IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON TRUSTED COMPUTING

Received September 28, 2018, accepted October 13, 2018, date of publication October 23, 2018,
date of current version November 19, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2876893

On the Design of Computation Offloading in Cache-Aided D2D Multicast Networks

DONGYU WANG[1], YANWEN LAN[1], TIEZHU ZHAO[2], ZHENPING YIN[2], AND XIAOXIANG WANG[1]

[1]Key Laboratory of Universal Wireless Communications, Ministry of Education, Beijing University of Posts and Telecommunications, Beijing 100876, China
[2]Samsung R&D Institute of China, Beijing 100102, China

Corresponding author: Dongyu Wang (dy_wang@bupt.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61701038 and in part by the Fundamental Research Funds for the Central Universities.

ABSTRACT As the demand of data-hungry and computing intensive tasks grows dramatically, cache-aided device-to-device multicast (D2MD) networks are introduced, which can offload traffic from the base stations to D2D users (DUEs) directly to alleviate the heavy burden on backhaul links and improve the energy and spectrum efficiency. However, most previous works ignored the limitation of battery power and scarce computing capabilities of DUEs. In this paper, we study the computation and traffic offloading in cache-aided D2MD networks for the content delivery and delay sensitive task offloading services. Firstly, in order to provide stable multicast links and enhanced computing resources, a D2D cluster head (DCH) selection strategy is proposed that jointly considers the social attributes, available energy, and transfer rate of DUEs. Secondly, to improve the efficiency of content distribution and optimize the energy consumption of content delivery, we propose a novel multicast-aware coded and cooperative caching scheme, which may increase the opportunity for D2D multicasting to obtain the desired contents. Thirdly, considering the DUEs association, uplink full duplex DCH transmission power allocation, and mobile edge computing computation resource scheduling, an optimization computation offloading model is formulated. On this basis, we model the computation offloading and resource allocation optimization problem. Furthermore, we transform this problem into user allocation optimization problem and resource allocation optimization problem (RAOP), and RAOP is proved as a convex problem, and the optimal resource allocation solution is found. Finally, the simulation results show that our proposed schemes can effectively decrease the energy consumption and computing costs.

INDEX TERMS D2D Multicast, content offloading, cooperative caching, coded caching, computation offloading, MEC.

I. INTRODUCTION

Nowadays, with the rapid evolution of mobile communication technologies, more and more intelligent devices, such as smart phones, tablets, laptops, etc., have installed more and more mobile applications, e.g., multimedia videos sharing and real-time online games, which generate disseminate various types data traffic [1]–[3]. With the explosive increase of data traffic, the mobile devices are facing different wireless network access requirements for bandwidth-intensive and extensive-computing. As far as we know, the increase of these new applications and services puts a heavy burden on backhaul links which connect base stations (BSs) with the core network [4], [5]. Existing works focus on offloading the data-hungry or computing intensive tasks to the cloud for execution. However, it will undergo a long latency when mobile devices are connected to the cloud relayed by BS, which does not satisfy the requirements of delay sensitive tasks.

As a promising technique to offload cellular traffic, the emerging device-to-device (D2D) communication is regarded as an effective approach to satisfy the above massive demands and to achieve higher network capacity [6], [7]. It is quite common for people to share interested contents or play multi-player online games with others in their close vicinity by using D2D communication. Furthermore, for our focused content caching and computing offloading cases in which a cluster of D2D users (DUEs) in geographically close proximity request for the same contents e.g. 4K videos, or

computing tasks e.g. real-time virtual reality (VR) games, it is reasonable to leverage the broadcasting nature of wireless transmission [8]. In D2D multicast (D2MD) mode, knowledge of information about physical location and social ties strongly affects the formation of D2MD clusters and selection of D2MD cluster heads (DCHs). In practical cellular network, DUEs connect with each other directly using the licensed spectrum, thus offloading the traffic generated from DUE-BS links, which is under the control of a cellular BS. However, the mobile devices face some inevitable limitations including storage, battery and capacity.

In D2D multicast (D2MD) mode, knowledge of information about physical location and social ties strongly affects the formation of D2MD clusters and selection of D2MD cluster heads (DCHs). In practical cellular network, DUEs connect with each other directly using the licensed spectrum, thus offloading the traffic generated from DUE-BS links, which is under the control of a cellular BS. However, the mobile devices face some inevitable limitations including storage, battery and capacity.

To the best of our knowledge, based on new features and unique advantages provided by caching and computing at mobile devices, existing works related to cache-enabled D2D multicast network and D2D computing offloading mainly focus on the following three aspects.

A. D2D CLUSTER HEAD SELECTION

Selecting appropriate DCHs can improve D2MD performance in the cache-enabled D2D multicast network. Physical location, social ties and content popularity impose significant impact on D2MD cluster formation and DCHs selection [10]–[13]. An effective pricing-based multicast video distribution system and a grid-based clustering method in [10] are proposed which consider users and social characteristics to alleviate the BS traffic load. Choi *et al.* [11] propose a code-based discovery protocol for D2D to realize proximity based services. The works of [12] exploit the network knowledge extracted from underlying mobile social networks (MSNs) and propose a social aware D2D communication scheme to improve the spectrum and energy efficiency. Liu *et al.* [13] model the multicast group from a combination of geographic and non-geographic (GN) perspectives and propose the GN Model to characterize social relationship.

B. D2D CONTENT CACHING

Currently, the advanced research on caching is prefetcahing the most popular uncoded or coded files into the edge network (e.g. DCHs). When users make requests, files would be offloaded in the local network (e.g. DUEs), which can largely alleviate the network overheads. In addition, what to cache and where to cache are important issues in designing a caching strategy. There have been lots of efforts on D2D caching networks [14]–[28]. In [14], the offloading gain is calculated by considering the energy limit and data rate requirements of D2D members to maximize successful offloading probability. A jointly optimizing caching and scheduling policies for cache-enabled D2D communications is proposed in [15]. Given that the average long-term received power of the requester is not less than the D2D establishment threshold, authors propose a joint caching policy to maximize the content-related energy efficiency in [16]. Considering the content similarity and distributed among caching nodes in the scenario that nodes not only cache files from the base station, but also can cache files from nearby nodes, the cost reducing problem is formulated as a local cooperative game model [17], [18]. To maximize the probability of finding the desired contents within the maximal coverage range of D2D communications, the optimized and proactive caching schemes are proposed [19], [20]. Zhao *et al.* [21] propose a non-orthogonal multiple access-based multicast (NOMA-MC) scheme to improve the spectrum efficiency in which content objects can be pushed and multicasted simultaneously. Wang *et al.* [22] investigate the cost-optimal caching problem with user mobility for D2D networks. It optimizes caching placement so as to minimize the expected cost of obtaining files of interest by collecting file segments. In paper [23], for the scenario that mobile helpers (Candidate DCHs) with caching ability and DUEs with content requests are distributed as two independent homogeneous Poisson point process (HPPP) in the cell, based on the obtained results of successful content delivery probability, the optimal probabilistic caching strategy of mobile helper is investigated. In [24]–[26], to maximize the D2D system throughput with minimal data rate constraints and to minimize the transmission delay, opportunistic cooperation schemes for cache-enabled D2D communication are proposed. To optimize resource allocation, a cluster content caching structure is proposed in [27], which takes full advantages of distributed caching and centralized signal processing. To maximize the data offloading ratio, which is defined as the percentage of the requested data that can be delivered via D2D links rather than through BSs, a mobility-aware caching placement strategy is proposed by taking advantage of the user mobility pattern by the inter-contact times between different users in [28].

C. D2D COMPUTING OFFLOADING

Due to finite computing power and battery life of DCHs and DUEs, the development in the new applications and services is limited. In recent years, mobile edge computing (MEC) is proposed to solve the computational capability issues [29], [30]. By deploying computing servers at the edge of the network, MEC system can enable the users to offload tasks, offering users abundant wireless resources and vast computation capabilities, which help them enjoy shorter delay and higher performance computing services. There have been some previous works on computing offloading for MEC [31]–[38]. For instance, Zhang *et al.* [31] propose a task offloading scheme to determine which task should be remotely offloaded to the MEC servers or performed locally. The aim of their work is to minimize the energy consumption. The works in [32] and [33] study the task offloading in two-tier 5G small cell networks (SCNs). The forward

link and backward link are jointly considered to optimize the energy efficiency. By jointly considering computation offloading, resource allocation and content caching in cellular networks with MEC, an alternating direction method of multipliers (ADMM) based distributed algorithm is proposed to maximize the system revenue [34]. Their works are further studied by Wang *et al.* [35], in which a caching and full duplex (FD) communication enabled MEC framework is proposed. Tan *et al.* [36] propose a new concept of task caching and formulate an efficient task caching problem and efficient task offloading problem. An alternative iterative-based algorithm is proposed to solve the energy efficient problem. The works in [37] consider the user mobility and propose a coded caching scheme in MEC-Enabled SCNs. A MEC-enhanced adaptive bitrate video delivery scheme is studied in [38]. By the way, the utility function of the system is maximized. Although many works have been done on D2D and MEC, most of them only consider the two technologies separately, which leaves their potential advantages unexplored.

Unfortunately, the main concern of D2D content caching and computing offloading is the storage capacity of the DUEs. The current works seldom consider the computing and storage capabilities of DCH and MEC at the same time. Furthermore, most current works does not consider the gains of D2D multicast opportunities from adopting coded cooperative caching and DCHs selection schemes, but only obtains the local gain. In addition, for the task offloading, in the above works, it is assumed that the MEC has enough hardware and software resources to support computing tasks. In fact, this assumption is impractical as MEC computing power is limited, which cannot support all of the computing tasks.

Focusing on the above insufficiency, we study the computation and traffic offloading schemes in cache aided D2D multicast networks for high data rate and delay sensitive services. The main contributions of this paper are presented as follows.

- Propose a D2D cluster head selection scheme by jointly considering users' social attributes, available energy and storage of their devices, and the transfer rate from the BS to the user. In the scheme, we model the probability a user selected to be a DCH based on the Chinese Restaurant Process (CRP) and the weighted sum method, which is dependent on the influence factors.
- Propose a novel multicast-aware coded and cooperative caching scheme to alleviate the congestion of backhaul links and improve the quality of experience at peak hours. The scheme consists of a modified coded caching scheme on the end user level and a cooperative caching scheme on the DCH level. Based on the two schemes, an optimization problem to minimize the average energy consumed is formulated and is solved by a cooperation-based greedy caching algorithm (CBCA).
- Propose a joint computation offloading and resource allocation optimization scheme. The users' revenue maximization optimization problem is formulated which takes the user association, computation offloading strategy policy, uplink FD-DCH transmission power allocation, and computation resource scheduling into account. To solve the problem, we transfer it into two sub problems, named user allocation optimization problem (UAOP) and resource allocation optimization problem (RAOP), and the optimal resource allocation scheme is proposed. The effectiveness of the proposed schemes is demonstrated by simulation results.

