CACHING IN HETEROGENEOUS ULTRADENSE 5G NETWORKS

A Comprehensive Cooperation Approach

Peng Lin, Komal S. Khan, Qingyang Song,
and Abbas Jamalipour

Rich multimedia services have significantly increased the traffic load over mobile networks. Emerging mobile content caching techniques can efficiently relieve overloaded network traffic by caching popular contents at intermediate nodes and routers. Furthermore, stimulated by network densification, coordinated multipoint (CoMP) joint transmission (JT) techniques are expected to have a significant role in 5G networks. In this article, we present a system architecture for a cache-enabled heterogeneous ultradense network (HUDN) that consists of CoMP-integrated ultradense cells and cluster-based device-to-device (D2D) networks. We propose three cooperative caching schemes in cellular networks, D2D networks, and cross-tier networks. In the proposed schemes, caching decisions will be made by considering the content's popularity, the device distribution, the transmission method, and the caching capability. Numerical evaluations reveal that the proposed schemes outperform existing ones in terms of delay and cache hit probability.

Evolution Toward Cache-Enabled Heterogeneous Ultradense Networks

Cache-enabled network densification is expected to be an effective approach to satisfy the explosive growth of mobile data traffic. The improvement is based on the fact that some very popular contents

Digital Object Identifier 10.1109/MVT.2019.2904748
Date of publication: 18 April 2019

Authorized licensed use limited to: CHONGQING UNIV OF POST AND TELECOM. Downloaded on April 25,2022 at 06:40:32 UTC from IEEE Xplore. Restrictions apply.

IMAGE LICENSED BY INGRAM PUBLISHING

are transmitted from a remote server repeatedly. Equipping cache-enabled access points with predictive capabilities, context awareness, and social networks can substantially reduce redundant traffic by proactively serving predictable user demands [1], [2]. Meanwhile, massive radio access points [i.e., small base stations (BSs), wireless routers, and D2D-enabled devices] are densely deployed to exploit the spatial reuse of spectrum. This pushes the 5G network in a heterogeneous and ultradense manner and prompts the network to include a cellular tier and a D2D tier [3]. Applying caching to mobile networks has attracted much attention from researchers. However, some problems for designing caching strategies in both cellular and D2D networks remain and need to be addressed.

Caching in Wireless Cellular Networks

To deliver contents to end users more quickly, while using bandwidth and storage resources more efficiently, caching in network edge nodes (e.g., small BSs, femtohelpers, and wireless routers) is considered to be a promising solution. Some recent works analyze possible challenges and propose some caching approaches [1]–[4]. For high cache hit probability, as many contents as possible are cached in users' nearby BSs. However, the storage space is limited and, as a result, content selection for caching must be performed. Traditional isolated caching policies, such as least recently used (LRU), least frequently used (LFU), and most popular content (MPC) strategies, have good performance in sparse networks. In the scenario in which caching nodes are densely deployed and hold close topological associations, these strategies will lead to content redundancy if each node does not share its caching information [2]. Therefore, exploiting cooperation between caching nodes is indispensable in mobile caching systems.

Several studies investigated a cooperation-based architecture for mobile caching systems. By utilizing the inherent topological associations between different caching nodes, the caching performance in a three-tier cooperative architecture is studied in [2]. This research analyzes how traffic is duplicated over the core network and can be reduced, resorting to cooperative caching. The results show the impacts of caching capacity, content popularity, and delivery techniques on the delay performance. However, it may be impaired due to a lack of high-capacity connections between caching nodes. Femtocaching [5] extends the caching deployment from a centrally controlled topology to a wireless distributed infrastructure. By deploying massive distributed caching helpers (e.g., femto access points and Wi-Fi points) that have high storage

Authorized licensed use limited to: CHONGQING UNIV OF POST AND TELECOM. Downloaded on April 25,2022 at 06:40:32 UTC from IEEE Xplore. Restrictions apply.

STIMULATED BY NETWORK DENSIFICATION, COORDINATED MULTIPOINT JOINT TRANSMISSION TECHNIQUES ARE EXPECTED TO HAVE A SIGNIFICANT ROLE IN 5G NETWORKS.

capacities, the devices in overlapping cellular networks can communicate with several helpers, and the corresponding caches can be regarded as ensembles to serve the devices. However, these studies only exploit cooperative caching in terms of storage space and content popularity. Therefore, to take full advantage of limited storage space, as many popular contents are cached as possible. However, as 5G networks support more diverse transmission techniques, the benefits of caching will depend not only on what content should be cached but also on how the cached content is transmitted.

CoMP-JT has become an important component of 5G networks, providing high-throughput transmissions by reducing intercell interference [6]. In a cache-enabled CoMP-integrated network, if content requested by a device is cached in several nearby BSs, a multi-input, single-output channel forms between the BSs and the device. The BSs can collaboratively transmit content to the device by JT instead of formal single transmission (ST). In this situation, from both the storage and transmission levels, determining how to exploit caching to provide high-quality services becomes a much more complicated issue.

D2D Caching for Future Wireless Networks

The idea of D2D communication is gaining attention for its ability to provide services for users without central control of the BSs. When considering content caching in D2D networks, the caching strategy design becomes more rigorous, not only because the storage capacity in the user devices is very small but also because the cached contents should be very specific and related to nearby user request patterns. Recent studies on D2D networks focus on designing content placement and delivery strategies, with most of the strategies designed to minimize network delay, transmission costs, or backhaul load. The main challenge is the interference issue within the D2D network. Aside from interference between D2D and cellular links, interference between D2D links themselves needs to be considered when designing content delivery strategies. To achieve better throughput, some research explores D2D caching from both coded and uncoded perspectives. In uncoded placement, the complete file or part of the file is stored in the available cache [7]. In a coded placement, each file usually is divided into small segments that are processed by certain coding methods [8], such as Fountain or Raptor, and then stored in the cache. Such placement schemes are desirable because they require less space to store these small file segments, especially when user devices have limited storage space.

To better schedule caching and transmission of devices in D2D networks, the devices can be registered with the cellular network to roughly predict the amount of data traffic that could be managed or served by the devices themselves. However, this is not easy because a registered D2D device might not be willing to participate in serving other devices in the cellular network. Although incentive-based mechanisms to increase a user's interest in participation have been introduced [9], factors such as the distance between the serving and requesting devices and the availability of content in the storage space of a device will play an important role in providing good performance.

