IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Digital Object Identifier

Content-Centric Mobile Edge Caching

TIANKUI ZHANG[1], (SENIOR MEMBER, IEEE), XINYUAN FANG[1], YUANWEI LIU[2], (SENIOR MEMBER, IEEE) AND ARUMUGAM NALLANATHAN[2], (FELLOW, IEEE)

[1]School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing, 100876, China
[2]School of Electronic Engineering and Computer Science Engineering, Queen Mary University of London, London, E1 4NS, U.K.

Corresponding author: Tiankui Zhang (e-mail: zhangtiankui@bupt.edu.cn).

This work was supported by the National Natural Science Foundation of China (No. 61971060 and No. 61502046).

ABSTRACT With the explosion of data volume, it becomes challenging to deliver high quality service to mobile users. Therefore, edge caching has received significant attentions since it can bring contents near mobile users, to boost spectral efficiency, and reduce backhaul load of mobile networks. Due to limited storage resources within mobile networks, it is important to improve efficiency of content management in edge caching. In this article, we propose an integrated content-centric mobile network framework for edge caching in 5G networks, which can leverage content-centric networking (CCN), achieve content-oriented information management, and increase content delivery efficiency. We elaborately design the cache-enabled mobile network architecture, CCN based function entities, CCN embedded protocol stack, and content retrieval process, and develop several effective approaches for tackling practical implementation constraints of CCN based edge caching. We demonstrate that our content caching strategies can significantly enhance edge caching performance. To further improve the performance of content-centric mobile edge caching, we identify promising open research directions.

INDEX TERMS Cache, content-centric networking, edge caching, mobile networks.

I. INTRODUCTION

RESEARCH on the next generation mobile networks (5G) is primarily driven by the combination of the expected increases in traffic capacity demands and the support for new use cases [1]. The future mobile Internet traffic will be dominated by content distribution and retrieval. With the explosion of data volume, it becomes challenging to deliver high quality service to the mobile users efficiently in fifth generation (5G) networks. It is predicted that data traffic in the global mobile cellular network will reach 2 ZB in 2021 [2], of which 71% of mobile data traffic is used for content distribution. In today's cellular networks, a large amount of mobile traffic is generated by social and mobile applications. However, a significant portion of such traffic consists of popular contents that are repeatedly transmitted to mobile users and unnecessarily consume extra backhaul bandwidth and resource of radio access networks (RAN). Although wireless technologies have been improved tremendously in the past couple decades, backhaul connections of RAN have not shown such rapid evolution. Consequently, backhaul networks become a bottleneck of content delivery in 5G networks. In order to meet the huge demands for content distribution in mobile cellular networks, the concept of edge caching has been proposed [3], [4]. Caching at wireless edges brings contents near mobile users through edge caching [5], [6], and therefore can boost spectral efficiency and reduce backhaul load of mobile networks [7]. Intuitively, caching is an efficient way to exploit the inherent content reuse while coping with the asynchronism among requests [6].

From a practical perspective, there are several approaches to cache contents at edge base stations (BSs) or user terminal (UTs) in wireless mobile networks. In [8], proactive caching at BSs in small cells alleviates backhaul usage, where files are cached during off-peak hours based on their popularity. UTs with storage space can be also used as caching nodes. Multiple UTs can collaboratively download and cache different parts of the same content from the serving BS and then share the cached content by device-to-device (D2D) communications [9], which can offload the backhaul traffic, reduce content access delay, and improve the user experience. Due to a limited amount of storage on each device, the main challenge is how backhaul traffic can be maximally offloaded by using D2D communications to satisfy content requests and to share messages among neighboring UTs [10]. From [11] and the references therein, it is demonstrated a carefully designed edge caching strategy is capable of improving network performance. Prior works [12]–[16] have developed some feasi-

ble frameworks on cache-enabled mobile networks. In [12], a big-data-aided cache architecture has been proposed, where a big data platform at the core site is in charge of tracking/predicting users' demands and cache-enabled BSs store strategic contents predicted by the big data platform. A general architecture of a mobile edge caching network is proposed in [13], which dynamically predicts and updates users' content demands using the matrix factorization technique. For the cooperative hierarchical caching framework in cloud RAN (C-RAN) in [14], contents are jointly cached at the central baseband unit (BBU) and at remote radio heads (RRHs). In [15], hierarchical-caching-based content-centric network architectures have been proposed. A edge caching architecture realized by service function chaining with the aid of edge computing techniques and virtualized resources is proposed in [16]. The above works pay more attention to the caching locations. However, in such cache-enabled mobile networks, proper network architecture, protocol for content delivery, and content-oriented information management are still open issues, and need to be further investigated.