The rest of this paper is organized as follows. In Section II, we present the system model. A D2D multicast cluster head selection scheme is proposed in Section III. The coded and cooperative caching scheme is presented in Section IV. The task offloading and resource allocation scheme is presented in Section V. The simulation results and analysis are presented soon afterwards in Section VI. Section VII concludes this paper and presents future work directions.

II. SYSTEM MODEL

The system model is presented in Fig.1. A macro-cell BS connects with the Internet via the core network which consists of the mobility management entity (MME) and the serving gateway or packet data network gateway (S/P GW). In the coverage area of a macro-cell BS, the users are grouped into multiple D2D multicast clusters according to geographical location. In each cluster, a user is selected as the DCH which saves some contents proactively and has ability to distribute traffic to the members in the same cluster via D2D multicast communication. When DUEs request for some contents, the DCH delivers the contents in a coded multicast way if it has cached them. Otherwise, the contents would be delivered by normal multicast through the cooperation of DCHs or BS. To achieve an efficient distributed caching, the clusters are in the same size and a member in one cluster cannot belong to another one.

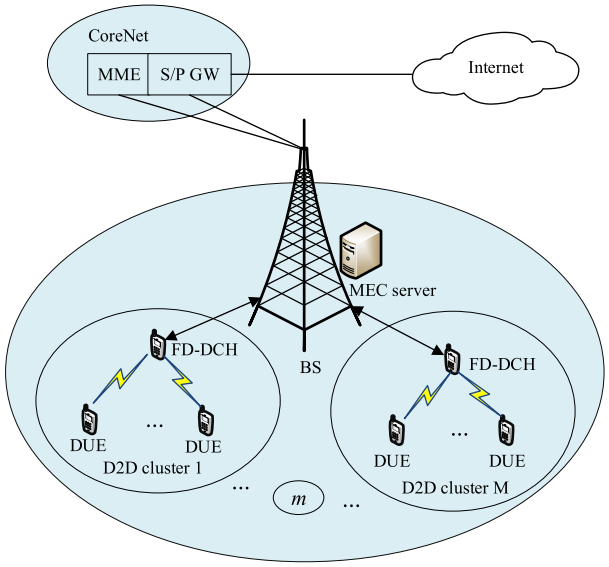


FIGURE 1. MEC-enabled D2D communication network.

Since different D2D multicast clusters have the same policy in cache placement phases and contents delivery phases,

we only take one cluster for analysis in this paper. As depicted in Fig.1, the BS acquires files from the content server via a high-speed backhaul link and storage these files in MEC. Assume there are N files with the same size of B bits, denoted by $\mathrm{F} = \{F_1, F_2, \cdots, F_N\}$. For file i, its requested probability p_i follows the Zipf distribution, which can be expressed as (1). The sum requested probability of all the files is 1, i.e. $\sum_{i=1}^{N} p_i = 1$.

$$p_j = \frac{\left(j^{-r}\right)}{\sum_{i=1}^{N}\left(i^{-r}\right)} \tag{1}$$

where the r denotes the popularity parameter.

In a cluster, there are N users and among them k users are equipped with heterogeneous local caches. The DCHs are selected according to the D2D multicast cluster head selection scheme proposed in Section III, denoted as $\mathbf{S} = \{1, 2, \cdots, S\}$. Once the DCH is selected, it caches M_{ch} files, while the DUEs only cache M_{cm} files. To be more efficient, it is assumed that the DCHs are able to share their cached contents with each other.

By deploying computing servers at the edge of the network, MEC system can enable the users to offload tasks and provide users the abundant wireless resources and vast computation capabilities. The process of the MEC-enabled D2D communication consists of two aspects:

- Content Caching
  For DUEs who require the data-hungry service, e.g. video or content downloading. In the placement phase, DCHs select a part of files to cache and simultaneously push some bits to DUEs in off-peak hours. In the delivery phase, according to whether the content to be shared is stored on the CHs or not, there are three different cases. First of all, if the request contents of DUEs have been cached in the associate DCH, coded multicast is applied for delivery contents from DCH to DUEs. Secondly, if the requested contents are stored in other DCHs instead of the associated DCH, the contents will be transmitted to the associated DCH via D2D link and then the associated DCH multicasts them to the DUEs. Last but not the least, if none of DCHs have stored the requested contents, BS will multicast the contents to DUEs.
- Computation Offloading
  For DUEs who require delay-sensitive tasks, e.g. Virtual Reality computing task, according to whether the status of the associate DCHs is idle or not, the computation task offloading can be categorized into two cases, local-direct D2D offloading and remote-MEC offloading. In the former case, if a DUE, which have a computation task, is able to find a idle DCH, its task can be partly offloaded to the DCH by D2D link. In the latter case, if all the requested DCHs are busy, a proportion of its task will be sent to the carefully selected associated DCH and then relayed to the MEC who will execute the task computation.

III. JOINT SOCIAL ATTRIBUTE, ENERGY AND RATE CLUSTER HEAD SELECTING SCHEME

D2D multicast plays an important role in improving system capacity, reducing transmission delay and improving resource utilization efficiency. In this section, a D2D multicast cluster head selection scheme is proposed. In a cluster, the DCH works as a relay to transmit contents to its members. If the DCH and DUEs distrust each other, they may be less interested in distributing and receiving the contents. If the DCH does not have enough energy and storage, the communication may interrupt. In addition, if the DCH and DUEs are located at the edge of the BS, the channel quality may be too bad to transfer contents with required quality. Therefore, users' social attributes, available energy and storage of their devices and the transfer rate from the BS to DUEs are taken into consideration in the scheme.

To describe the possibility a certain DUE being selected as a DCH, CRP is adopted. CRP is a famous stochastic process which is widely used in nonparametric topic models. In cluster c_n, the probability of user j to be selected as the DCH is represented as (2).

$$P_j^{c_n} = \left[P^{c_n}\left(z_j = 1 \left| Z_{-j}, \alpha_w\right.\right), \cdots, P^{c_n}\left(z_j = m \left| Z_{-j}, \alpha_w\right.\right)\right] \tag{2}$$

where $m\,(m \geq 3)$ is the size of cluster c_n and $P^{c_n}(z_j = i \left| Z_{-j}, \alpha_w\right.)$ is the probability that DUE j selecting DUE i to be the DCH.

Then the selection probability matrix can be defined as follows.

$$P^{c_n} = \left[P_1^{c_n T}, P_2^{c_n T}, \cdots, P_m^{c_n T}\right] \tag{3}$$

where the i-th row represents the probability that DUE i being selected as the DCH by all DUEs. The sum of elements in the i-th row is defined as the selection probability of DUE i. The DUE who has the largest selection probability is selected to be the DCH.

In cluster c_n, the probability that DUE j selecting DUE i as the DCH can be calculated by (4).

$$P^{c_n}\left(z_j = i \left| Z_{-j}, \alpha_w\right.\right) = \begin{cases} \dfrac{w_{i,j}^{c_n}}{\sum_{l \neq j} w_{l,j}^{c_n} + \alpha_w}, & i \neq j \\ \dfrac{\alpha_w}{\sum_{l \neq j} w_{l,j}^{c_n} + \alpha_w}, & i = j \end{cases} \tag{4}$$

where α_w is the parameter of CRP and $w_{l,j}^{c_n}$ is the influence factor of DUE i on DUE j. The definition of $w_{l,j}^{c_n}$ is presented as (5).

$$w_{i,j}^{c_n} = w_S s_{i,j}^{c_n} + w_E e_{i,j}^{c_n} + w_R r_{i,j}^{c_n} \tag{5}$$

where $w_S + w_E + w_R = 1$. $s_{i,j}^{c_n}$,$e_{i,j}^{c_n}$ and $r_{i,j}^{c_n}$ represent social influence factor, energy influence factor and transfer rate influence factor of DUE i on DUE j, respectively. The influence factors are defined as follows.

A. SOCIAL INFLUENCE FACTOR

$s_{i,j}^{c_n} \in [0, 1]$ represents the social influence of DUE i on DUE j. By analyzing the social connections between DUEs, we propose a social familiarity factor between DUE i and DUE j, denoted as (6).

$$s_{i,j}^{'c_n} = \frac{1}{-\ln\left(a_{i,j}^{c_n}\right)} \tag{6}$$

where $a_{i,j}^{c_n} \in [0, 1)$ is the social familiarity between DUE i and DUE j. Greater $a_{i,j}^{c_n}$ means higher social familiarity. If $a_{i,j}^{c_n} < \frac{1}{2}$, DUE i and DUE j will refuse to establish D2D communication link with each other. By taking all DUEs in cluster c_n into account, a normalized social influence factor can be obtained as follows.

$$s_{i,j}^{c_n} = \frac{s_{i,j}^{'c_n}}{\sum_{l\neq j} s_{i,j}^{'c_n}} \tag{7}$$

where $\sum_{l\neq j} s_{i,j}^{'c_n}$ represents other DUEs' social influence on DUE i in c_n.

B. ENERGY INFLUENCE FACTOR

$e_{i,j}^{c_n} \in [0, 1]$ represents DUE i's social influence on DUE j in cluster c_n. The maximum available time for transmission is adopted to measure the energy influence of DUE i on DUE j, denoted as (8).

$$T_{i,j}^{c_n} = \frac{E_{i,j}^{c_n}}{P_{i,j}^{c_n} + P_0} \tag{8}$$

where $E_{i,j}^{c_n}$, P_0 and $P_{i,j}^{c_n}$ are the available energy of DUE i, the circuit power and the transmission power, respectively. $P_{i,j}^{c_n}$ is calculated by (9).

$$P_{i,j}^{c_n} = \frac{\sigma^2\gamma_0}{G_{i,j}^{c_n}} \tag{9}$$

where σ^2 and γ_0 are the noise power and the received signal noise ratio (SNR) threshold, respectively. To guarantee the transmission quality, the received SNR must be no less than γ_0. $G_{i,j}^{c_n}$ is the channel gain between DUE i and DUE j, denoted as (10).

$$G_{i,j}^{c_n} = \left|h_{i,j}^{c_n}\right|^2 d_{i,j}^{c_n - \alpha_h} \tag{10}$$

where $h_{i,j}^{c_n}$, α_h and $d_{i,j}^{c_n}$ are the Rayleigh fading, the path loss factor and the physical distance between i and j, respectively.

By substituting (9) and (10) into (8), the maximum transmission time from i to j can be derived as (11).

$$T_{i,j}^{c_n} = \frac{E_{i,j}^{c_n}\left|h_{i,j}^{c_n}\right|^2 d_{i,j}^{c_n-\alpha_h}}{P_{i,j}^{c_n} + P_0\left|h_{i,j}^{c_n}\right|^2 d_{i,j}^{c_n-\alpha_h}} \tag{11}$$

Greater $T_{i,j}^{c_n}$ means greater energy influence of DUE i on DUE j. By taking into account of all DUEs in cluster c_n, a normalized energy influence factor is given by

$$e_{i,j}^{c_n} = \frac{T_{i,j}^{c_n}}{\sum_{l\neq j} T_{i,j}^{c_n}} \tag{12}$$

where $\sum_{l\neq j} T_{i,j}^{c_n}$ represents other DUEs' energy influence on j.