To improve the efficiency of D2D caching and communications, many works on caching have introduced clustering. It has proven to be an effective strategy for improving traffic efficiency. In conventional clustering methods, user devices are placed into clusters based on the similarity of their interests [10]. In recent literature, spatial distribution of devices is considered to effectively cache popular contents [11]. Over a given region, the device distribution can be either sparse or dense. If the users are sparsely distributed, most of the D2D communication links are expected to be longer, which makes it difficult to carry a transmission out successfully and will lead to a wider blockage region. The longer the communication distance between any two devices, the bigger the blockage region around them, which will affect nearby devices that are willing to participate in the D2D communication process. For dense user distribution, the collaboration distance for most of the links is shorter due to users' proximity to each other, which is constructive for more successful D2D transmissions.

However, not all user devices are able to carry out successful D2D communications due to their caching states. The closer the content is to a requesting user, the shorter the D2D communication link. Therefore, to ensure that content delivery is available nearby, content caching for a D2D device should be very specific to its surrounding request patterns. As discussed previously, some caching schemes may consider clustering and spatial distribution for user devices. However, these methods do not evaluate the probabilities of successful D2D transmissions. We believe that D2D caching schemes should consider the communication process in D2D networks. To achieve an efficient caching model that supports highly successful D2D content transmissions, we suggest the concept of virtual clusters to distribute content in D2D networks more efficiently.

Motivation and Contribution

Based on the studies of caching in both cellular and D2D networks, this article jointly addresses the existing problems in a cache-enabled HUDN, which includes a

Authorized licensed use limited to: CHONGQING UNIV OF POST AND TELECOM. Downloaded on April 25,2022 at 06:40:32 UTC from IEEE Xplore. Restrictions apply.

CoMP-integrated cellular tier and a D2D tier. To better deploy caching in the HUDN, we present a heterogeneous caching architecture that supports comprehensive cooperation. Under this architecture, we first explore caching in the CoMP-integrated system with hybrid transmissions (i.e., JT and ST). More specifically, we propose a cooperative caching scheme in which the storage- and transmission-level cooperation between BSs is exploited to jointly optimize content placement for JT and ST. This scheme reduces the content redundancy among BSs while ensuring a good opportunity for CoMP-JT. We then study content caching in the D2D network. To promote smooth D2D communications and enable a high cache hit probability, a cooperative caching scheme based on virtual clustering is proposed. In addition, considering the lack of caching information exchanged between the cellular and D2D tiers, we propose a cross-tier feedback scheme to reduce cache redundancy between cellular and D2D networks.

System Architecture for Cache-Enabled HUDN

To enable efficient content caching and delivery in a heterogeneous network (HetNet), a system architecture that supports comprehensive cooperation between caching nodes is needed. Existing works have provided some innovative principles and concepts for caching management in HetNets, involving some technical issues such as transmission mode selection, content scheduling, and interference management [4]. Figure 1 shows the system architecture of the cache-enabled HUDN, which consists of CoMP-integrated ultradense cells and cluster-based D2D networks. The HUDN provides high-quality services with the following function units.

- *Baseband unit (BBU) pool*: In the data center (DC), the BBU holds powerful computing capacity with low power consumption. 5G C-RAN also moves the baseband signal processing tasks of small-cell BSs, such as CoMP, centralized resource allocation, and request and data flow scheduling, into the BBU pool.
- *Remote radio heads (RRHs)*: The RRHs are connected with the BBU via optical fibers and are randomly located over coverage areas to provide content services and enhance connectivity.
- *Intelligent storage unit (ISU)*: The ISUs are deployed at small BSs, helper nodes, and wireless routers. For the sake of description, we uniformly call these cached-enabled nodes *BSs* in this article. Each ISU can select popular contents for caching and is responsible for

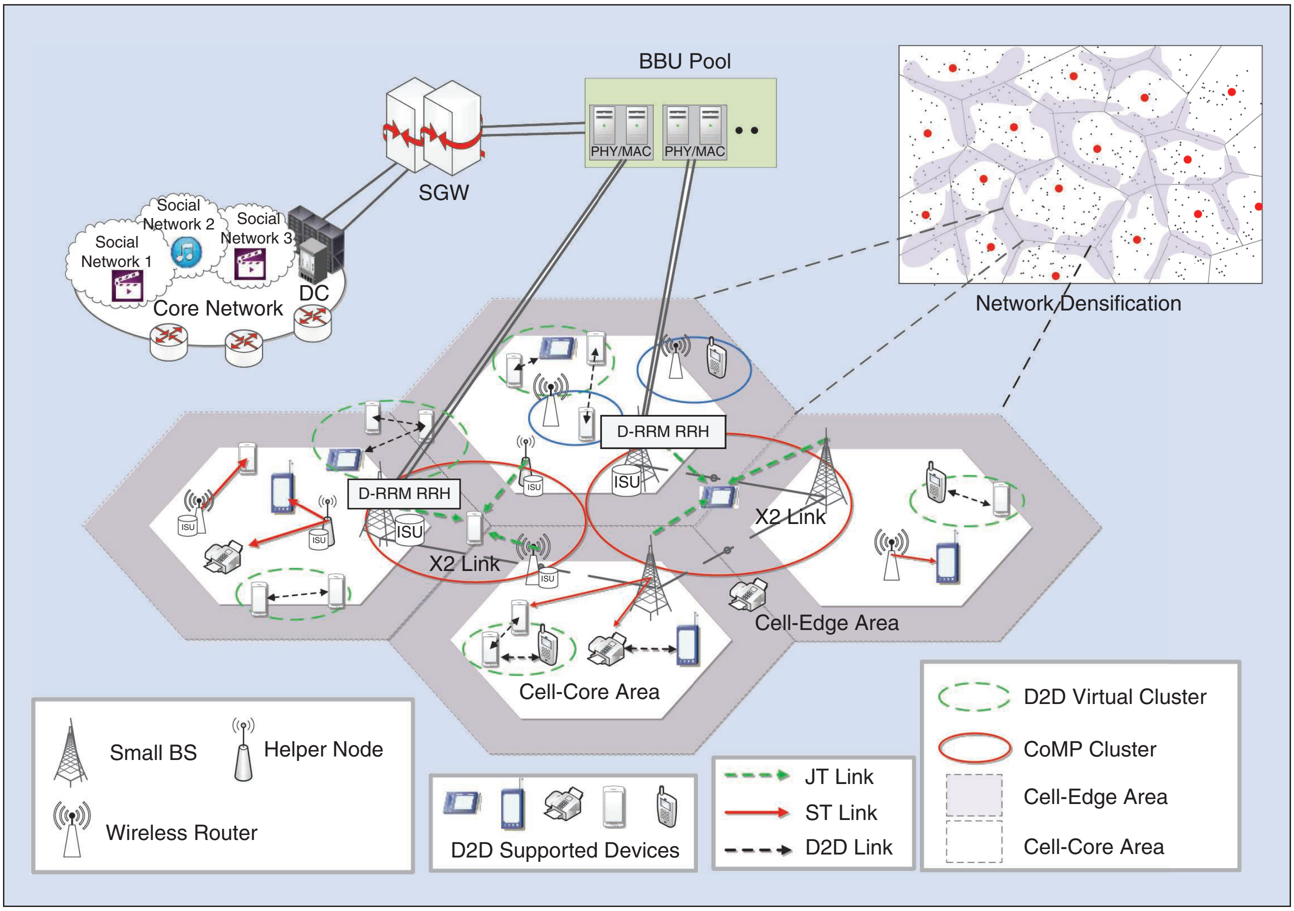


FIGURE 1 The cache-enabled HUDNs. PHY: physical layer; MAC: media access control layer.