In current wired and wireless networks, traditional caching, such as Web and content delivery network (CDN), is either application-oriented or deployed with static cache location. Therefore, it lacks of flexibility and scalability. Content-centric networking (CCN), which is a typical case of information-centric networking (ICN), provides a revolutionary architectures on efficient content delivery [17]. Different from the existing internet protocol (IP), which is host based content delivery techniques, CCN is aware of content name instead of the host address, utilizes cache space of every forwarding node as in-network caching, and is able to be deployed autonomously, regardless of applications [17]. Content caching has been research a lot in literatures in varying network frameworks, such as arbitrary network topology [18] and ad hoc networks [19]. The authors in [18] has proposed a caching placement algorithm for arbitrary topology in ICN. In content-centric wireless ad hoc networks, a cache space efficient caching scheme is proposed [19]. The potential advantages and significant research challenges of ICN in mobile wireless environments has been discussed in [20]. CCN based caching in 5G has shown performance gain of traffic offloading and content access delay reduction [8]. Content distribution on top of D2D technology based on ICN is studied in [21], which can provide seamless support to mobile services and decouple contents from node identifiers, thus providing a promising match with D2D requirements. The authors in [22] propose a novel framework with information-centric wireless virtualization and D2D communications. A device-level ICN architecture has been presented in [23] that is able to perform intelligent content distribution operations according to necessary context information on user mobility and content characteristics.

Cache-enabled mobile networks can leverage CCN to address content discovery, movement, delivery, management, and protection of information within content-oriented networks. Motivated by the efficiency and flexibility of content delivery paradigm in 5G, we integrate CCN with cache-enabled mobile networks for content delivery. The contributions of this article are summarized as below:

1) An integrated content-centric mobile network framework for edge caching in 5G networks is proposed. The essential parts of the proposed framework, namely, the cache-enabled mobile network architecture, CCN based function entities, CCN embedded protocol implementation, and content retrieval process are elaborated. In the proposed framework, the CCN protocol is embedded into UTs, BBU pools of BSs, as well as core networks.
2) Some practical implementation issues of CCN based edge caching are identified and some promising solutions are provided, including naming data content management in the content-centric mobile network framework, mobility management of cache-enabled cellular networks, edge caching incentive for mobile users, and radio resource management with edge caching in 5G networks.
3) An auction based caching strategy is introduced to facilitate edge caching at UTs via D2D communications in order to achieve content delivery efficiency by means of the performance gain of the traffic offloading and content access delay.

The rest of this paper is organized as follows. In Section II, we introduce the content-centric mobile framework. After discussing some practical implementation issues in Section III, we describe edge caching at UTs and evaluate performance in Section IV. Finally, we provide some open issues and conclusion remarks in Section V.

II. CONTENT-CENTRIC MOBILE NETWORK FRAMEWORK

In order to achieve content-oriented information management and increase content delivery efficiency, we propose a framework for CCN based edge caching in 5G networks. The proposed framework includes cache-enabled mobile network architecture, CCN based function entities, CCN embedded protocol implementation, and content retrieval process.

We consider cache-enabled mobile networks, as shown in Fig. 1, where both BSs and UTs are capable of storing content data in their local caches. When contents are cached at UTs, D2D communications can be utilized for content sharing and delivery among users.