C. TRANSFER RATE INFLUENCE FACTOR

$e_{i,j}^{c_n} \in [0, 1]$ represents the transfer rate influence of DUE i. If the BS sends data at a certain power, the transfer rate of DUE i can be calculated by (13).

$$R_{B,i}^{c_n} = W\log_2\left(1 + \frac{P_B G_{B,i}^{c_n}}{N_0}\right) \tag{13}$$

where $G_{B,i}^{c_n} = \left|h_{B,i}^{c_n}\right|^2 d_{B,i}^{c_n - \alpha_B}$ is the channel gain between the BS and DUE i. $d_{B,i}^{c_n}$, W, $h_{B,i}^{c_n}$ and α_B are the distance between DUE i and BS, the channel bandwidth, the Rayleigh fading and the path loss factor, respectively. The higher the transfer rate is, the greater influence of DUE i is. By taking all users in cluster c_n into account, a normalized transfer rate influence factor is derived as (14).

$$r_{i,j}^{c_n} = \frac{R_{B,i}^{c_n}}{\sum_{l\neq j} R_{B,l}^{c_n}} \tag{14}$$

where $\sum_{l\neq j} R_{B,l}^{c_n}$ represents other users transfer rate influence on DUE j.

By substituting (7), (12) and (14) into (5), the influence factor of DUE i on DUE j is presented as (15).

$$w_{i,j}^{c_n} = w_S\frac{s_{i,j}^{'c_n}}{\sum_{l\neq j} s_{i,j}^{'c_n}} + w_E\frac{T_{i,j}^{c_n}}{\sum_{l\neq j} T_{i,j}^{c_n}} + w_R\frac{R_{B,i}^{c_n}}{\sum_{l\neq j} R_{B,l}^{c_n}} \tag{15}$$

By substituting (15) into (4), we can obtain the probability to select DUE i. Then based on descending order of selection probability, the DUEs are sorted. If the storage of the user at the top is no less than the required storage space M_{ch}, the DUE will be selected as the DCH. If not, the DUE in the second place will be taken into consideration. Repeat in the above process until the DCH is selected out.

IV. MULTICAST-AWARE CODED AND COOPERATIVE CACHING SCHEME

To alleviate the congestion of backhaul links and improve the quality of experience at peak hours, a multicast-aware cooperative caching scheme for traffic offloading is proposed in this section. Firstly, the system model with caching in DCHs and DUEs is presented. Secondly, a modified coded caching scheme on the end user level and a cooperative caching scheme on the D2D cluster level are proposed. Based on the proposed schemes, a model of the average energy consumption is formulated. To minimize the consumed energy, a cooperation-based greedy caching algorithm is proposed.

A. CACHING SCHEMES IN DCHs AND DUEs

The system model with caching in DCHs and DUEs is presented in Fig. 2. In the model, both DCHs and DUEs are able to cache contents. When users have requests, they firstly ask their own DCHs. Let $c_{i,j}$ indicates whether the j-th file have been cached in the i-th DCH. Assume all the storage space of

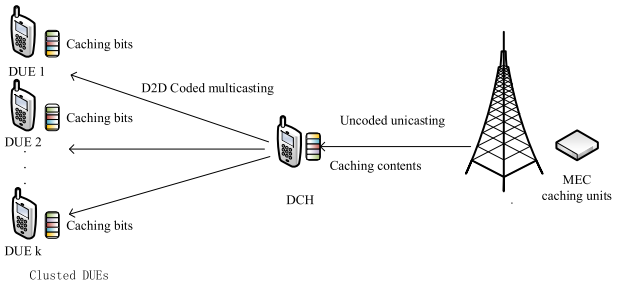


FIGURE 2. Illustration of content caching for D2D multicast network.

the DCHs is occupied, then we have $\sum_{j=1}^{N} c_{i,j} = M_{ch}, \forall i \in S$. Define c_j^s to indicate whether F_j has been cached in the DCHs, denoted as (16).

$$c_j^s = \begin{cases} 1, & \sum_{i=1}^{S} c_{i,j} \geq 1 \\ 0, & otherwise \end{cases} \tag{16}$$

If the DUEs' requests cannot be satisfied by its associated DCH, cooperation between DCHs is in need. The associated DCH requests the contents from the nearest DCH firstly and then delivers the contents to the requesting DUEs by means of D2D multicasting.

To cache distinguishing contents in each DCH, we divide the cached files into L file sets according to their popularity. It is noted that L cannot be determined before dividing the files. Assume the size of set l is N_l and the total size of all the sets is the size of contents requested by DUEs (M_{cm}), i.e. $\sum_{l=1}^{L} N_l = M_{cm}$. Denote the set of the files cached in a DCH by $Z = \{z_1, z_2 \cdots, z_{M_{ch}}\}$. In the set, the files are stored in descending order based on the requested probability, i.e. $p_{z_1} > p_{z_2} > \cdots > p_{z_{M_{ch}}}$. The files are divided with a factor of two to ensure a similar popularity in each set and to limit the decrement due to ignoring coding opportunities across different sets. Take the l-th file set for example to illustrate the process of files division. Assume the first file and the last file in set l are $F_{Z_1}^l$ and $F_{N_l}^l$, respectively. The files to be divided into set l must satisfy the following conditions: $P_{z_1}^l > p_{z_2}^l > \cdots > p_{z_{N_l}}$ and $2p_{z_{N_l}}^l \leq p_{z_1}^l \leq 2p_{z_{N_l+1}}^l$. The storage space in a DCH is associated with the cumulative probability of contents in each set. For set l, the amount of storage in need can be expressed as follows.

$$M_l = \min(M_{cm}^*, \sum_{j=1}^{N_l} p_{z_j}^l, N_l) \tag{17}$$

$$M_{cm}^* = \begin{cases} M_{cm} - \sum_{i=1}^{l-1} M_l, & l \succ 1 \\ M_{cm}, & else \end{cases} \tag{18}$$

where M_{cm} and M_{cm}^* denote the caching capacity of DUEs and the current unallocated amount of memory of the DUEs, respectively.

Based on the caching schemes in DCHs and DUEs, the transmission process can be described as follows. In the placement phase, each DUE randomly selects $100 * M_l/N_l$ percent of file set l from their DCH, where N_l determines the division of the cache placement of DCHs by a factor of two. In the deliver phase, if the DUEs in a cluster requests contents cached by their DCH, the DCH delivers the requested file-bits to every subset of request DUEs as [39] in which the users who request the files in the same file group would be satisfied by the coded multicasting to ensure all DUEs decode their desired content successfully. If the request files have not been cached in the associate DCH, the DUEs' demand will be fulfilled by the cooperation of DCHs or by the BS. The modified coded caching scheme is shown in Algorithm1.

Algorithm 1 A Modified Coded Caching Scheme for DUEs

Cache placement period:

for $s \in S$ **do**
  initialize $M_l = M_{cm}, L = 0$
  Cluster the files in CH s with a factor of two and upate the number of clusters L.
  for l from 1 to L **do**
    DUEs randomly cache $100 * M_l/N_l$ percent of file set l in cluster l.
    Update M_{l+1} according (17) and (18).
  end for
end for

Delivery period:

for $s \in S$, do **do**
  Get the set of DUEs $\mathbf{K}^*$ whose requests can be satisfied by DCH s.
  for $\mathbf{K_1}^* \subset \mathbf{K}^*$ **do**
    DCH s transmits the corresponding file-bits to DUEs in $K_1{}^*$.
  end for
end for

B. TRANSMISSION IN DELIVER PHASE

In the deliver phase, three cases are considered. In case 1, the DUEs can acquire contents from their home DCHs directly. In case 2, the home DCH asks other DCHs for help. In case 3, the BS delivers the requested contents to the DUEs directly. The details of the transmission schemes in these cases are presented as follows.

1) CODED D2D MULTICAST IN DCHs

As mentioned above, if the requested contents have already been cached in the home DCH, the DCH delivers the data using coded multicast. Generally, the probability of a DUE's request can be satisfied by its home DCH (e.g. DCH i) can be

express as follows.

$$P_i^{ch} = \sum_{j=1}^{N} p_j c_{i,j} \tag{19}$$

For DCH i, if the amount of data transmitted in coded multicast is R_c, the energy consumed can be obtained by (20).

$$E_i^{ch}(u) = R_c(M_u, M_{ch}, u) W_m \tag{20}$$

where u is the number of requests satisfied by the home DCH and W_m is the consumed energy to transmit a file from DCH i to its DUEs via D2D. R_c has been normalized by the file size.

In (20), $R_c(M_u, M_{ch}, u)$ can be derived as follows.

$$R_c(M_u, M_{ch}, u) = \sum_{l=1}^{L} K_l (1 - M_l/N_l) R_m \tag{21}$$

subject to:

$$\begin{cases} \sum_{l=1}^{L} K_l = u \\ \sum_{l=1}^{L} M_l = M_{cm} \\ \sum_{l=1}^{L} N_l = M_{ch} \end{cases} \tag{22}$$

where N_l and M_l denote the number of files in cluster l and its allocated memory size, respectively. R_m indicates the amount of data should be transmitted in each cluster, expressed as(23). K_l denotes the number of requests in cluster l, a random variable associated with the popularity of the files in cluster l. As it is complicated to deal with all the random requests in each DCH, we use its expectation to approximate, represented as (24).

$$R_m = min\{\frac{N_l}{K_l M_l}(1 - (1 - M_l/N_l)^{K_l}), \frac{N_l}{K_l}\} \tag{23}$$

$$K_l = u \sum_{j=1}^{N_l} p_{z_j^l} \tag{24}$$

2) COOPERATIVE CONTENTS SHARING AND MULTICAST

If the requested contents are cached by other DCHs rather than the home DCH, the contents will be transmitted from the nearest DCHs in the same group and finally be delivered by D2D normal multicast. It is assumed that there are u requests in DCH i and these requests can be satisfied by the cooperation of DCHs.

With the probability p_i^{cl} that a local DUE's request can be satisfied by the cooperation of DCHs, the energy consumption can be derived as (26).

$$p_i^{cl} = \sum_{j=1}^{N} p_j (1 - c_{i,j}) c_{i,j}^s \tag{25}$$

$$E_i^{cl}(u) = \sum_{j=1}^{N} 2(1 - (1 - p_j)^u)(1 - c_{i,j}) C_j^s W_m \tag{26}$$

where W_m denotes the energy consumption to transmit contents from one DCH to the other DCH via D2D unicast in the same BS group.

3) BS ASSISTED MULTICAST

If the requests can't be satisfied by both home DCH and other DCHs in the same group, the BS acquires the contents from the server and then directly delivers them to DUEs using multicast.

The probability that a DUE's request can be satisfied by the cooperation of BS is presented as (27).

$$P_i^{BS} = \sum_{j=1}^{N} p_j (1 - c_{i,j}) c_j^s \tag{27}$$

Then the energy consumption for transmitting contents to u DUEs in cluster i with the cooperation of BS can be expressed as (28).

$$E_i^{bs}(u) = \sum_{j=1}^{N} (1 - c_{i,j}^s)(1 - (1 - p_j)^u)(W_s) \tag{28}$$

where W_s denotes the energy consumption for transmitting a content from the BS to the DUE.