Authorized licensed use limited to: CHONGQING UNIV OF POST AND TELECOM. Downloaded on April 25,2022 at 06:40:32 UTC from IEEE Xplore. Restrictions apply.

EQUIPPING CACHE-ENABLED ACCESS POINTS WITH PREDICTIVE CAPABILITIES, CONTEXT AWARENESS, AND SOCIAL NETWORKS CAN SUBSTANTIALLY REDUCE REDUNDANT TRAFFIC BY PROACTIVELY SERVING PREDICTABLE USER DEMANDS.

request-information collection, cache scheduling and sharing, and cache updating.

- *D2D radio resource management unit (D-RRM)*: The D-RRM is responsible for managing D2D networks that involve D2D link scheduling, potential device discovery, access mode selection, and interference avoiding.
- *High-capacity X2 interface*: Adjacent BSs are connected by a high-capacity link through the X2 interface and the X2 link. Then the cached contents and corresponding control instructions can be transmitted between BSs in a timely way.

In the cache-enabled HUDN, the densely deployed small BSs provide global wireless coverage. They are under the centralized control of a service gateway (SGW) that connects to a DC. The content providers (CPs) in the core network provide a library of multimedia contents throughout the whole network. The ISU is built at each BS and has a certain amount of storage space to cache contents. Network densification brings difficulties to interference management. To simplify network management, the network is divided into the cell-core area and the cell-edge area, according to the long-term averaged received power (LAP). We assume an association strategy in which each mobile device is associated only with the BS that has the highest LAP, the anchored BS. The other BSs that cover the device are called *assisted BSs*, and they form a CoMP cluster to provide JT services. The BSs in the cluster are assumed to support both JT and ST. The devices located in the cell-core area (the cell-core devices) are served with ST. The devices located in the cell-edge area (the cell-edge devices) can obtain not only ST service but also JT service to improve cell-edge throughput. In addition, each device is able to cache the content it received earlier, to assist the nearby devices with direct D2D communications.

In the cache-enabled HUDN, multimedia contents are pushed to the network's edge nodes (i.e., BSs and devices). The BSs and D2D devices with cached contents can serve as access points to provide multimedia contents. When a device requests content, it can be obtained from an access point directly to significantly reduce the content's delivery delay and alleviate the backhaul traffic. The key challenge is to achieve optimal content placement in the HUDN. This requires an integrated caching scheme that considers request patterns, topology characteristics, and transmission methods. Achieving this goal requires addressing the problems, including devices' transmission mode selection, content request scheduling, and potential cache discovery. Hence, to exploit caching efficiently and provide devices with low-latency services in the cache-enabled HUDN, we will explore the impact of nodes' cooperation on the HUDN's performance and propose a comprehensive cooperative caching scheme.

Cooperative Caching for CoMP-Integrated Cellular Networks

Content Popularity and Device Distribution

The key to designing a caching strategy is to accurately evaluate content popularity and analyze the device distribution. Content popularity can be interpreted as the ratio of the number of requests for particular content to the total number of requests. It is both time varying and space varying, i.e., the popularity of the content may be different at different times and in different locations. Content popularity can be divided into global and local popularities, based on the scope of the investigation. Most studies have proven that global content popularity follows the Zipf distribution [12], which has a skewness parameter $\alpha \in [0,1]$ that characterizes the aggregation degree of the content. Local popularity indicates the probability that content will being requested in a local area. Generally, the popularity of local content in a certain region can be obtained during a given period with the help of big data analytics.

Device distribution indicates the profile of devices' locations in the network. It can be used to predict the spatial composition of content requests. We denote a parameter $\beta \in [0,1]$ as the ratio of the number of cell-edge requests to the number of cell-core requests, i.e., the request distribution coefficient. In the CoMP-integrated network, a larger value of β indicates that the additional requests received by a BS are from the cell-edge area. To improve the cell-edge throughput, focus the content placement in a CoMP cluster on creating more JT opportunities so that more cell-edge devices can be jointly served. Conversely, a smaller β indicates that more requests are from the cell-core area and will be served with ST. In this situation, the content placement in adjacent BSs should focus on ensuring high cache hit probability with ST. Therefore, device distribution is an important factor to consider for caching deployment in CoMP-integrated systems.

Cooperation Strategy Among BSs

In conventional cache-enabled cellular networks, when a BS receives a request, it first checks its local storage and serves the device with ST if the content is cached. Otherwise, the BS will retrieve the content from the CP. In this context, an isolated caching strategy (e.g., LRU, LFU, or MPC) can usually provide good performance. However, in the CoMP-integrated system, a cell-edge device can

Authorized licensed use limited to: CHONGQING UNIV OF POST AND TELECOM. Downloaded on April 25,2022 at 06:40:32 UTC from IEEE Xplore. Restrictions apply.

obtain the desired content either from its anchored BS by ST or from both anchored and assisted BSs by JT, based on the caching state of the desired content. Therefore, to make the caching deployment efficient and adaptive to the hybrid transmission, an integrated caching scheme with two types of cooperation is considered.

- *Storage-level cooperation*: One way to exploit cooperative caching is to enable the BSs to share the cached contents through X2 links. Accordingly, for a certain device (in the cell-core or cell-edge area) served with ST, its request can be satisfied by its anchored BS directly (if the BS has cached the desired content) or through a neighboring BS that is inaccessible for the device but has the content cached. Otherwise, the desired file will be retrieved from the CP. To ensure a high cache hit probability, the BSs should cache more and different contents to reduce content redundancy among BSs. Specifically, each BS should be able to determine what contents need to be cached by the anchored BS and what contents need to be cached as supplements in its neighboring BSs.
- *Transmission-level cooperation*: In the CoMP-integrated system, in addition to the performance improvement resulting from reducing content redundancy, transmission-oriented caching provides JT gain. When the device is in the cell-edge area, its accessible BSs jointly decide which transmission method (i.e., ST or JT) to select to serve the device based on the caching state of the desired content. If the anchored BS and multiple assisted BSs cache the content, the device will be served with JT. Otherwise, the request will be satisfied by ST provided by the anchored BS. Because JT can provide higher throughput for cell-edge devices compared with ST, BSs are encouraged to cache more of the same contents to facilitate more JT opportunities.