A. CACHE-ENABLED MOBILE NETWORK ARCHITECTURE

At present, functions of BS mostly are divided by the BBU and the RRH. In 5G networks, the interface between them is changed from the circuit to the packet fronthaul [14]. The BBU splits into a distributed unit (DU) and a centralized unit (CU), where the DU is in charge of the physical layer and real-time medium access control (MAC) layer process while the CU focuses on higher layer computation. For the case of

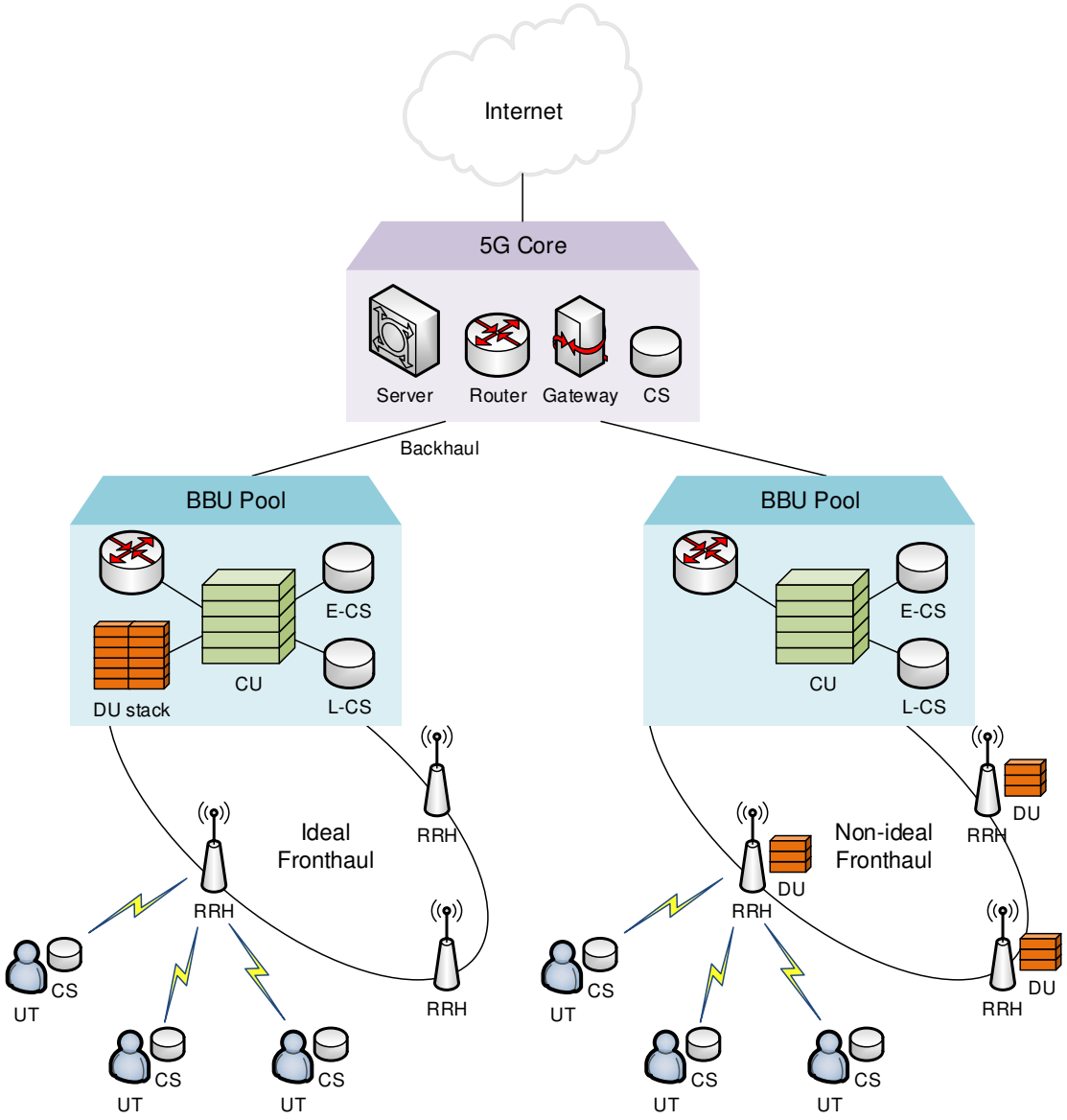


FIGURE 1: Cache-enabled mobile network architecture.

non-ideal fronthaul, the DU can be co-located with the RRH to reduce fronthaul traffic.