C. ENERGY CONSUMPTION MINIMIZATION

In a cluster, assume k_1, k_2 and k_3 ($k_1, k_2, k_3 \in [0, k]$) are the number of DUEs in the three delivery cases, respectively. The combination of these three variables is denoted as k^*, i.e. $k^* = \{k_1, k_2, k_3\}$. Define $\mathcal{K}$ to denote the set of k^* with all possible values i.e. $k^* \in \mathcal{K}$. Then in cluster i, the probability of k^* can be expressed as (29).

$$P_i(k^*) = (p_i^{ch})^{k_1} (P_i^{cl})^{k_2} (1 - P_i^{ch} - P_i^{cl})^{k_3} \tag{29}$$

subject to:

$$k_1 + k_2 + k_3 = K \tag{30}$$

where K is the number of DUEs.

If there are S DCHs in a cluster, the energy consumption minimization problem of a cluster can be modeled as (31).

$$\begin{aligned} \text{Minimize } E_{total} = \frac{1}{SK} \sum_{i=1}^{S} \sum_{k^* \in \mathcal{K}} P_i(k^*) \{E_i^{sm}(k_1) \\ + E_i^{cm}(k_2) + E_i^{SBS}(k_2)\} \\ \text{subject to: (22) and (30).} \end{aligned} \tag{31}$$

In order to minimize the total energy consumption, a low complexity greedy-based caching algorithm is proposed, presented as Algorithm 2.

When initiating, the files are cached in each DCH based on their popularity under some constraints, which is the optimal scheme without cooperation between DCHs. To be fair for each DCH in coordination, generate a DCHs sequence randomly in each iteration. During the iteration, the cache placement scheme with the minimized energy consumption is selected as the optimal one. When all iterations complete,

Algorithm 2 Cooperation-Based Greedy Caching Algorithm(CBCA)

Input: $N, M_{cm}, M_{ch}, k, S, W_m, W_s$
Output: $C_{S\times N}, E_{total}$

```
Initialize C_{S×N}
Compute E_total according to (31) and define E_pre = E_total
for j = 1 to j = N do
    Generate a sequence for DCHs randomly, denoted as
    S ≜ {DCH_1, DCH_2, ··· , DCH_S} in a D2D cluster
    for i = 1 to i = S do
        Change content F_j and update the cache placement
        as C_{i×N} = C†_{i×N} and its E_total
        if E_total ⩽ E_pre then
            E_pre = E_total, C_{i×N} = C†_{i×N}
        end if
    end for
end for
E_total = E_pre
return C_{S×N}, E_total.
```

the sub-optimal cooperation cache placement solution will be output.

V. ENERGY EFFICIENT TASK OFFLOADING AND RESOURCE ALLOCATION

To optimize the system revenue, FD technology [36], [40] is integrated in the MEC-enabled D2D Network. The BS is equipped with a MEC server and connects to the core network via wired optical fibers as shown in Fig.3. In this section, we formulate the user association, the task offloading strategy, FD uplink power allocation and computing resource allocation as an optimization problem.

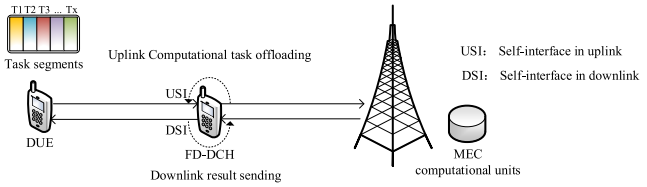


FIGURE 3. Illustration of computational task offloading for D2D multicast network.

A. SYSTEM MODEL WITH MEC

The DCHs with FD antennas (FD-DCHs) connect to the DUEs and the BS via wireless links. These FD-DCHs are able to help DUEs to access network and relay tasks to the MEC server on BS. It is assumed that tasks are divisible and can be processed at both local DUEs and edge MEC at the same time, which is more practical compared with [36]. The set of DUEs and FD-DCHs are denoted by $\mathcal{K} = \{1, 2, \cdots, K\}$ and $\mathcal{M} = \{1, 2, \cdots, M\}$ respectively. Each DUE in the D2D clusters has one task to process. We denote $x_{i,m} \in \{0, 1\}(\forall i \in K, \forall m \in M)$ as the associated decision of DUE i. Specifically, we assume $x_{i,m} = 1$ if DUE i is associated with FD-DCH m, and $X = \{x_{i,m}\}$ indicates the associate decision of DUE. In our setting, tasks can be divided and partly relayed to the MEC (e.g., Data analysis task in VR). We define $L_i = (\sigma_i, s_i, T_i)$ as the task of DUE i, where σ_i (in CPU cycles per Mbits) denotes total computing resource required by the task, s_i (in bits) indicates the data size of the task, T_i is the maximal tolerable delay of the task. Moreover, it is assumed that the required computational resources and data size of a task are proportional to its segmented ratio, such as $L_i(\theta) = (\theta\sigma_i, \theta s_i, T_i)$, where θ indicates the segmented factor. It's worth noting that the DUEs cannot complete the whole tasks they required within the maximal tolerable delay due to the limited computational ability. Therefore they should relay a part of segments of the task to the MEC.

The computation offloading steps of the MEC-enabled D2D network are summarized as follows.

- Firstly, DUEs sent a proportion of the tasks to their associated FD-DCHs.
- Secondly, the tasks are further relayed to th BS at the same time with the same frequency band of the forward link by the FD-DCHs. The offload proportion of task for user i is denoted by $o_i \in [0, 1]$. Specially, $o_i = 1(o_i = 0)$ means the task will be totally offloaded to the MEC (local of DUEs) for processing.

1) COMMUNICATION MODEL

We assume DUEs and FD-DCHs work on the orthogonal spectrums both in the forward link and backward link. Therefore there is no mutual interference with each other. The bandwidth resources of forward link and backward link are the same, denoted as B. For the forward link from DUE $i \in \mathcal{K}$ to the associated FD-DCHs $n \in \mathcal{M}$, the achievable data rate can be calculated by (32).

$$R^a_{i,m} = B\log(1 + \frac{p_i g_{i,m}}{\sigma^2 + SI}) \tag{32}$$

where p_i denotes the transmission power of DUE i. $g_{i,m}$ means the channel gain from the DUE i to FD-DCH m. SI denotes the self-interference of the FD antenna, and $SI = Ib_{i,m}p_m$, where I denotes the residual SI gain. p_m and $b_{i,m}$ are the allocated power and the power ratio in FD-DCH m for task relaying respectively [41]. We regard SI as a constant which can limit the self-interference within a small range based on interference cancellation schemes.

Similarly, the data rate of the backward link from FD-DCH m to BS can be presented as (33).

$$R^b_{i,m} = B\log(1 + \frac{b_{i,m}p_m g_m}{\sigma^2}) \tag{33}$$

where p_m is the maximal transmission power of FD-DCH m. $b_{i,m} \in (0, 1]$ denotes the power factor allocated for task offloading of DUE i. g_m means the channel gain between the FD-DCH m to the BS.

Referred to [42], the whole available data rate of the uplink is denoted as (34).

$$R_{i,m} = \min(R^a_{i,m}, R^b_{i,m}) \tag{34}$$

Due to the transmission rate in the input link is higher than the output link for FD communication, i.e. $R^a_{i,m} \geq R^b_{i,m}$, $R^s_{i,m} = R^a_{i,m}$, where $R^s_{i,m}$ denotes the total offload data rate to the MEC.

Note that the downlink transmission delay is not taken into consideration in this paper because the result of task processing is very small compared to that in the uplink transmission [43].

2) COMPUTION MODEL

In the computation model, w^l_i denotes the local computational capability (CPU cycles per Mbits) of DUE i, and σ_i (in CPU cycles per Mbits) denotes total computing resource required for the task of DUE i. The local execution latency of the whole own task by local computing is denoted as (35).

$$t^l_i = \frac{\sigma_i}{w^l_i} \tag{35}$$

The execution latency of the whole task from DUE i by FD-DCH m computing is defined as

$$t^l_{i,m} = \frac{\sigma_i}{w^l_m} \tag{36}$$

where w^l_m denotes the local computational capability (CPU cycles per Mbits) of FD-DCH m.

Similarly, the execution latency of the whole task from DUE i by the MEC server computing can be presented as

$$t^e_i = \frac{\sigma_i}{a_i w^e} \tag{37}$$

where w^e and a_i denote the computational capability of the MEC server and the computing factor assigned to the task of DUE i, respectively.

The transmission latency of the task $L_i = (\sigma_i, s_i, T_i)$ from DUE i to FD-DCH m can be calculated by

$$t^d_{i,m} = \frac{s_i}{R^a_{i,m}} \tag{38}$$

The transmission latency of the task $L_i = (\sigma_i, s_i, T_i)$ from DUE i to MEC through the FD-DCH m can be calculated by

$$t^s_{i,m} = \frac{s_i}{R^s_{i,m}} \tag{39}$$

As mentioned above, tasks are able to be divided. Let o^s_i and o^d_i denote the proportion of the offloading task to MEC or a FD-DCH from DUE i respectively. Therefore, the local processing time of FD-DCH and MEC can be given by

$$t^l_{i,m} = (1 - o^d_i)t^l_i \tag{40}$$

$$t^{ls}_{i,m} = (1 - o^s_i)t^l_i \tag{41}$$

The total execution latency of the offloading task from DUE i to FD-DCH m and MEC can be presented as (42) and (43), respectively.

$$\overset{*}{t}{}^d_{i,m} = o^d_i(t^d_{i,m} + t^l_{i,m}) \tag{42}$$

$$\overset{*}{t}{}^s_{i,m} = o^s_i(t^s_{i,m} + t^e_i) \tag{43}$$

It assumed that tasks can be handled at both local MT and MEC server at the same time, therefore the total complete time of task L_i should be the larger value of local processing time and MEC edge or FD-DCH total execution time, which can be calculated as (44) and (45), respectively.

$$t^{MEC}_i = \max(t^{ls}_{i,m}, \overset{*}{t}{}^s_{i,m}) \tag{44}$$

$$t^{D2D}_i = \max(t^l_{i,m}, \overset{*}{t}{}^d_{i,m}) \tag{45}$$

B. REVENUE MAXIMIZATION

To maximize the revenue of D2D users in the cluster, we model and solve the revenue maximization problem in this section. Firstly, define the utility function as the subtraction between service revenues and costs. Based on the utility function, the revenue maximization problem is formed. Secondly, reformulate the original problem by decomposing it into two sub optimization problems. Finally, the optimization problem is solved by applying Greedy Algorithm.

