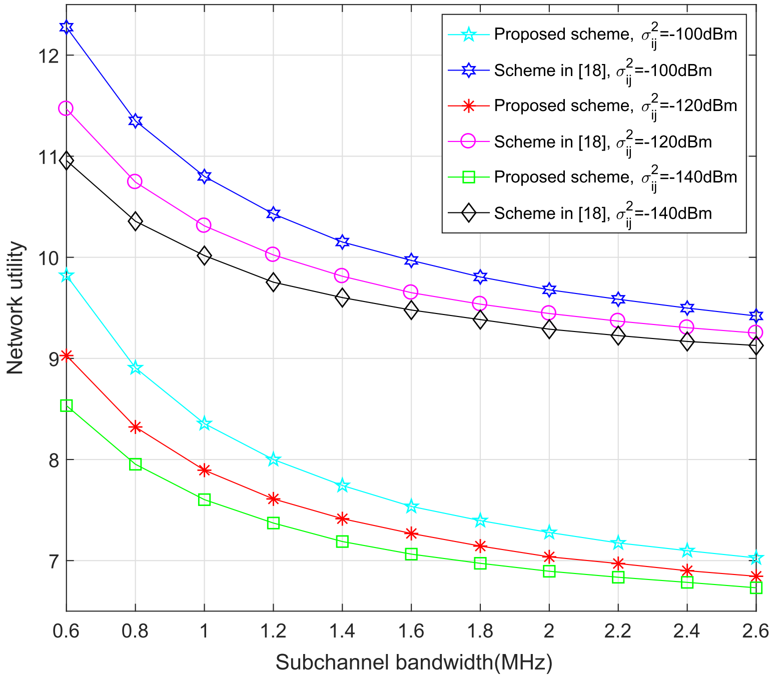

Fig. 6 Network utility versus subchannel bandwidth (different noise power). The network utility of the system versus the subchannel bandwidth for different noise power

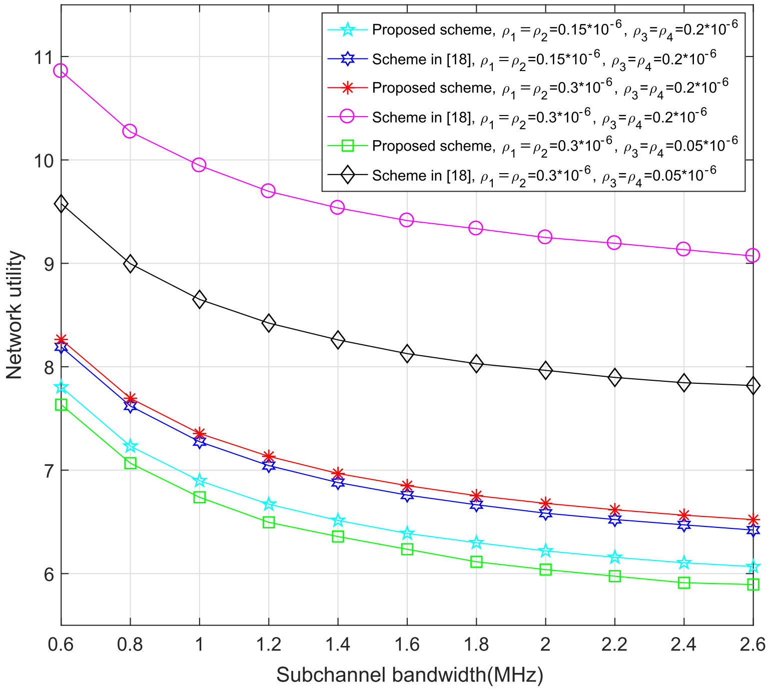

Fig. 7 Network utility versus subchannel bandwidth (different unit price of caching contents at GAPs). The network utility of the system versus the subchannel bandwidth for different unit price of caching contents at GAPs

the curves, we set $\lambda = 0.5$ and $N = 6$. For given subchannel bandwidth, we examine the performance of the proposed scheme and the scheme proposed in [18] for different noise power. It can be seen from the figure that for a fixed noise power, the network utility decreases with the increase of subchannel bandwidth. It can also be seen from the figure that the network utility decreases as the noise power decreases. This is because the transmission delay of the wireless link decreases with the decrease of the noise power, thus resulting in smaller network utility. Comparing the curves obtained from the two algorithms, we can see that the proposed scheme outperforms the scheme proposed in [18].