The cache-enabled mobile network architecture in our proposed framework is depicted as in Fig. 1, where traffic offloading gain is achieved by the cache instruments. The cache storage (CS) is embedded into UTs, BBU pools, and 5G core networks. The size of the CS in the UT is limited but is closest to content requests. As a result, users' requirements can be satisfied without occupying any fronthaul and backhaul resources. There are two kinds of CS in the BBU pool: local cache store (L-CS) for cached contents on the BBU and extend content store (E-CS) recording all the cached contents of UTs under its coverage. The CS in the core networks can save expense of inter-carrier communications or reduce fees for purchasing content from Internet.

B. FUNCTION ENTITY DEFINITION

The integration of CCN with cache-enabled mobile networks aims to improve the abilities of content management, such as content discovery and content cache. For this purpose, a CCN protocol is embedded into UTs, BBU pools, and 5G core networks. In CCN, content delivery is initiated by a consumer sending an content request, which is defined as Interest packet. Then the Interest is replied by either content provider or cache node by a series of Data packets [17]. The basic protocol of CCN contains three items:

- a forwarding information base (FIB) used to guide Interests routing towards Data;
- a pending interest table (PIT) to keep track of forwarded Interest that is not yet satisfied with a returned Data packet; and
- CS to store the cached content replicas.

The CCN functionality embedded in UTs and core networks can exploit the basic protocol of CCN directly. We design the functionalities of the CCN based BBU pool in the proposed framework since it is in charge of cache store allocation, content information management, and conversion the CCN to IP protocol. This is the main feature of CCN based edge caching. As shown in Fig. 2, the CCN based BBU pool in the proposed framework has four logical function entities: CCN core, intra-cell content registration, cache store allocation, and protocol conversion.

- **CCN core:** Based on CCN functionalities, CCN core maintains three fundamental data structures of CCN: a FIB, a PIT and a L-CS for local cached content replicas.
- **Intra-cell content registration:** The BBU pool uses an E-CS to record all the cached contents of the UTs under its coverage. In contrast to the L-CS, the E-CS does not maintain the metadata of each content. Instead, it records the owners of the contents. When a piece of record matches a request, a BBU pool will establish a D2D transmission link for the content consumer and the content provider so that the providers can reply requests of the consumers by D2D communications.
- **Content store allocation:** In the proposed framework, all available cache storage resources are virtualized as a pool. The content store allocation entity is in charge of reassigning suitable cache space for each CS to achieve an optimal edge caching performance. The CS size is a key parameter affecting cache hit ratio and content replacement rate. Therefore, it is essential to determine a suitable virtual CS for each cell according to the analysis of the user distribution and behaviors.
- **Protocol conversion:** When the BBU pool needs to communicate with the core network, an IP-embedded C-CN protocol conversion entity assists IP protocol translation and information interactive to support successful end-to-end communications.

C. CCN EMBEDDED PROTOCOL

The CCN embedded protocol stack from UTs to core networks are shown in Fig. 3. The layers in the figure including packet data convergence protocol (PDCP), radio link control (RLC), medium access control (MAC), physical (PHY), radio frequency (RF), GPRS tunnelling protocol user plane (GTP-U), etc.

- **Protocol of UT:** UTs are the exogenic request generators. The requests are initiated by the application layer and Interest packets are produced by the CCN layer.
- **Protocol of BBU:** The BBU is a bounder between RANs and core networks. Therefore, its protocol stack includes internal (on the left column) and external (on the right column) parts. The internal protocol helps extract the requested content name and other necessary information, which is used to match cached replicas and search for forwarding routes. If a request cannot be satisfied inside the BBU pool, it will be forwarded to the core network looking for content providers. The external protocol encapsulates the request to another CCN-based packet or an IP-based packet, according to the external network architecture. When an external packet is transmitted to the BBU pool, it processes the packet inversely and sends to the assigned RRH.
- **Protocol of core network:** Since carriers can generate data within their networks, some content servers are deployed in the core networks. Therefore, the protocol stack of the core networks covers application layer as well. The other layers in its protocol stack are same as the BBU's external protocol.