1) UTILITY FUNCTION AND PROBLEM FORMATION

The utility function is expressed as the subtraction between service revenues and costs. The service revenues can be denoted by the benefits rooted in the intrinsic value of tasks including the profits of data size and the computational resource of tasks. The costs are consisted of the price of the allocated computational resource and the consumed power for data transmission for MEC operators. The detailed formulation of utility for task $L_i = (\sigma_i, s_i, T_i)$ can be represented as (46).

$$\begin{aligned} u_{i,m} = x_{i,m}[&d_m(\kappa s_i + \rho\sigma_i - \eta b_{i,m} P_m - \beta a_i w^e) \\ &+ (1 - d_m)(\kappa s_i + \rho\sigma_i)(1 - o^d_i)] \end{aligned} \tag{46}$$

where κ denotes the revenue coefficient per unit of offloading data size. η denotes the revenue coefficient per unit of power of FD-DCHs. ρ and β represent price coefficient per unit of computational resource and price coefficient of allocated computational capability per unit time respectively.

$$\begin{aligned} &maximize\ U = \sum_{i=1}^{K}\sum_{m=1}^{M} u_{i,m} \\ &s.t.\ C1 : t_i \leq T_i \\ &\qquad C2 : \sum_{m=1}^{M} x_{im} \leq 1 \\ &\qquad C3 : \sum_{i=1}^{K} x_{im} \leq N, \quad \forall m \in \mathcal{M} \\ &\qquad C4 : \sum_{i=1}^{K} b_{i,m} \leq 1, \quad \forall m \in \mathcal{M} \\ &\qquad C5 : \sum_{i=1}^{K} a_i \leq 1, \quad \forall i \in \mathcal{K} \\ &\qquad C6 : R^a_{i,m} \geq R^b_{i,m}, \quad \forall i \in \mathcal{K},\ \forall m \in \mathcal{M} \end{aligned} \tag{47}$$

where the objective function computes the maximum of the revenue of DUEs through the deployment of task caching and offloading. The first constraint C1 ensures all the tasks are parallely computed at both the MEC server and local DUEs. Additionally, all the tasks must be completed within their maximal tolerable delays. The constraint C2 is proposed to guarantee each DUE can access only one FD-DCH at most. The constraint C3 requires that the number of accessed DUEs cannot exceed the maximum access number of each FD-DCH, where N means the maximum access quantity of DUEs. The constraint C4 means the allocated power of each FD-DCH which cannot exceed its maximum transmission power. The constraint C5 states that the allocated computational resource should be less than the maximum computational capability of MEC. The constraint C6 limits the data rate of backward link which shouldn't be more than the frontward link for each DUE.

2) PROBLEM TRANSFORMATION

As $x_{i,m}$ ($\forall i \in \mathcal{K}$) is a binary variable, the objective function (47) is non-convex. Moreover, the constraint C1 includes a maximization function which is difficult to deal with. The original problem is a mixed discrete and non-convex optimization problem, which is NP-hard. Some transformation is necessary to make the objective function to be solvable in polynomial time. In this section, we reformulate the original problem by decomposing it into two sub optimization problems which are named user allocation optimization problem (UAOP) and resource allocation optimization problem (RAOP).

Based on a fixed X, the RAOP can be expressed as

$$\begin{aligned} &\underset{X}{maximize}\ U(A,B,O)=\sum_{i=1}^{K}\sum_{m=1}^{M}u_{i,m} \\ &s.t.\ C1,\ C2,\ C5,\ C6 \end{aligned} \tag{48}$$

Given $\underset{X}{Z}(A,B,O) = -\underset{X}{U}(A,B,O)$, (43) can be reformulated as:

$$\begin{aligned} &\underset{X^*}{min}\ Z(A,B,O)=-x_{i,m}[d_m(\kappa s_i+\rho\sigma_{\mathrm{i}}-\eta b_{i,m}P_m \\ &\quad-\beta a_i w^e)+(1-d_m)(\kappa s_i+\rho\sigma_{\mathrm{i}})(1-o_i^d)] \\ &s.t.\ C1,\ C2,\ C5,\ C6 \end{aligned} \tag{49}$$

Proposition 1: For task L_i which should be offloaded to MEC or FD-DCH, the optimal offload ratio o_i^{best} is $1-\frac{T_i w_i^l}{\sigma_i}$, and the optimal total execution time of computing offloading is T_i.

Proof: Firstly, we analyze the computing offloading in the MEC. According to C1, we can obtain

$$\begin{aligned} t_i^{MEC} &= \max(t_{i,m}^{ls}, , \overset{*}{t}_{i,m}^{s}) \\ &= \max\{\frac{(1-o_i^s)\sigma_i}{w_i^l}, \frac{o_i^s s_i}{B\log(b_{i,m}p_m h_{i,m}^b)}+\frac{o_i^s\sigma_i}{a_i w^e}\} \\ &\le T_i \end{aligned} \tag{50}$$

where $h_{i,m}=\frac{g_m}{\sigma^2}$.

Due to $o_i^s \in [0,1]$, there must exist an o_i^s* ($o_i^s* \in [0,1]$ which satisfies (51).

$$\frac{(1-o_i^s)\sigma_i}{w_i^l}=\frac{o_i^s s_i}{B\log(1+b_{i,m}p_m h_{i,m}^b)}+\frac{o_i^s\sigma_i}{a_i w^e} \tag{51}$$

where o_i^s decreases from o_i^s* to 0, $t_{i,m}^{ls}$ increases and $\overset{*}{t}_{i,m}^{s}$ decreases. The optimal value o_i^{best} must meet the condition $t_{i,m}^{ls}=T_i$, because $\frac{\sigma_i}{w_i^l} \succ T_i$ and $t_{i,m}^{ls} \ge \overset{*}{t}_{i,m}^{s}$. Then we get

$$o_i^{best}=1-\frac{T_i w_i^l}{\sigma_i} \tag{52}$$

Combined with (39), the total execution latency of offloading can be rewritten as follows.

$$\overset{*}{t}_{i,m}^{s}=\frac{{o_i}^{best} s_i}{B\log(1+b_{i,m}p_m h_{i,m}^b)}+\frac{{o_i}^{best}\sigma_i}{a_i w^e} \tag{53}$$

From (53), we can find that the larger $\overset{*}{t}_{i,m}^{s}$ means the larger allocated transmission power or computational resource. A rise in $\overset{*}{t}_{i,m}^{s}$ will increase the costs instead of the benefits of DUEs. Therefore the optimal value of $\overset{*}{t}_{i,m}^{s}$ must meet $t_{i,m}^l=\overset{*}{t}_{i_m}^{s}=T_i$.

When offloading the computing to FD-DCHs, the benefits from completing tasks decrease based on (49) because the ratio of offloading increases. ■

Furthermore, we define $\xi_{i,m}=R_{i,m}^s/B$, and by integrating (47) and (48) and substituting related variables, we can obtain

$$a_i(\xi_{i,m})=\frac{-B{\sigma_i}^2\xi_{i,m}+Bw_i^l T_i\sigma_i\xi_{i,m}}{w^e(s_i\sigma_i-BT_i\sigma_i\xi_{i,m}-s_i T_i w_i^l)} \tag{54}$$

$$b_{i,m}(\xi_{i,m})=\frac{2^{\xi_{i,m}}-1}{h_{i,m}} \tag{55}$$

The problem (44) can be rewritten as follows.

$$\begin{aligned} &\underset{X}{minimize}\ Z^*(A,B,O)=-x_{i,m}\{d_m[\kappa s_i+\rho\sigma_{\mathrm{i}} \\ &\quad-\eta b_{i,m}(\xi_{i,m})P_m-\beta a_i(\xi_{i,m})w^e] \\ &\quad+(1-d_m)(\kappa s_i+\rho\sigma_{\mathrm{i}})(1-o_i^d)\} \\ &s.t.\ C7:\sum_{i=1}^{K}\frac{2^{\xi_{i,m}}-1}{h_{i,m}}\le 1,\quad \forall m\in M \\ &\quad C8:\sum_{i=1}^{K}a_i(\xi_{i,m})\le 1 \\ &\quad C9:\xi_{i,m}\le\log(1+p_i h_{i,m}^a) \end{aligned} \tag{56}$$

3) PROBLEM SOLUTION

Firstly, the resource allocation optimization problem is discussed. The second derivative of $Z_{i,m}$ for $\xi_{i,m}$ can be calculated as

$$\frac{\partial^2 Z_{i,m}}{\partial^2\xi_{i,m}}=\frac{2s_i(B)^2T_i(\sigma_i-\omega_i^l T_i)^2}{\omega^e(B\xi_i T_i\sigma i-s_i\sigma_i+s_i\omega_i^l T_i)^3}+\frac{(\ln 2)^2 2^{\xi_{i,m}}}{h_{i,m}} \tag{57}$$

Derivation from Proposition 1, when $t_{i,m}^{s*} = T_i$, we can obtain

$$s_i\omega_i^l T_i = (1 - o_i^{best})\sigma_i s_i \tag{58}$$

Similarly, when $t_{i,m}^s = T_i - t_i^e$ and $t_i^e \succ 0$, it is holding that:

$$\begin{aligned} B\xi_{i,m}\sigma_i T_i &= R_i^s T_i \sigma_i \\ &\succ R_i^s t_{i,m}^s \sigma_i \end{aligned} \tag{59}$$

Moreover, when $R_i^s t_{i,m}^s = o_i^{best} s_i$, by integrating (54) and (55), we have

$$B\xi_{i,m}T_i\sigma_i + s_i\omega_i T_i - s_i\sigma_i \succ (1 - o_i^{best})\sigma_i s_i + o_i^{best} s_i \succ 0 \tag{60}$$

Thus we can derive $\frac{\partial^2 Z_{i,m}}{\partial^2 \xi_{i,m}} \succ 0$ and (56) is convex. As the second order derivative of (56) is strictly negative, the optimization problem can be solved by adopting the Karush-KuhnTucker (KKT) conditions. The Lagrange function of (56) can be expressed as

$$\begin{aligned} L(\xi, \lambda, \mu) = &\sum_{i=1}^{K}\sum_{m=1}^{M} Z_{i,m}(\xi_{i,m}) + \lambda(\sum_{i=1}^{K}\frac{2^{\xi_{i,m}}-1}{h_{i,m}} - 1) \\ &+ \mu[\sum_{i=1}^{K} a_i(\xi_{i,m}) - 1] \\ &+ \tau_i[\xi_{i,m} - B\log(1 + p_i h_{i,m}^a)] \end{aligned} \tag{61}$$

For $\forall i \in K$, $\forall m \in M$, the KKT conditions are as follows.

$$\frac{\partial L}{\partial \xi_{i,m}} = \frac{(\eta p_i + \lambda)2^{\xi_{i,m}}\ln 2}{h_{i,m}} + (\beta w^e + \mu)a_i(\xi_{i,m}) + \tau_i = 0 \tag{62}$$

$$\lambda(\sum_{i=1}^{i}\frac{2^{\xi_{i,m}} - 1}{h_{i,m}} - 1) = 0 \tag{63}$$

$$\mu(\sum_{i=1}^{K}\frac{-B\sigma_i{}^2\xi_{i,m} + Bw_i^l T_i\sigma_i\xi_{i,m}}{w^e(s_i\sigma_i - BT_i\sigma_i\xi_{i,m} - s_i T_i w_i^l)} - 1) = 0 \tag{64}$$

$$\tau_i[\xi_{i,m} - B\log(1 + p_i h_{i,m}^a)] = 0 \tag{65}$$

where $a_i{}'(\xi_{i,s}) = \frac{JV}{(J - C\xi_{i,m})^2}$, in which $J = w^e s_i\sigma_i - w^e s_i T_i w_i^l$, $C = w^e BT_i\sigma_i$, and $V = B^b w_i^l T_i\sigma_i - B\sigma_i{}^2$.