Delay Analysis

We incorporate the two types of cooperation discussed into a comprehensive cooperative caching scheme that focuses on reducing content delivery delay. In the cache-enabled HUDN, when devices obtain contents through the cellular network, content delivery delay is composed of the following three parts.

- *Intracell delivery delay*: The transmission delay from a BS to a device, or *intracell delay*, depends on the transmission method, i.e., ST or JT. When a cell-edge device u requests content, the transmit rate using JT is denoted by r_{JT}, which can be calculated based on superposing multiple signal-to-interference-plus-noise ratios (SINRs) from BSs in the CoMP cluster to device u. For a certain content f_k with size S_k, the transmission delay for using JT is S_k/r_{JT}. Similarly, the transmission delay using ST is obtained as S_k/r_{ST}, where r_{ST} is the transmit rate using ST and can be calculated based on the SINR from the anchored BS to device u. To reduce the impact of delay fluctuation caused by the instability of wireless transmissions, each BS uses the average transmit rate to compute the expected delays when JT and ST are used.
- *Intercell sharing delay*: This delay appears when content is not cached in the device's anchored BS but has been cached in one of the anchored BS's neighboring BSs. Denote the transmit rate between adjacent BSs by r_{NE}. Then, the content delivery delay of a BS fetching the desired content from a neighboring BS can be calculated according to r_{NE}. All the BSs calculate the delay according to the average transmission rate between BSs.
- *Outer-cell delay*: The outer-cell delay occurs when the requested content is not cached in any BS. In this situation, the content has to be retrieved from the CP. The length of the delay depends on the state of the core network. For instance, the average transmit rate is generally high at night and low during the day. Denote the transmit rate of the backhaul link by r_{CP}; then, the delay of the BS fetching content f_k from the CP through the core network can be calculated using S_K/r_{CP}.

Problem Formulation

To support delay-sensitive content delivery in the CoMP-integrated cellular network, the purpose now is to minimize the content delivery delay by considering joint content placement for JT and ST. The caching decision is defined by $x_{i,k}$ to denote whether content f_k is cached in BS i. According to the delay analysis in the previous section, the content delivery delay of a device that accesses contents via the cellular link includes two parts.

One is the wireless transmission delay D_{cell} in the cell, which includes the transmission delay that results from using JT and the transmission delay that results from using using ST. The other is the backhaul delivery delay D_{bh}. This comprises the delivery delay from neighboring BSs and the CP. All of these parameters are evaluated based on the caching states of BSs. Next, the optimization problem becomes how best to deploy the content placement $x_{i,k}$ to minimize the content delivery delay, i.e., $\min D_{\mathrm{cell}} + D_{\mathrm{bh}}$, which should satisfy the following constraints.

- For BS i with storage capacity C_i, it can cache, at most, C_i/S_k contents in the local storage space.
- Content f_k can be transmitted to a cell-edge device by JT on the condition that all the BSs in the CoMP cluster have cached the desired content f_k.
- The traffic load produced by content sharing between adjacent BSs should not exceed the X2 link capacity.
- The content placement decision for a BS is a binary variable.

The second constraint is that all of the BSs in the same CoMP cluster must have cached the content that is intended to be transmitted using JT, so that they can provide devices with JT service in a timely manner. Note that there is another opportunity for BSs to perform JT. We wait for the

Authorized licensed use limited to: CHONGQING UNIV OF POST AND TELECOM. Downloaded on April 25,2022 at 06:40:32 UTC from IEEE Xplore. Restrictions apply.

TO SUPPORT EFFICIENT D2D COMMUNICATIONS, WE CONSTRUCT A VIRTUAL CLUSTERING STRATEGY IN WHICH THE REGIONS WITH A DENSE DEVICE DISTRIBUTION ARE CONSIDERED A VIRTUAL CLUSTER.

BS that does not have the desired content to download it from a cooperative BS or the CP. Then, the BSs transmit the content to a cell-edge device using JT. However, compared with the case in which the content has been cached and ready for CoMP-JT, this method requires an extra delay and backhaul load. Therefore, we do not consider it as a possible option. The third constraint guarantees that content sharing cannot produce an extra burden on the X2 link, to ensure smooth cooperation among BSs. The investigated problem is nonlinear integer programming, which is NP-hard. In the next section, we use a method based on a genetic algorithm (GA) to solve it.

Caching Decision

To make the caching strategy work in an effective way, each BS periodically evaluates the request distribution coefficient β and the average transmit rate r_{NE} and r_{CP}. These parameters, combined with the content-related information, are aggregated and sent to the SGW. The optimization problem is solved by the GA-based method in the SGW. The flowchart of our proposed scheme is shown in Figure 2. In the GA-based algorithm, the BS caching decisions of are first encoded together to be a matrix, specifically an individual matrix. Then, multiple individuals are constructed to be the first-generation population. The effect of caching decisions depends on the fitness function, which is constructed by the content-delivery delay. By iteratively addressing the population with crossover, mutation, and selection, we can obtain an excellent individual matrix that is close to the optimal result. The output matrix is then decoded into caching decisions for each BS and each BS cache's contents according to the optimized caching decisions.

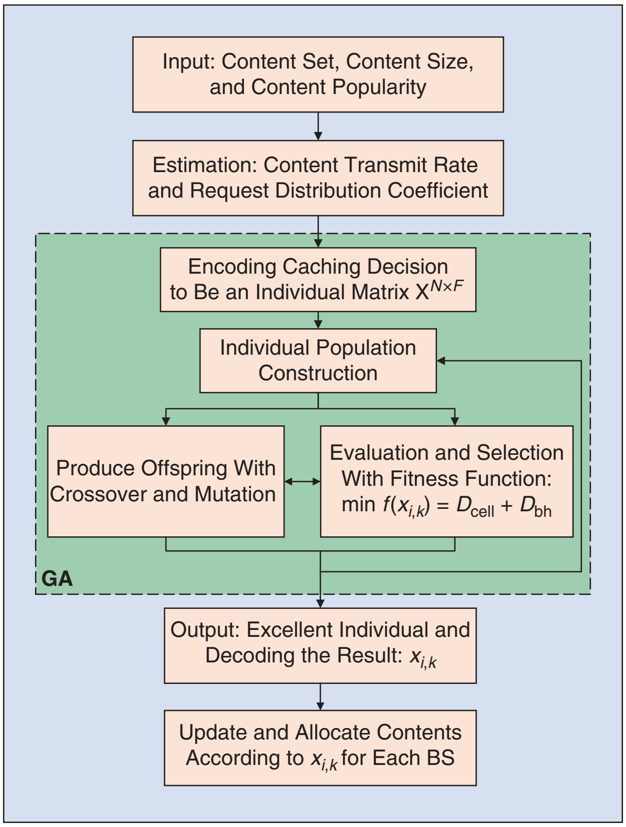

Input: Content Set, Content Size, and Content Popularity
Estimation: Content Transmit Rate and Request Distribution Coefficient
Encoding Caching Decision to Be an Individual Matrix $X^{N\times F}$
Individual Population Construction
Produce Offspring With Crossover and Mutation
Evaluation and Selection With Fitness Function: min $f(x_{i,k}) = D_{\mathrm{cell}} + D_{\mathrm{bh}}$
GA
Output: Excellent Individual and Decoding the Result: $x_{i,k}$
Update and Allocate Contents According to $x_{i,k}$ for Each BS

FIGURE 2 The GA-based algorithm for caching.