D. CONTENT RETRIEVAL PROCESS

As shown in Fig. 4, content dissemination in the proposed framework is triggered by the requests from UTs. The forwarding process goes upwards from UTs to its affiliated BBU, and further up to the core networks, even to Internet step-by-step. At each step, the request process is terminated and the demanded content is replied along with the opposite way of the request if a request is able to be satisfied either by server or caching. Otherwise, the request keeps on forwarding upwards. In this section, we discuss the procedures of content retrieval given that cache hits BBU pool's L-CS and E-CS and the core networks.

- **Hit at BBU L-CS:** When an Interest is generated by UT.B and mis-matches to L-CS.B (as well as PIT.B), it is forwarded to the BBU pool through the RRH, as shown in case 1 of Fig. 4. If the targeting content exists in L-CS.BBU, the Data packet is sent back to UT.B. In this case, only the fronthaul sustains the traffic while the backhaul is liberated.
- **Hit at BBU E-CS:** If L-CS.BBU does not match an Interest, the BBU searches its own E-CS. The E-CS records all cached contents in the L-CS of the UTs within its coverage. For instance, UT.A in Fig. 4 sends a request of content C3 to the BBU pool. L-CS.BBU does not include C3 while E-CS.BBU records the targeting data name and its owner ID, which is UT.B, as shown in case 2 of Fig. 4. Then, the BBU builds a D2D transmission link between UT.B and UT.A for the content C3 delivery. After receiving C3, UT.A is able to find a record of delivering content C3 to its application layer. This record is generated when UT.A sends an Interest to the BBU and is utilized to process Data packet delivered by UT.B. It reflects the advantage of the CCN architecture focusing on the content itself instead of its provider. In this circumstance, content retrieval is completed by cooperation between the UTs with D2D communications and there is no backhaul link cost.
- **Hit at core network:** If neither L-CS.BBU nor E-CS.BBU includes the required content of UT.C, Interest of UT.C is forwarded to the core network through the backhaul link. When a content replica is cached in L-CS.Core, it will be replied backwards to UT.C, for instance, case 3 of Fig. 4. Otherwise, Interest of UT.C

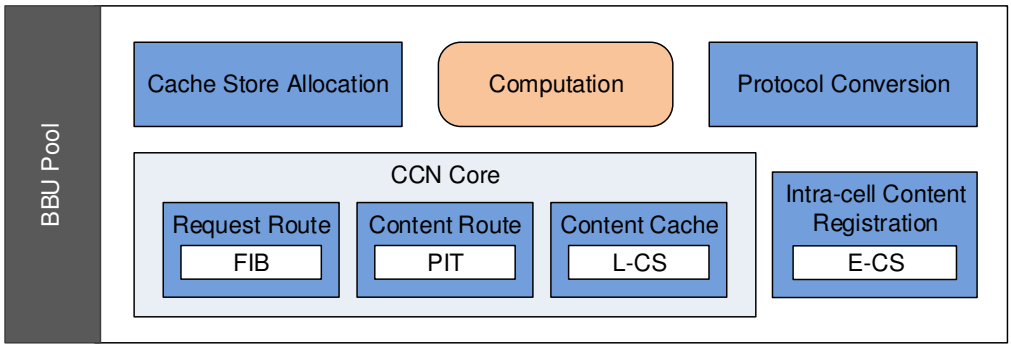


FIGURE 2: Function entities in CCN based BBU pool.