Figure 7 depicts the network utility versus the subchannel bandwidth for different unit price of caching contents at GAPs. The number of users is chosen as 6 in examining the results. For comparison, we plot the network utility obtained from our proposed scheme and the scheme proposed in [18]. It can be seen from the figure that for fixed unit price of caching contents at GAPs, the network utility decreases with the increase of subchannel bandwidth. Comparing the network utility obtained for different unit price of caching contents at GAPs, we can see that the network utility decreases with the decrease of unit price of caching contents at GAPs. This is because the cache cost decreases with the decrease of unit price of caching contents at GAPs, resulting in smaller network utility. Comparing the results obtained from our proposed algorithm and the algorithm proposed in [18], we can see that our proposed scheme outperforms the one proposed in [18].

7 Conclusions

In this paper, we study the joint user association and cache content placement problem in cache-enabled HetNets. By introducing a utility function defined as a weighted sum of user download delay and the caching cost, the joint user association and content placement problem is formulated as a utility function optimization problem which is transformed equivalently into three convex subproblems by applying McCormick envelopes and Lagrange partial relaxation. We then propose an iterative algorithm which compute the Lagrange multipliers and the optimal solution of the subproblems iteratively. Numerical results demonstrate the effectiveness of the proposed algorithm.

Acknowledgements

This work is supported by the National Science and Technology Specific Project of China (2016ZX03001010-004), National Natural Science Foundation of China (No. 61571073), the Joint Scientific Research Fund of Ministry of Education and China Mobile (MCM20160105), Wenfeng Talent Program of CQUPT (W2015-19), the special fund of Chongqing Key Laboratory (CSTC), and the project of Chongqing Municipal Education Commission (Kjzh11206).

Authors' contributions

The authors have equal contributions. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

[1] Key Lab of Mobile Communication Technology, Chongqing University of Posts and Telecommunications, Chongwen Road, Chongqing, People's Republic of China. [2] Department of Electrical and Computer Engineering, McMaster University, 1280 Main Street West, Hamilton L8S 4L8, Canada.

Received: 12 February 2018 Accepted: 30 April 2018
Published online: 16 May 2018

References

1. M Jo, T Maksymyuk, RL Batista, TF Maciel, ALF de Almeida, M Klymash, A survey of converging solutions for heterogeneous mobile networks. IEEE Wirel. Commun. **21**(6), 54–62 (2014). https://doi.org/10.1109/MWC.2014.7000972
2. H Boostanimehr, VK Bhargava, Unified and distributed qos-driven cell association algorithms in heterogeneous networks. IEEE Trans. Wirel. Commun. **14**(3), 1650–1662 (2015). https://doi.org/10.1109/TWC.2014.2371465
3. H Beyranvand, W Lim, M Maier, C Verikoukis, JA Salehi, Backhaul-aware user association in fiwi enhanced lte-a heterogeneous networks. IEEE Trans. Wirel. Commun. **14**(6), 2992–3003 (2015). https://doi.org/10.1109/TWC.2015.2399308
4. X Luo, Delay-oriented qos-aware user association and resource allocation in heterogeneous cellular networks. IEEE Trans. Wirel. Commun. **16**(3), 1809–1822 (2017). https://doi.org/10.1109/TWC.2017.2654458
5. G Hattab, D Cabric, in *2016 IEEE Global Communications Conference (GLOBECOM)*. Joint resource allocation and user association in multi-antenna heterogeneous networks, (2016), pp. 1–7. https://doi.org/10.1109/GLOCOM.2016.7841931
6. F Wang, W Chen, H Tang, Q Wu, Joint optimization of user association, subchannel allocation, and power allocation in multi-cell multi-association OFDMA heterogeneous networks. IEEE Trans. Commun. **65**(6), 2672–2684 (2017). https://doi.org/10.1109/TCOMM.2017.2678986
7. L Zhao, H Wang, X Zhong, in *2017 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*. An association approach that combines resource consumption and load balancing in heterogeneous cellular networks, (2017), pp. 776–783. https://doi.org/10.1109/iThings-GreenCom-CPSCom-SmartData.2017.119
8. C Liang, FR Yu, H Yao, Z Han, Virtual resource allocation in information-centric wireless networks with virtualization. IEEE Trans. Veh. Technol. **65**(12), 9902–9914 (2016). https://doi.org/10.1109/TVT.2016.2530716
9. X Wang, M Chen, T Taleb, A Ksentini, VCM Leung, Cache in the air: exploiting content caching and delivery techniques for 5G systems. IEEE Commun. Mag. **52**(2), 131–139 (2014). https://doi.org/10.1109/MCOM.2014.6736753
10. E Bastug, M Bennis, M Debbah, Living on the edge: the role of proactive caching in 5G wireless networks. IEEE Commun. Mag. **52**(8), 82–89 (2014). https://doi.org/10.1109/MCOM.2014.6871674
11. J Wen, K Huang, S Yang, VOK Li, Cache-enabled heterogeneous cellular networks: Optimal tier-level content placement. IEEE Trans. Wirel. Commun. **16**(9), 5939–5952 (2017). https://doi.org/10.1109/TWC.2017.2717819
12. S Krishnan, M Afshang, HS Dhillon, Effect of retransmissions on optimal caching in cache-enabled small cell networks. IEEE Trans. Veh. Technol. **66**(12), 11383–11387 (2017). https://doi.org/10.1109/TVT.2017.2721839
13. J Li, J Sun, Y Qian, F Shu, M Xiao, W Xiang, A commercial video-caching system for small-cell cellular networks using game theory. IEEE Access. **4**, 7519–7531 (2016). https://doi.org/10.1109/ACCESS.2016.2582836
14. T Liu, AA Abouzeid, in *2016 IEEE Global Communications Conference (GLOBECOM)*. Content placement and service scheduling in femtocell caching networks, (2016), pp. 1–6. https://doi.org/10.1109/GLOCOM.2016.7841691