D2D Caching Based on Virtual Clustering

Virtual Cluster Formation

For the D2D communication that underlays the cellular network, we know that there can be sparse distribution of devices in one subregion and dense distribution of devices in another subregion within a BS's coverage, depending on users' locations. This variable distribution affects the D2D communication process because the sparsely distributed devices might find that it is difficult to carry out successful D2D communications due to the longer distance between devices. In this situation, even the requested content is cached in a device but still is not able to participate in the request-serving process due to the distance requirement. To support efficient D2D communications, we construct a virtual clustering strategy in which the regions with a dense device distribution are considered a virtual cluster. This can be seen in Figure 1.

To construct the virtual cluster, we refer to D2D devices as points in a given region under the management of a BS. The D2D devices are modeled by a homogeneous Poisson point process (HPPP) with density λ. Considering frequency reuse, we assume that multiple D2D links can be activated in a given cluster simultaneously. Considering that the location of each point is known to the BS, the goal is to find which point should be included in one virtual cluster. We use an average-minimum-distance method to construct the virtual cluster. The BS first determines a point (user device), the original cluster center, in its coverage area. For a nearby point n_a that wants to join in the cluster, the BS evaluates the average minimum distance $\bar{d}_V$ of the virtual point set V,

$$\bar{d}_V = \frac{1}{M}\sum_{n\in V}\operatorname*{dist}_{n_a\in V/n}\{n, n_a\}, \qquad (1)$$

where $\mathrm{dist}\{n, n_a\}$ is the distance between points n and n_a and M is the total number of points in the point set. Parameter $\bar{d}_V$ measures the overall spacing of the point set. We assume that point n_a can be included in set V only when $\bar{d}_V < d_{\mathrm{th}}$, where d_{th} is the threshold factor for

Authorized licensed use limited to: CHONGQING UNIV OF POST AND TELECOM. Downloaded on April 25,2022 at 06:40:32 UTC from IEEE Xplore. Restrictions apply.

efficient transmissions in D2D networks. This dynamic parameter can be adjusted according to the network state. The other qualified points are added into the cluster according to the evaluation strategy until the point set converges to be stable. In this way, the devices are divided into multiple virtual clusters under the management of BSs.

Cluster-Based Cooperative Caching

We explore content placement in the virtual D2D cluster in this section. Because the devices in the same virtual cluster can maintain smooth communications, reducing content redundancy plays an important role in improving cache hit probability. We propose a scheme in which devices within a virtual cluster will have different contents cached, because user devices can easily request different contents from nearby devices. This scheme can be achieved by making the devices broadcast their local caching information in the virtual cluster on a periodic basis. Specifically, for each device that obtains desired content from the cellular network, it first checks the caching table, which records the caching information of the surrounding devices in the cluster. If the content is not cached, the device will replace old material with newly acquired material in descending order of content popularity. All the devices update contents in this way, ensuring high utilization of the limited storage space.

The BSs are responsible for managing the caching and communication of devices, which involves caching and updating in the virtual cluster, as shown in Figure 3. Using the location information of the D2D devices, every BS periodically updates and keeps track of the devices within a virtual cluster. A BS adds or removes D2D devices from its virtual cluster by evaluating the average minimum distance $\bar{d}_V$. Once a D2D device is added to the virtual cluster, it can participate in D2D communication. On the other hand, if a D2D device moves outside a virtual cluster's region, it will rely on self-service from its own cache or request its desired content from the BS.

Cross-Tier Cooperative Caching

When a device intends to request content, it can be obtained from neighboring devices in the virtual D2D cluster or, if the requested content is not locally cached, from the anchored BS. In such a two-tier caching network, cooperative caching among BSs and devices becomes necessary to reduce content redundancy. Specifically, some popular contents that a large proportion of devices have cached should less likely be cached by the local BSs. Therefore, a feedback scheme between the cellular and D2D tiers is needed. The cross-tier feedback scheme includes the following two steps.

- *Information collection from D2D networks*: Devices in the same D2D cluster periodically broadcast their caching information to each other. Based on the shared messages, the caching table that records the surrounding caching information updates immediately once the contents in the devices are replaced. Then, each device sends its caching table to the local BS.
- *Content popularity updating*: Based on the caching table, the popularities of contents in the local BS are updated to form a new content-popularity listing, which indicates the probability that the devices in a cell will request the corresponding contents through cellular links.

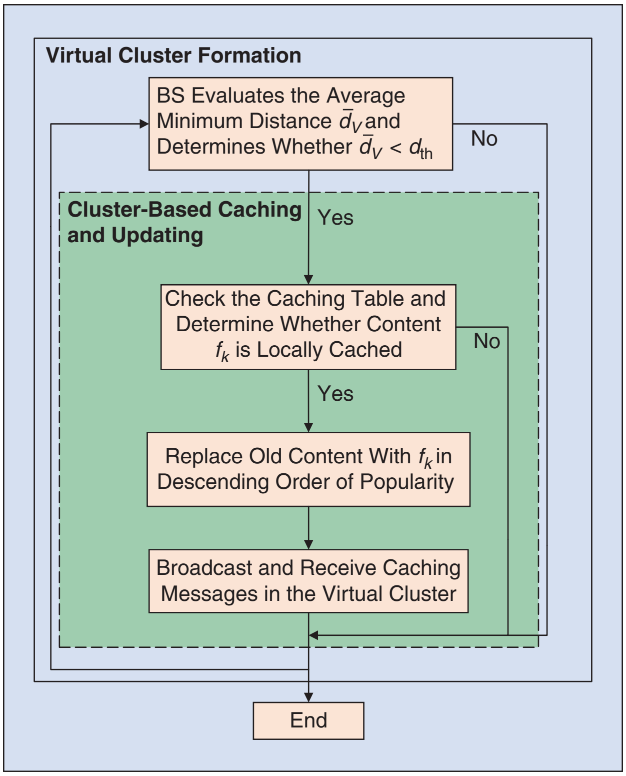
The newly acquired content popularity is imported into the optimization process for content placement in BSs, which is then updated and the content redundancy between cellular and D2D networks reduced.