Let $[y]^+ = \max\{y, 0\}$, combined with the constraints (58)-(60), the Lagrange multipliers update as below.

$$\lambda(t+1) = \left[\lambda(t) + \delta(t)\left(\sum_{m=1}^{M}\frac{2^{\xi}-1}{h_{i,m}} - 1\right)\right]^+ \tag{66}$$

$$\mu(t+1) = \left[\delta(t)\left(\sum_{i=1}^{K}\frac{-B\sigma_i{}^2\xi_{i,s} + Bw_i^l T_i^{\max}\sigma_i\xi_{i,s}}{w^e(s_i\sigma_i - BT_i^{\max}\sigma_i\xi_{i,m} - s_i T_i^{\max} w_i^l)} - 1\right) + \mu(t)\right]^+ \tag{67}$$

$$\tau_i(t+1) = \left[\tau_i(t) + \delta(t)(\xi_{i,m} - \log(1 + p_i h_{i,m}^a))\right]^+ \tag{68}$$

where t denotes the current times of iteration and $\delta(t)$ represents the step of the t-th iteration. By utilizing the KKT condition, the optimal resource allocation solution can be found.

The optimal ξ_{im} can be obtained from (62)-(65). According to (49) and (50), the optimal a_i^{best} and $b_{i,m}^{best}$ will be obtained.

To solve the allocation optimization problem, we can adopt the method above to optimize resource allocation scheme. Let A_X^{opt} and B_X^{opt} denote the decision of allocated power and computational resource in a access scheme of X respectively. The allocation problem for the left DUEs is a 0-1 nonlinear optimization problem, which is NP-complete. There have been quite a few existing meta-heuristic algorithms to solve such NP-hard problems, such as Ant Colony Optimization (ACO), Genetic Algorithm (GA), Simulated Annealing (SA) and Greedy Algorithm. Among these algorithms, the Greedy Algorithm is effective and with low complexity in approximating global optimal solution by searching a series of locally optimal alternatives. Therefore, we adopt the Greedy Algorithm to solve the user allocation optimization problem in this paper. The details of the algorithm are presented in Algorithm 3.

Algorithm 3 Process of the Greedy-Based Algorithm

Require:
set of DUEs $\mathcal{K} = \{1, 2, \cdots, K\}$;
Maximum iterative number I;
The current busy status of FD-DCHs $\mathcal{D} = \{d_i, d_2, \cdots, d_M\}$
The tasks of DUEs $L_i = (\sigma_i, s_i, T_i)$;
$B\ p_i, p_m, w^e, \kappa, \nu, \rho, \eta, \beta$.

Ensure:
Determined resource allocation scheme A^*, B^*, O^*;
Determined user allocation schemeX^*.

```
Initialize μ, λ, τ
for i = 1 to i = K do
    Caculate the optimal offloading ratio o_i by (52);
    Set U_p = 0;
    for m = 1 to m = M do
        Set x_{i,m} = 1;
        for t = 1 to t = I do
            Set δ(t) = 1/(10 + t);
            Calculate ξ_{i,m} by (62);
            Update Lagrange multiplier μ, λ, τ by (66)-(68);
        end for
        Calculate current optimal a*_{i,m}, b*_{i,m} by (54) and (55)
        and the current utility U_c by (47);
        if U_c ≤ U_p then
            Set x_{i,m} = 0;
        else
            Set U_p = U_c, a_{i,m} = a*_{i,m}, b_{i,m} = b*_{i,m};
        end if
    end for
end for
```

VI. PERFORMANCE EVALUATION

In this section, we evaluate the performance of the proposed schemes by MATLAB. The simulation setup and performance analysis are presented as follows.

A. SIMULATION SETUP

The simulation topology is depicted in Fig. 1, in which the BS is loacated in the center of a 120×120 m^2 area and the FD-DCHs are distributed around it uniformly. The clusters are grouped in the same size according to their location. In the system, there are 40 DUEs and 4 FD-DCHs. The transmission power of DUEs, p_i, is set to be 10 dBm, while the transmission power of FD-DCHs, p_m, is set as 20 dBm. For the wireless links, the channel power gain of UEs follows the Gaussian distribution $CN(10; 5)$. In addition, the thermal noise power is set to be -100 dbm. Other simulation parameters are shown in table 1, in which most of the settings and values are set referred to [44].

TABLE 1. Summary of the simulation parameters.

Parameters	Values
Number of requests in a D2D cluster (K)	20∼90
Number of DCHs (S)	1∼7
Energy consumption to transmission a content by the cooperation ofSBS (W_s)	16.776 J
Energy consumption of D2D link (W_m)	6.711 J
Content library size (N)	300
Caching capacity of uers (M_{cm})	6
Caching capacity of CH (M_{ch})	$10 \sim 100$
Popularity parameter (r)	0.2∼2
Revenue coefficient per unit of offloaded data size(κ)	0.05$/Mbit
Revenue coefficient of power (η)	0.1$/mW
Price coefficient of allocated computational resource (ρ)	0.05$/Gcycles/s
Price coefficient per unit of total leaving required computational resource (β)	2$/Gcycles
Bandwith of backward and forward link for each UE B	0.5 MHz

B. SIMULATION RESULTS AND ANALYSIS

In order to analyze the performance of the content offloading scheme based on the cluster selection scheme in Section III, we apply the following three different schemes to make a comparison.

- *Random Caching:* Each DCH in a group random caches the contents until their storage spaces are full;
- *CCBCS [44]:* The caching capacity of each DCH in a group is divided into two parts, the first parts cache the most popular contents, while the other cache the contents not so popular;
- *PBCS [45]:* Every DCH caches the most popular contents until their storge spaces are full.

The computation offloading scheme based on the cluster selection scheme in Section III is compared with the following three different schemes.

- Proposed scheme w.o. FD: The proposed scheme with optimal task execution and resource allocation, completed within the limit of maximal tolerable delay, which is without (w.o.) the adoption of FD;
- RECA w. FD: Random ratio of execution and optimal resource allocation, completed within the limitation of maximal tolerable delay, which is with (w.) the adoption of FD;
- CECA w. FD: Cloud execution and optimal resource allocation, completely within the limitation of maximal tolerable delay, which is with (w.) the adoption of FD.

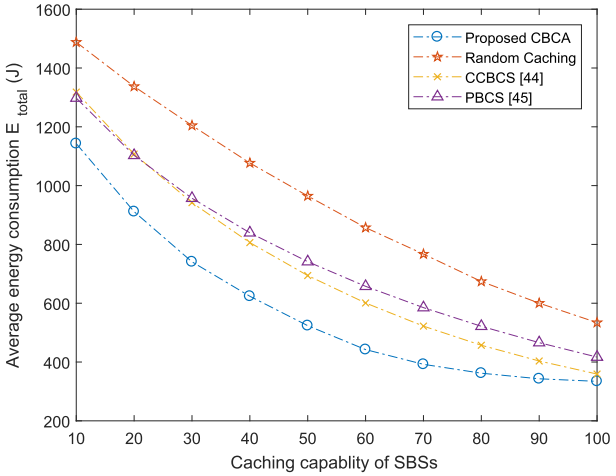


FIGURE 4. The average energy consumption VS the caching capability of DCH, with $S = 2$, $K = 100$, $r = 0.5$.

Fig. 4 compares the average energy consumption with different caching capablity of DCHs. It can be noticed that the proposed CBCA is better than the other schemes, since it combines the cooperation caching and the coded caching, trading off the cost of contents delivery from the backhual, the D2D clusters and the associated DCH. Furthermore, the random caching with cooperation and CCBCS perform better than PBCS as the caching capability increases. The possible reason is that PBCS doesn't take advantage of the cooperation of DCHs in heterogeneous networks. Furthermore, the average of energy consumption decreases as the caching capabilities of DCHs increase. The descending trend can be explained as below. With the increase of caching capability, more contents can be cached in the local DCHs, so that much more requests would be satisfied by the local coded multicast or the cooperation of the D2D clusters, which finally saves energy consumption.

The average energy consumption with different number of average requests is presented in Fig. 5. Obviously, the more requests are, the higher energy consumption will be. The reason is that the traffic increase linearly with the requests, which lead to a linear increase of energy consumption. As depicted in Fig. 5, the proposed scheme performs the best no matter how K changes compared with other schemes.

Fig. 6 depicts the influence of DCH group size on the average energy consumption. With the group size increasing, the energy consumption of all the schemes except PBCS decreases. Obviously, through cooperation of DCH, more

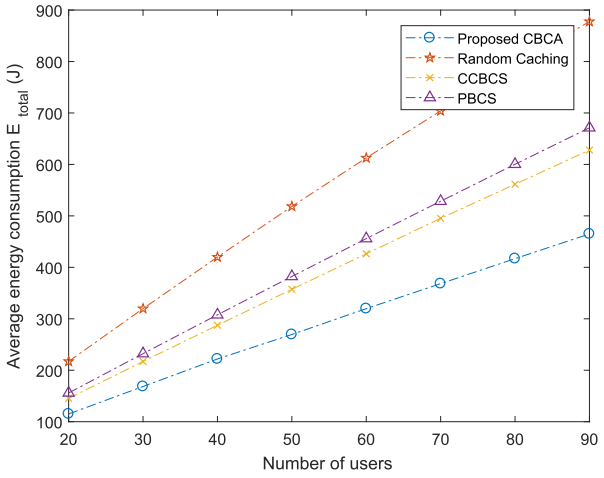


FIGURE 5. The average energy consumption VS the number of users k, with $S = 2$, $M_S = 50$, $r = 0.5$.

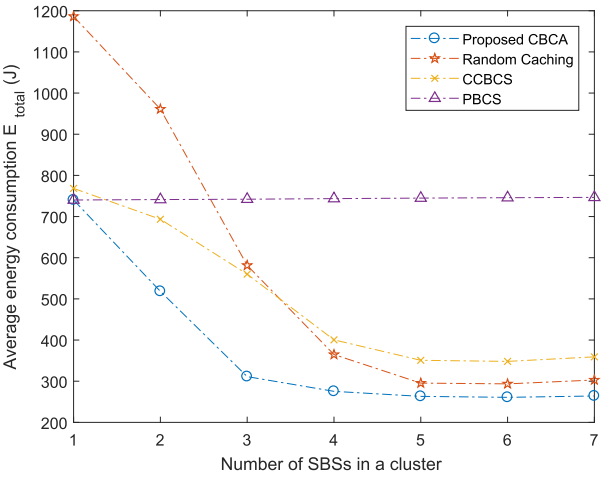


FIGURE 6. The average energy consumption VS the number of DCHs in a cluster, with $M_S = 50$, $k = 100$, $r = 0.5$.

traffic will be offloaded, which reduces energy consumption. Moreover, it can be observed that the proposed scheme can save the most energy compared with the other schemes.