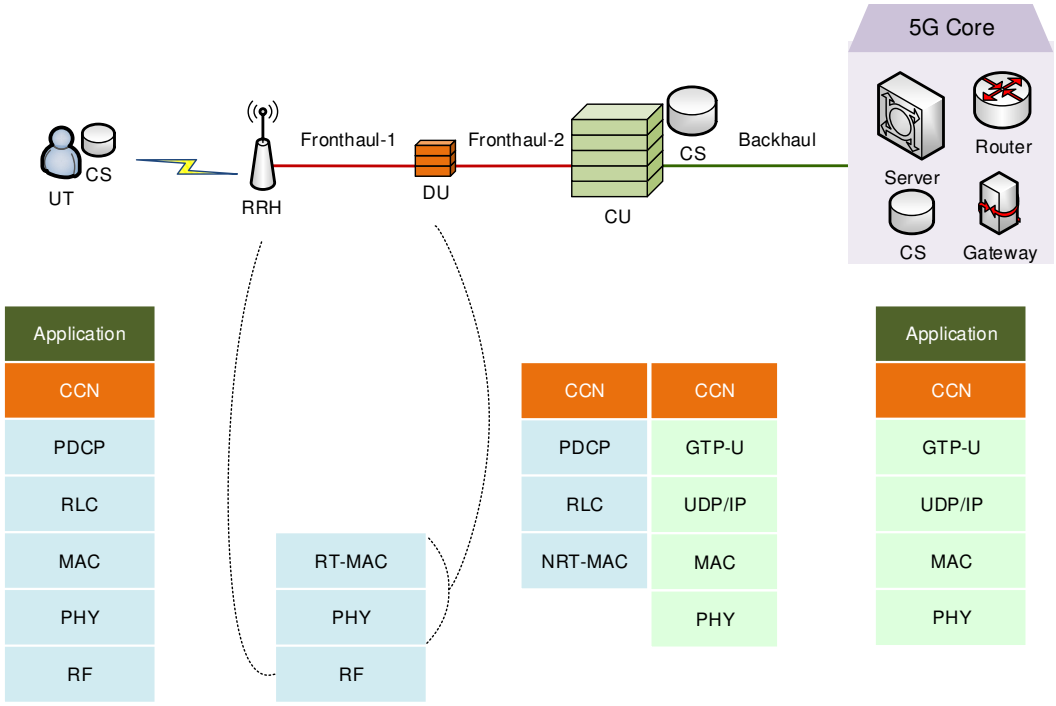


FIGURE 3: CCN embedded protocol stack of 5G mobile networks.

will be forwarded to the Internet. When the content is delivered from the Internet to UT.C, a content replica will be cached in the L-CS.Core for future requests. The caching decision depends on the prediction of the popularity distribution and user preference. Big data analysis is a promising solution [9].

The aforementioned framework provides a feasible solution to realize CCN but with minimum changes in the current communication protocols and network infrastructures. The conventional IP-based mobile network architecture is mostly reserved. All non-content dissemination services are processed just as usual.

III. PRACTICAL IMPLEMENTATION CONSIDERATIONS

Although CCN based edge caching in 5G networks potentially alleviates backhaul traffic load and reduces content access delay, it also imposes several implementation constraints in practical scenarios, such as naming data management, mobility and incentive of UTs, radio resource management with edge caching, etc. In this section, we will provide some effective solutions to these implementation constraints.

A. NAMING DATA MANAGEMENT

Data naming is an important feature of CCN. Particularly, CCN requires a unique name for each individual data object to identify content regardless of its location and provider. Therefore, there should be an entity maintaining and updating naming data list. Both the BSs and UTs are limited by small cache space compared to a huge number of contents from Internet and the content replacement will occur frequently with the varying popularity. As a result, the naming data list will update frequently. Prediction of the popularity distribution and user preference using big data analytics is a