Mode Selection and Content Scheduling

In the comprehensive cooperative caching framework we have discussed, when a device generates a content request, there are three transmission modes for content delivery: D2D mode, cellular-ST (C-ST) mode, and cellular-JT (C-JT) mode. Figure 4 shows how the device obtains its desired content by using different transmission methods based on the content placement in the cache-enabled HUDN.

Performance Evaluation

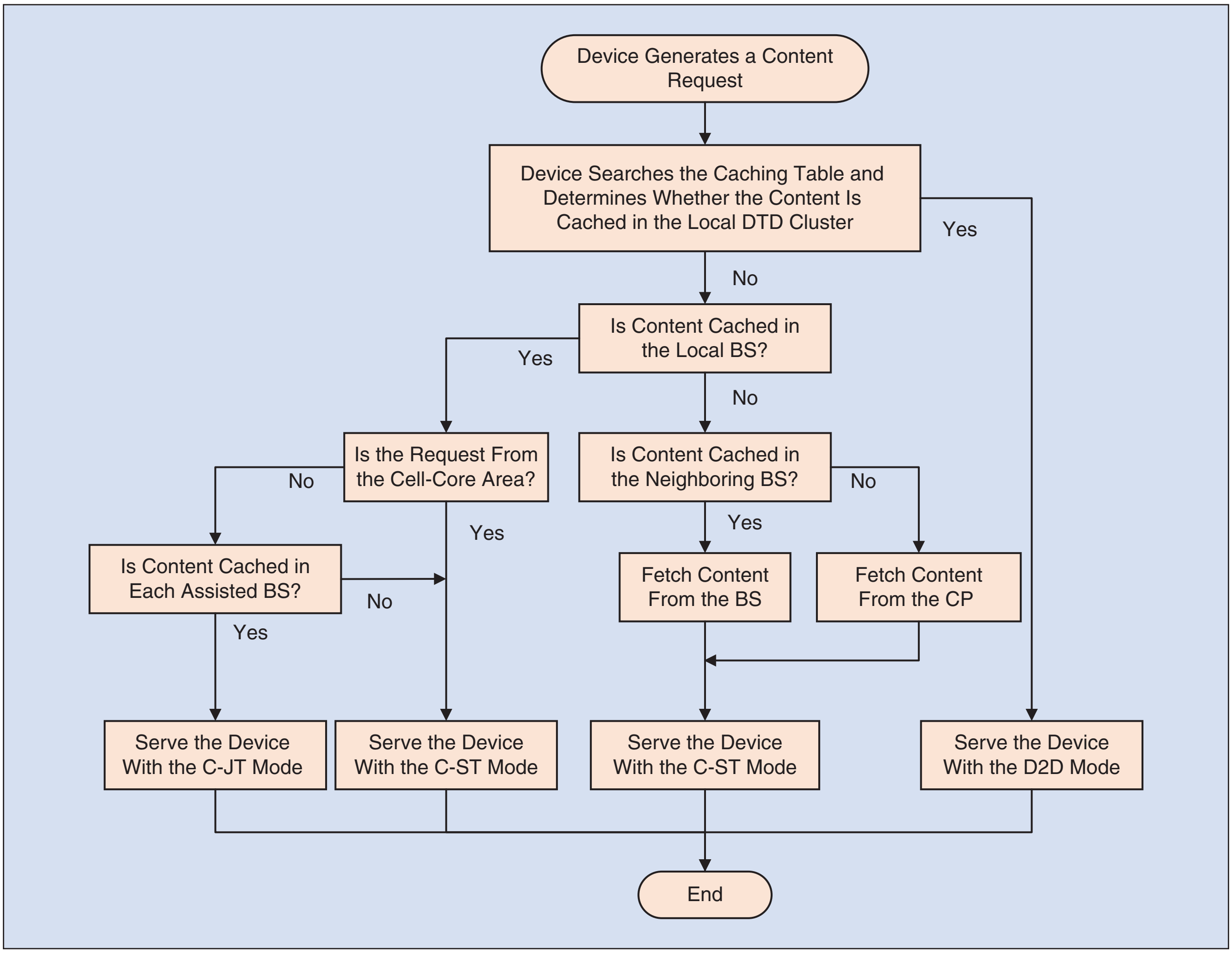
An emulator jointly developed with the MATLAB and C environments uses simulations to evaluate the



Virtual Cluster Formation
BS Evaluates the Average Minimum Distance $\bar{d}_V$ and Determines Whether $\bar{d}_V < d_{th}$
No
Cluster-Based Caching and Updating
Yes
Check the Caching Table and Determine Whether Content f_k is Locally Cached
No
Yes
Replace Old Content With f_k in Descending Order of Popularity
Broadcast and Receive Caching Messages in the Virtual Cluster
End

FIGURE 3 The caching and updating in a virtual cluster.

Authorized licensed use limited to: CHONGQING UNIV OF POST AND TELECOM. Downloaded on April 25,2022 at 06:40:32 UTC from IEEE Xplore. Restrictions apply.



Device Generates a Content Request
Device Searches the Caching Table and Determines Whether the Content Is Cached in the Local DTD Cluster
Yes
No
Is Content Cached in the Local BS?
Yes
No
Is the Request From the Cell-Core Area?
Is Content Cached in the Neighboring BS?
No
No
Yes
Yes
Is Content Cached in Each Assisted BS?
No
Yes
Fetch Content From the BS
Fetch Content From the CP
Serve the Device With the C-JT Mode
Serve the Device With the C-ST Mode
Serve the Device With the C-ST Mode
Serve the Device With the D2D Mode
End

FIGURE 4 The mode selection and content scheduling.

performance of the proposed schemes for the cellular, D2D, and cross tiers.

Evaluation for CoMP-Integrated Cellular Networks

This section evaluates the proposed caching scheme for the CoMP-integrated system. We consider five BSs with cellular topology under the centralized control of one SGW. The caching capacity of each BS is set to 5 GB. The coverage of each BS is divided into cell-core and cell-edge areas. The cellular channel has bandwidth $W = 100$ MHz. For simplicity, the received SINR of a cell-core device with a single cellular link is set to [3~15], and that of a cell-edge device is set to [1~3]. The transmit rate between adjacent BSs via X2 link is $r_{NE} = [20\sim50]$ Mb/s. The transmit rate in the core network is $r_{CP} = [5\sim10]$ Mb/s. The total number of multimedia contents in the network is 1,000, and each content is sized within [100~200] Mb. The Zipf parameter is $\alpha = 0.7$, and the request distribution coefficient is $\beta = 0.8$.