Fig.7 presents the performance with r changing. As r increases from 0.2 to 2, the energy consumption decreases. Even with large r, the proposed scheme still has a better performance than other schemes.

Fig.8 presents the utility and costs of DUEs achieved with different ratio of computational costs to power costs per unit. In the simulation, the power costs per unit ρ is fixed and the computational costs per unit κ changes. The simulation results show that the utility of DUEs decreases with ratio increasing. That is because the total cost increases as the computational costs per unit increases.

Fig. 9 shows the influence of average input task data size on the utility of DUEs achieved with different schemes. As depicted in Fig.9, the utility of DUEs increases as the average data size increases at the beginning. Afterwards, the growth rate of utility slows and even experiences a negative growth like the scheme without FD. That can be

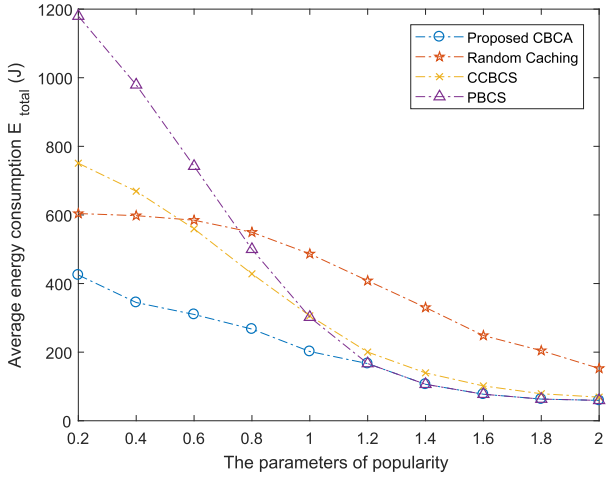


FIGURE 7. The average energy consumption VS the popular parameter r, with $S = 3$, $M_S = 50$, $k = 100$.

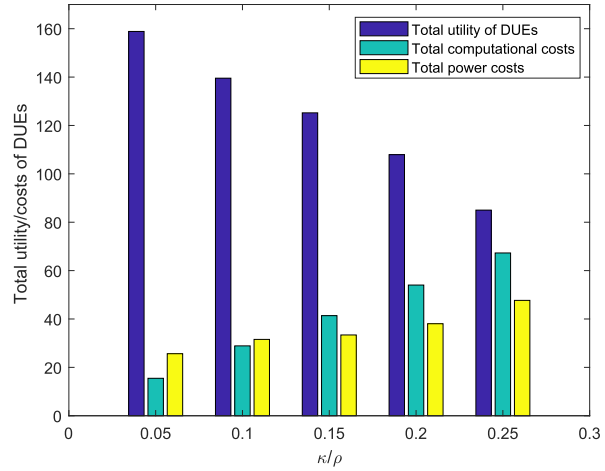


FIGURE 8. The influence of ratio of computational costs to power costs per unit.

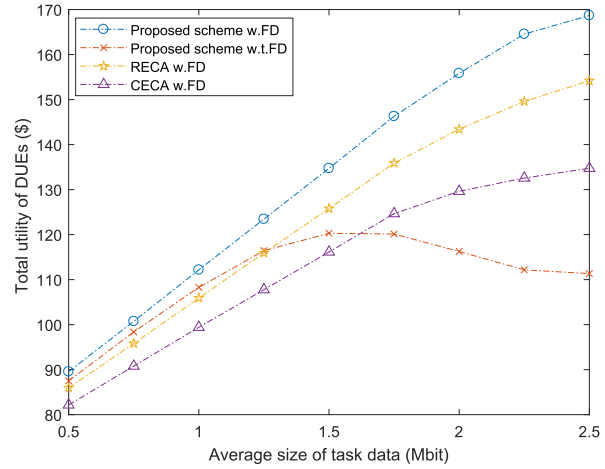


FIGURE 9. Total utility of DUEs with different average size of data.

explained as follows. When the data size is small, the cost caused by resource occupation is relatively small compared with the value of data. As the size of tasks increases, the cost

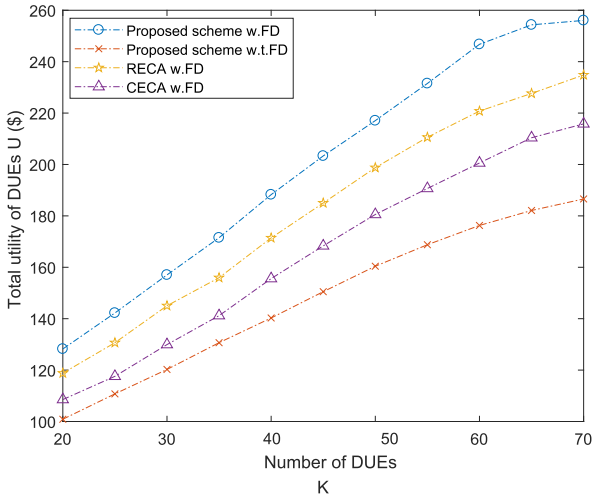


FIGURE 10. The influence of number of DUEs.

increases as additional power and computational resources are in need to ensure all the tasks are transmitted completely within the maximum tolerance time. Compared with other schemes, the proposed scheme with FD achieves the best performance.

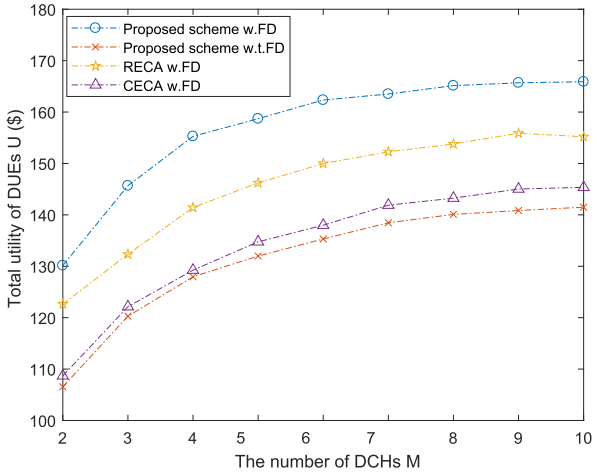


FIGURE 11. The influence of the number of FD-DCHs.

The influence of the number of DUEs and FD-DCHs are presented in Fig. 10 and Fig. 11, respectively. From Fig.10, it can be observed that as the number of DUEs increase, the utility increases but the growth rate decreases slowly. The reason is that with limited power of FD-DCHs and computing resources of MEC server, the system should balance the cost between DUEs to maximize the total utility of DUEs as there exists a competition in resources among DUEs. In Fig. 11, as the number of FD-DCHs increases, the utility increases to some extent at the beginning and the growth rate tends to be flat afterwards. It is because the increase in FD-DCHs enables DUEs to select the FD-DCHs with better channel conditions, which improves the system performance.

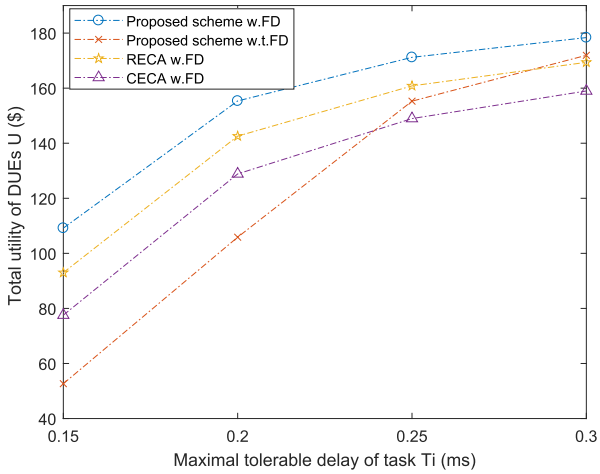


FIGURE 12. The influence of the average maximal tolerable delay.

The influence of the average maximal tolerable delay of tasks is depicted in Fig.12. It can be seen that as the average maximal tolerable delay increases, the total utility of DUEs increases due to the decrease in power and computational resources. The proposed scheme without FD performs better in some intervals compared with other schemes.

VII. CONCLUSION

In this paper, we investigate the computation and traffic offloading in cache aided D2D multicast networks for the content delivery and delay sensitive task offloading services. Firstly, based on the social attributes, available energy and transfer rate of DUEs, we propose a DCH selection strategy to provide stable multicast links and enhanced computing resources. Then, a novel multicast-aware coded and cooperative caching scheme is proposed to improve the efficiency of content distribution and optimize the energy consumption of content delivery. Finally, we formulate an optimization problem and propose computation offloading and resource allocation optimization schemes. Furthermore, the optimization problem is transformed it into two sub problems, UAOP and RAOP. The later one is proved as a convex problem, and its optimal resource allocation solution is found. The effectiveness of our proposed schemes are demonstrated by simulation results with different system parameters. For future work, we will consider the influence of DUEs mobility on caching and computation offloading.

REFERENCES

[1] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A survey on mobile edge computing: The communication perspective," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 4, pp. 2322–2358, 4th Quart., 2017.

[2] D. Wu, S. Si, S. Wu, and R. Wang, "Dynamic trust relationships aware data privacy protection in mobile crowd-sensing," *IEEE Internet Things J.*, vol. 5, no. 4, pp. 2958–2970, Aug. 2018.

[3] S. Guo, D. Wu, H. Zhang, and D. Yuan, "Resource modeling and scheduling for mobile edge computing: A service provider's perspective," *IEEE Access*, vol. 6, pp. 35611–35623, Jun. 2018.

[4] T. Zhang, "Data offloading in mobile edge computing: A coalition and pricing based approach," *IEEE Access*, vol. 6, pp. 2760–2767, Dec. 2017.