15. D Liu, C Yang, in *2016 IEEE Global Communications Conference (GLOBECOM)*. Optimal content placement for offloading in cache-enabled heterogeneous wireless networks, (2016), pp. 1–6. https://doi.org/10.1109/GLOCOM.2016.7842078
16. J Liao, KK Wong, MRA Khandaker, Z Zheng, Optimizing cache placement for heterogeneous small cell networks. IEEE Commun. Lett. **21**(1), 120–123 (2017). https://doi.org/10.1109/LCOMM.2016.2612197
17. S He, H Tian, X Lyu, G Nie, S Fan, Distributed cache placement and user association in multicast-aided heterogeneous networks. IEEE Access. **5**, 25365–25376 (2017). https://doi.org/10.1109/ACCESS.2017.2769664
18. Y Cui, F Lai, S Hanly, P Whiting, in *2016 IEEE Global Communications Conference (GLOBECOM)*. Optimal caching and user association in cache-enabled heterogeneous wireless networks, (2016), pp. 1–6. https://doi.org/10.1109/GLOCOM.2016.7842077
19. Y Wang, X Tao, X Zhang, G Mao, Joint caching placement and user association for minimizing user download delay. IEEE Access. **4**, 8625–8633 (2016). https://doi.org/10.1109/ACCESS.2016.2633488
20. E Bastug, M Kountouris, M Bennis, M Debbah, in *2016 IEEE 17th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*. On the delay of geographical caching methods in two-tiered heterogeneous networks, (2016), pp. 1–5. https://doi.org/10.1109/SPAWC.2016.7536893
21. G Zhang, TQS Quek, M Kountouris, A Huang, H Shan, Fundamentals of heterogeneous backhaul design-analysis and optimization. IEEE Trans. Commun. **64**(2), 876–889 (2016). https://doi.org/10.1109/TCOMM.2016.2515596
22. L Liberti, CC Pantelides, An exact reformulation algorithm for large nonconvex NLPS involving bilinear terms. J. Glob. Optim. **32**(2), 161–189 (2006). https://doi.org/10.1007/s10898-006-9005-4
23. S Boyd, L Vandenberghe, *Convex Optimization*. (Cambridge University Press, Cambridge, 2004)
24. X Zhou, L Yang, D Yuan, Bipartite matching based user grouping for grouped OFDM-IDMA. IEEE Trans. Wirel. Commun. **12**(10), 5248–5257 (2013). https://doi.org/10.1109/TWC.2013.090413.130097
25. DP Bertsekas, *Nonlinear Programming*. (Athena Scientific, Belmont, 1999)