To demonstrate the superiority of the proposed scheme for the CoMP-integrated cellular network, we compare it with the caching schemes that support only storage-level cooperation (CSC) [13] and transmission-level cooperation (CTC) [14], respectively. The former strategy performs caching by taking account of the content sharing between caching nodes while ignoring the opportunity of performing JT. The latter uses the MPC strategy to promote JT opportunity but without considering joint optimization to reduce content redundancy. Moreover, we also implement the MPC strategy in an ST scenario (MPC-ST). This method can be used as a baseline model to evaluate the performance improvement from cooperative caching (i.e., either storage-level or transmission-level cooperation).

Figure 5 shows the normalized delay performance of the proposed scheme for the cellular network compared with the CSC, CTC, and MPC-ST schemes. The total cache size of BSs in the radio access networks (RANs) is set from 5 to 30%. The normalized content-delivery delay is denoted as the ratio of the content delivery delay under a caching strategy to that without caching. We can observe that the content-delivery delays of all of these schemes decrease with the increased caching capacity.

Authorized licensed use limited to: CHONGQING UNIV OF POST AND TELECOM. Downloaded on April 25,2022 at 06:40:32 UTC from IEEE Xplore. Restrictions apply.

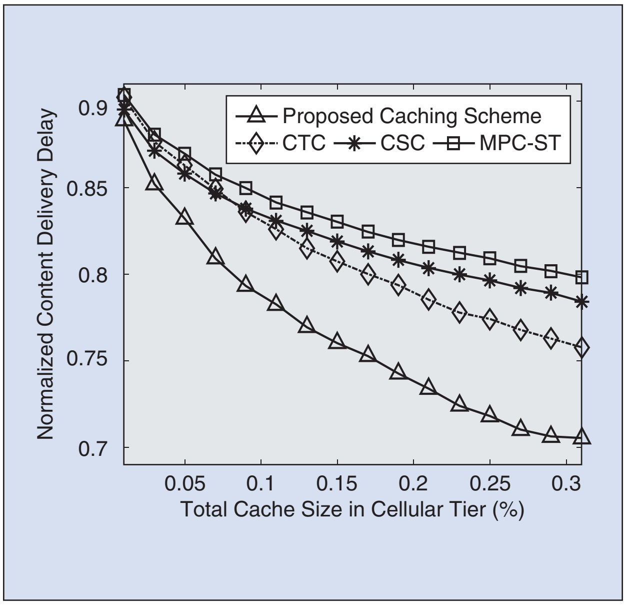


FIGURE 5 The normalized content delivery delay versus caching capacity in different caching schemes with $\beta = 0.8$.

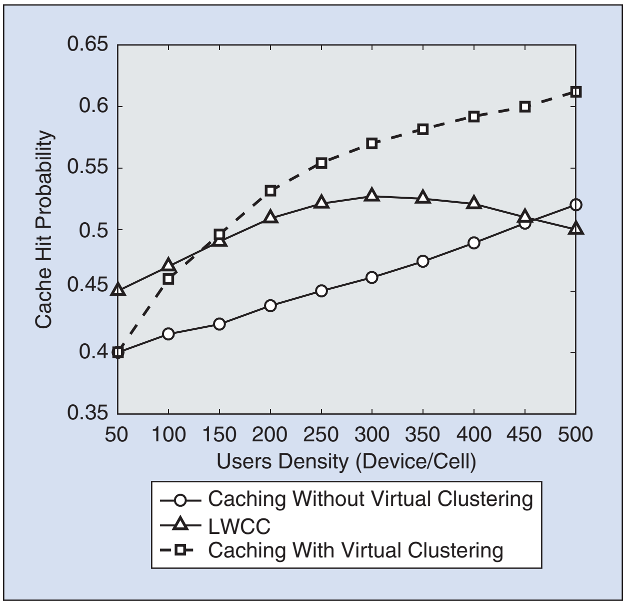


FIGURE 6 The cache hit probability versus device density for caching in D2D networks.

In addition, the CSC scheme has better delay performance than the MPC-ST scheme, illustrating that storage-level cooperation can squeeze more storage space to improve the cache hit probability. Besides, when the storage capacity of the cellular network is small (below 0.1), the CSC scheme holds similar delay performance compared with the CTC scheme. As the storage capacity increases, the CTC scheme outperforms the CSC scheme, demonstrating the performance improvement contributed by JTs. Meanwhile, as expected, the proposed jointly optimized caching scheme has a much better delay performance than all of the other schemes, especially when there is adequate storage space.

Evaluation for Virtual-Clustering-Based D2D Networks

We evaluated the cache-hit probability of the proposed cooperative caching strategy with and without virtual clustering. We also compared our scheme with a lightweight cooperation caching strategy (LWCC) [15]. In this scheme, the D2D devices and contents are assigned group IDs for efficient caching. The D2D devices cache contents based on matching their own group IDs with the content IDs. The radius of the cell coverage is taken as 200 m. We consider a HPPP for the device distribution. The device density (device/cell) is set from 50 to 500. The storage capacity of a device is set to 1 GB. The Zipf parameter is $\alpha = 0.7$. The threshold distance d_{th} for the virtual cluster is set to 10 m.

Figure 6 shows the variation in cache-hit probability with increasing device density. The increasing number of devices increases the chances of D2D communication, due to which the cache-hit probability is increased. On the other hand, the LWCC scheme performs better initially, with fewer devices. However, when the number of devices exceeds 150, its performance starts to decrease. The performance degrades with the increasing number of devices, while our scheme maintains a better performance even in dense environments, because D2D communication is limited to virtual clusters only. Hence, the proposed scheme increases the chances of successful D2D communication by avoiding large blockage regions. This is also shown through the higher cache hit probability with virtual clusters as compared to the one without virtual clustering.

Evaluation for Cross-Tier Networks

For cross-tier cooperative caching between the cellular and D2D tiers, we evaluate the delay performance of the

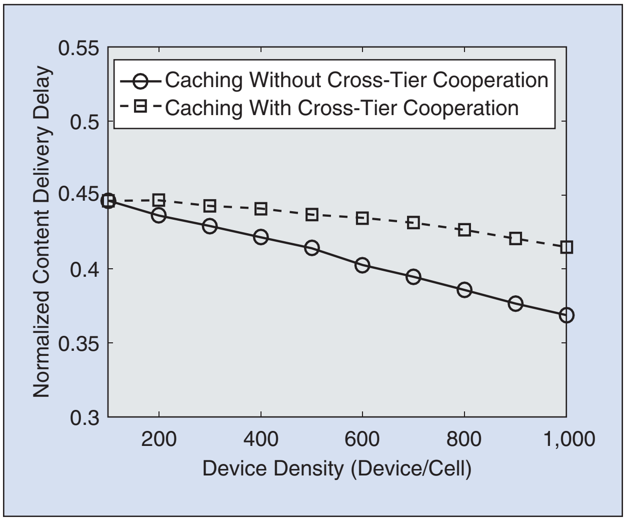


FIGURE 7 The normalized content delivery delay versus device density for caching with and without cross-tier cooperation.