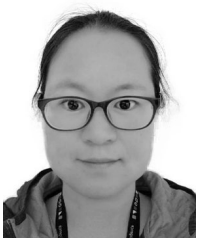
[5] M. Chen, Y. Qian, Y. Hao, Y. Li, and J. Song, "Data-driven computing and caching in 5G networks: Architecture and delay analysis," *IEEE Wireless Commun.*, vol. 25, no. 1, pp. 70–75, Feb. 2018.
[6] R. Wang, J. Yan, D. Wu, H. Wang, and Q. Yang, "Knowledge-centric edge computing based on virtualized D2D communication systems," *IEEE Commun. Mag.*, vol. 56, no. 5, pp. 32–38, May 2018.
[7] C.-M. Huang, M.-S. Chiang, D.-T. Dao, W.-L. Su, S. Xu, and H. Zhou, "V2V data offloading for cellular network based on the software defined network (SDN) inside mobile edge computing (MEC) architecture," *IEEE Access*, vol. 6, pp. 17741–17755, Mar. 2018.
[8] E. Bastug, M. Bennis, M. Médard, and M. Debbah, "Toward interconnected virtual reality: Opportunities, challenges, and enablers," *IEEE Commun. Mag.*, vol. 55, no. 6, pp. 110–117, Jun. 2017.
[9] L. Feng *et al.*, "Resource allocation for 5G D2D multicast content sharing in social-aware cellular networks," *IEEE Commun. Mag.*, vol. 56, no. 3, pp. 112–118, Mar. 2018.
[10] D. Wu, J. Yan, H. Wang, D. Wu, and R. Wang, "Social attribute aware incentive mechanism for device-to-device video distribution," *IEEE Trans. Multimedia*, vol. 19, no. 8, pp. 1908–1920, Aug. 2017.
[11] K. W. Choi, D. T. Wiriaatmadja, and E. Hossain, "Discovering mobile applications in cellular device-to-device communications: Hash function and Bloom filter-based approach," *IEEE Trans. Mobile Comput.*, vol. 15, no. 2, pp. 336–349, Feb. 2016.
[12] G. Zhang, K. Yang, and H. H. Chen, "Socially aware cluster formation and radio resource allocation in D2D networks," *IEEE Wireless Commun.*, vol. 23, no. 4, pp. 68–73, Aug. 2016.
[13] J. Liu, L. Fu, J. Zhang, X. Wang, and J. Xu, "Modeling multicast group in wireless social networks: A combination of geographic and non-geographic perspective," *IEEE Trans. Wireless Commun.*, vol. 16, no. 6, pp. 4023–4037, Jun. 2017.
[14] B. Chen, C. Yang, and A. F. Molisch, "Cache-enabled device-to-device communications: Offloading gain and energy cost," *IEEE Trans. Wireless Commun.*, vol. 16, no. 7, pp. 4519–4536, Jul. 2017.
[15] B. Chen, C. Yang, and Z. Xiong, "Optimal caching and scheduling for cache-enabled D2D communications," *IEEE Commun. Lett.*, vol. 21, no. 5, pp. 1155–1158, 2017.
[16] Y. Long, D. Wu, Y. Cai, and J. Qu, "Joint cache policy and transmit power for cache-enabled D2D networks," *IET Commun.*, vol. 11, no. 16, pp. 2498–2506, Nov. 2017.
[17] X. Song, Y. Geng, X. Meng, J. Liu, W. Lei, and Y. Wen, "Cache-enabled device to device networks with contention-based multimedia delivery," *IEEE Access*, vol. 5, pp. 3228–3239, Feb. 2017.
[18] Y. Zhang, Y. Xu, Q. Wu, X. Liu, K. Yao, and A. Anpalagan, "A game-theoretic approach for optimal distributed cooperative hybrid caching in D2D networks," *IEEE Wireless Commun. Lett.*, vol. 7, no. 3, pp. 324–327, Jun. 2018.
[19] J. Rao, H. Feng, C. Yang, Z. Chen, and B. Xia, "Optimal caching placement for D2D assisted wireless caching networks," in *Proc. IEEE ICC*, Kuala Lumpur, Malaysia, May 2016, pp. 1–6.
[20] M. Ji, G. Caire, and A. F. Molisch, "Wireless device-to-device caching networks: Basic principles and system performance," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 1, pp. 176–189, Jan. 2016.
[21] Z. Zhao, M. Xu, Y. Li, and M. Peng, "A non-orthogonal multiple access-based multicast scheme in wireless content caching networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2723–2735, Dec. 2017.
[22] R. Wang, J. Zhang, S. H. Song, and K. B. Letaief, "Exploiting mobility in cache-assisted D2D networks: Performance analysis and optimization," *IEEE Trans. Wireless Commun.*, vol. 17, no. 8, pp. 5592–5605, Aug. 2018.
[23] T. Deng, G. Ahani, P. Fan, and D. Yuan, "Cost-optimal caching for D2D networks with user mobility: Modeling, analysis, and computational approaches," *IEEE Trans. Wireless Commun.*, vol. 17, no. 5, pp. 3082–3094, May 2018.
[24] B. Chen, C. Yang, and G. Wang, "High-throughput opportunistic cooperative device-to-device communications with caching," *IEEE Trans. Veh. Technol.*, vol. 66, no. 8, pp. 7527–7539, Aug. 2017.
[25] P. Lin, Q. Song, Y. Yu, and A. Jamalipour, "Extensive cooperative caching in D2D integrated cellular networks," *IEEE Commun. Lett.*, vol. 21, no. 9, pp. 2101–2104, Sep. 2017.
[26] D. Wu, L. Zhou, Y. Cai, and Y. Qian, "Collaborative caching and matching for D2D content sharing," *IEEE Wireless Commun.*, vol. 25, no. 3, pp. 43–49, Jun. 2018.
[27] Z. Zhao, M. Peng, Z. Ding, W. Wang, and H. V. Poor, "Cluster content caching: An energy-efficient approach to improve quality of service in cloud radio access networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 5, pp. 1207–1221, May 2016.
[28] R. Wang, J. Zhang, S. H. Song, and K. B. Letaief, "Mobility-aware caching in D2D networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 8, pp. 5001–5015, Aug. 2017.
[29] D. Wu, F. Zhang, H. Wang, and R. Wang, "Security-oriented opportunistic data forwarding in mobile social networks," *Future Gener. Comput. Syst.*, vol. 87, pp. 803–815, Oct. 2018.
[30] T. X. Tran, A. Hajisami, P. Pandey, and D. Pompili, "Collaborative mobile edge computing in 5G networks: New paradigms, scenarios, and challenges," *IEEE Commun. Mag.*, vol. 55, no. 4, pp. 54–61, Apr. 2017.
[31] K. Zhang *et al.*, "Energy-efficient offloading for mobile edge computing in 5G heterogeneous networks," *IEEE Access*, vol. 4, pp. 5896–5907, Aug. 2016.
[32] L. Yang, H. Zhang, M. Li, J. Guo, and H. Ji, "Mobile edge computing empowered energy efficient task offloading in 5G," *IEEE Trans. Veh. Technol.*, vol. 67, no. 7, pp. 6398–6409, Jul. 2018.
[33] H. Zhang, J. Guo, L. Yang, X. Li, and H. Ji, "Computation offloading considering fronthaul and backhaul in small-cell networks integrated with MEC," in *Proc. IEEE INFOCOM WKSHPS*, Atlanta, GA, USA, May 2017, pp. 115–120.
[34] J. Zhang, W. Xia, F. Yan, and L. Shen, "Joint computation offloading and resource allocation optimization in heterogeneous networks with mobile edge computing," *IEEE Access*, vol. 6, pp. 19324–19337, Mar. 2018.
[35] C. Wang, C. Liang, F. R. Yu, Q. Chen, and L. Tang, "Joint computation offloading, resource allocation and content caching in cellular networks with mobile edge computing," in *Proc. IEEE ICC*, Paris, France, May 2017, pp. 1–6.
[36] Z. Tan, F. R. Yu, X. Li, H. Ji, and V. C. M. Leung, "Virtual resource allocation for heterogeneous services in full duplex-enabled SCNs with mobile edge computing and caching," *IEEE Trans. Veh. Technol.*, vol. 67, no. 2, pp. 1794–1808, Feb. 2018.
[37] Y. Hao, M. Chen, L. Hu, M. S. Hossain, and A. Ghoneim, "Energy efficient task caching and offloading for mobile edge computing," *IEEE Access*, vol. 6, pp. 11365–11373, Mar. 2018.
[38] X. Liu, J. Zhang, X. Zhang, and W. Wang, "Mobility-aware coded probabilistic caching scheme for MEC-enabled small cell networks," *IEEE Access*, vol. 5, pp. 17824–17833, Aug. 2017.
[39] U. Niesen and M. A. Maddah-Ali, "Coded caching with nonuniform demands," *IEEE Trans. Inf. Theory*, vol. 63, no. 2, pp. 1146–1158, Feb. 2017.
[40] M. Heino *et al.*, "Recent advances in antenna design and interference cancellation algorithms for in-band full duplex relays," *IEEE Commun. Mag.*, vol. 53, no. 5, pp. 91–101, May 2015.
[41] S. Chen, F. Qin, B. Hu, X. Li, and Z. Chen, "User-centric ultra-dense networks for 5G: Challenges, methodologies, and directions," *IEEE Wireless Commun.*, vol. 23, no. 2, pp. 78–85, Apr. 2016.
[42] L. Chen, F. R. Yu, H. Ji, G. Liu, and V. C. M. Leung, "Distributed virtual resource allocation in small-cell networks with full-duplex self-backhauls and virtualization," *IEEE Trans. Veh. Technol.*, vol. 65, no. 7, pp. 5410–5423, Jul. 2016.
[43] J. N. Laneman, D. N. C. Tse, and G. W. Wornell, "Cooperative diversity in wireless networks: Efficient protocols and outage behavior," *IEEE Trans. Inf. Theory*, vol. 50, no. 12, pp. 3062–3080, Dec. 2004.
[44] M. Taghizadeh, K. Micinski, S. Biswas, C. Ofria, and E. Torng, "Distributed cooperative caching in social wireless networks," *IEEE Trans. Mobile Comput.*, vol. 12, no. 6, pp. 1037–1053, Jun. 2012.
[45] X. Wang, M. Chen, T. Taleb, A. Ksentini, and V. C. M. Leung, "Cache in the air: Exploiting content caching and delivery techniques for 5G systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 131–139, Feb. 2014.

DONGYU WANG received the B.S. and M.S. degrees from Tianjin Polytechnic University, China, in 2008 and in 2011, respectively, and the Ph.D. degree from the Beijing University of Posts and Telecommunications, China, in 2014. From 2014 to 2016, he was a Post Ph.D. with the Department of Biomedical Engineering, Chinese PLA General Hospital, Beijing.

In 2016, he joined the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications. His research interests include device-to-device communication, multimedia broadcast/multicast service systems, resource allocation, theory and signal processing, with specific interests in cooperative communications, and mobile edge computing.

YANWEN LAN received the B.S. degree in communications engineering from Henan University in 2013, and the M.E. degree in electrical and communications engineering from Hainan University in 2016. He is currently pursuing the M.S. degree with the Beijing University of Posts and Telecommunications. His research interests include mobile edge computing and cache-enabled heterogeneous networks.

TIEZHU ZHAO received the M.S. degree in telecom engineering from Xi'an Jiaotong University in 2001. He is currently with the Samsung China R&D Center. His research interests include radio network design, optimization and verification, and system evaluation for 4G and 5G products.



ZHENPING YIN received the M.S. degree in communication and information system from the Beijing University of Post and Telecommunications. She is currently with the Samsung China R&D Center. Her research interests include software development of 4G and 5G products.

XIAOXIANG WANG received the B.S. degree in physics from Qufu Normal University, Qufu, China, in 1991, the M.S. degree in information engineering from East China Normal University, Shanghai, China, in 1994, and the Ph.D. degree in electronic engineering from the Beijing Institute of Technology, Beijing, China, in 1998. In 1998, she joined the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications. From 2010 to 2011, she was a Visiting Fellow with the Department of Electrical and Computer Engineering, North Carolina State University, Raleigh. Her research interests include communications theory and signal processing, with specific interests in cooperative communications, multiple-input-multiple-output systems, multimedia broadcast/multicast service systems, and resource allocation.