Authorized licensed use limited to: CHONGQING UNIV OF POST AND TELECOM. Downloaded on April 25,2022 at 06:40:32 UTC from IEEE Xplore. Restrictions apply.

proposed scheme with and without cross-tier cooperation. The device density is set from 100 to 1,000. Figure 7 shows the delay performance of the schemes, with and without cross-tier cooperation, under increasing device density. It can be observed that caching with cross-tier cooperation significantly improves delay performance, and its advantage becomes more apparent as device density increases.

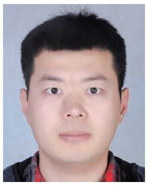
Conclusions

Exploiting caching in mobile networks is considered an important enabler for 5G wireless systems. This article presented a study on cooperative caching in HUDNs. We first provided an overview of existing research for caching in mobile networks. Then, to enable cooperative caching, we proposed an architecture that consists of CoMP-integrated ultradense cells and cluster-based D2D networks. The cache-enabled HUDN can provide high-quality services for users with the support of intelligent content caching, CoMP-JT, and cluster-based D2D content sharing. Under the proposed system architecture, we proposed three caching schemes for cellular networks, D2D networks, and D2D integrated cellular networks, respectively. For caching in cellular networks, we exploited storage- and transmission-level cooperation among BSs to jointly optimize content placement for ST and JT. For caching in D2D networks, we proposed a virtual cluster-formation scheme to improve content placement and D2D communication. In addition, to reduce content redundancy between the cellular and D2D networks, a cross-tier feedback scheme was proposed to improve utilization of limited storage space. We performed numerical simulations to evaluate the performance of the proposed caching schemes, and the results showed that our proposed schemes outperform the existing ones in terms of content delivery delay and cache hit probability.

Acknowledgments

Qingyang Song is the corresponding author for this article. This work was supported in part by the National Natural Science Foundation of China under grants 61771120 and 61775033 and the Fundamental Research Funds for the Central Universities under grants N161608001 and N171602002.

Author Information



Peng Lin (penglin11@foxmail.com) is pursuing his Ph.D. degree at Northeastern University, Shenyang, China. His research interests include content caching in mobile edge networks, device-to-device communications, cooperative transmissions, and machine learning. He is a Student Member of the IEEE.

Komal S. Khan (komal.khan@sydney.edu.au) is pursuing her Ph.D. degree at the University of Sydney, Australia. Her research interests include device-to-device (D2D) communications, content caching, and stochastic modelling for D2D caching.

Qingyang Song (songqy@cqupt.edu.cn) is a professor with Chongqing University of Post and Telecommunications, China. Her research interests include radio resource management, cognitive radio networks, cooperative communications, mobile caching, and wireless powering. She is a Senior Member of the IEEE.

Abbas Jamalipour (a.jamalipour@ieee.org) is the professor of ubiquitous mobile networking at the University of Sydney, Australia. His research interests include wireless communications, mobile cellular networks, and the Internet of Things. He is a Fellow of the IEEE.

References

[1] E. Bastug, M. Bennis, and M. Debbah, “Living on the edge: The role of proactive caching in 5G wireless networks,” *IEEE Commun. Mag.*, vol. 52, no. 8, pp. 82–89, Aug. 2014.

[2] 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.

[3] 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, Sept. 2017.

[4] M. Sheng, C. Xu, J. Liu, J. Song, X. Ma, and J. Li, “Enhancement for content delivery with proximity communications in caching enabled wireless networks: Architecture and challenges,” *IEEE Commun. Mag.*, vol. 54, no. 8, pp. 70–76, Aug. 2016.

[5] N. Golrezaei, A. F. Molisch, A. G. Dimakis, and G. Caire, “Femtocaching and device-to-device collaboration: A new architecture for wireless video distribution,” *IEEE Commun. Mag.*, vol. 51, no. 4, pp. 142–149, Apr. 2013.

[6] N. Rajatheva, *5G Mobile and Wireless Communications Technology.* Cambridge, U.K.: Cambridge Univ. Press, 2016.

[7] J. Jiang, S. Zhang, B. Li, and B. Li, “Maximized cellular traffic offloading via device-to-device content sharing,” *IEEE J. Select. Areas Commun.*, vol. 34, no. 1, pp. 82–91, Jan. 2016.

[8] F. Lázaro, E. Paolini, G. Liva, and G. Bauch, “Distance spectrum of fixed-rate raptor codes with linear random precoders,” *IEEE J. Select. Areas Commun.*, vol. 34, no. 2, pp. 422–436, Feb. 2016.

[9] S. Chu, J. Li, T. Liu, and F. Shu, “A contract-based incentive mechanism for data caching in ultra-dense small-cells networks,” in *Proc. IEEE Wireless Communications and Networking Conf.*, pp. 1–6. Mar. 2017.

[10] L. Yang, D. Wu, S. Xu, G. Zhang, and Y. Cai, “Social-energy-aware user clustering for content sharing based on D2D multicast communications,” *IEEE Access*, vol. 6, pp. 36,092–36,104, Jul. 2018.

[11] P. Yang, N. Zhang, S. Zhang, L. Yu, J. Zhang, and X. S. Shen, “Content popularity prediction towards location-aware mobile edge caching,” *IEEE Trans. Multimedia*, vol. 21, no. 4, pp. 915–929, Apr. 2019.

[12] S. H. Chae and W. Choi, “Caching placement in stochastic wireless caching helper networks: Channel selection diversity via caching,” *IEEE Trans. Wireless Commun.*, vol. 15, pp. 6626–6637, Oct. 2016.

[13] X. Li, X. Wang, K. Li, Z. Han, and V. C. M. Leung, “Collaborative multi-tier caching in heterogeneous networks: Modeling, analysis, and design,” *IEEE Trans. Wireless Commun.*, vol. 16, no. 10, pp. 6926–6939, Oct. 2017.

[14] Z. Chen, J. Lee, T. Q. S. Quek, and M. Kountouris, “Cooperative caching and transmission design in cluster-centric small-cell networks,” *IEEE Trans. Wireless Commun.*, vol. 16, pp. 3401–3415, May 2017.

[15] T. Shiroma, T. Nakajima, C. Wu, and T. Yoshinaga, “A light-weight cooperative caching strategy by D2D content sharing,” in *Proc. Fifth Int. Symp. on Computing and Networking*, pp. 159–165, 2017.

VT

Authorized licensed use limited to: CHONGQING UNIV OF POST AND TELECOM. Downloaded on April 25,2022 at 06:40:32 UTC from IEEE Xplore. Restrictions apply.