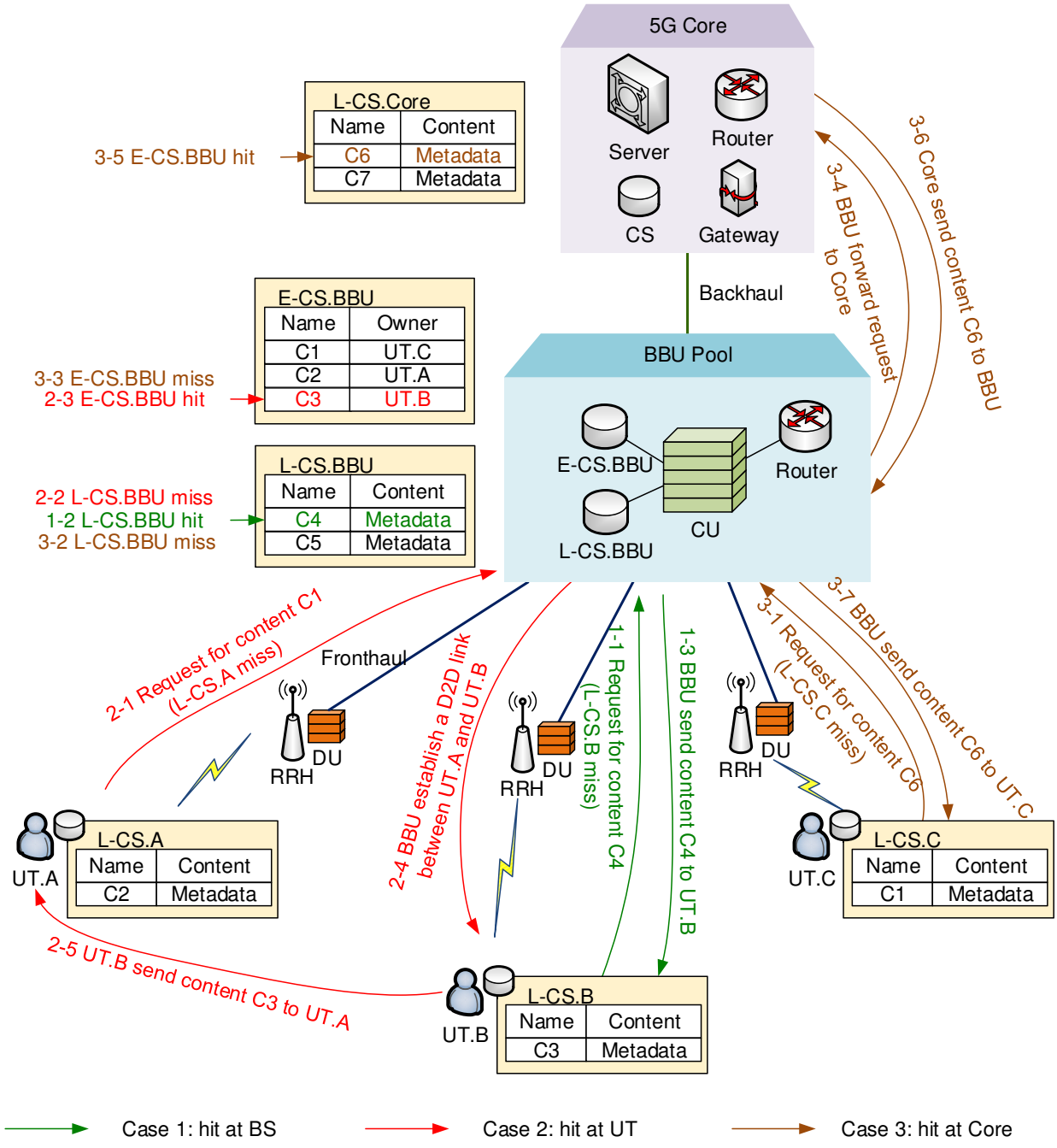


FIGURE 4: Content retrieval process of content-centric mobile networks.

promising solution [12], [13]. Statistical patterns of content requests, both in aggregate form and on a per-user basis, can be predicted so that most popular contents can be maintained in the naming data list proactively.

B. MOBILITY MANAGEMENT

In content-centric cellular networks, there are two kinds of UT mobility: content provider mobility and content consumer mobility, depending on whether a UT acts as a provider or a consumer. Content-centric cellular networks can inherently support the mobility of the content providers as content replicas can be maintained in networks by edge caching. For the content consumer mobility, a mechanism that triggers the discovery of new retrieval providers for the same content clearly outperforms a tunnel-based approach in the conventional cellular networks. One challenge is how to capitalize on location information while facilitating the naming and caching of CCN. Another challenge is how to maintain and update naming data lists when UTs' moving leads to a rapid topological changes. Fortunately, the naming data lists are maintained in the BBU pools by L-CS in Fig. 2, which can be updated locally and rapidly, and address the issue.

C. EDGE CACHING INCENTIVE FOR UT

From the perspective of network performance improvement, contents can be cached by the UTs for D2D content sharing. However, the nature of selfishness of mobile users becomes a major obstacle to utilize edge caching to offload data traffic of backhaul networks. Considering the heterogeneous preference, each user cares merely about his/her own preference and only caches the contents he/she likes most, which may result in duplicating caching and underutilization of cache spaces for all UTs. Therefore, it is essential to introduce incentive mechanisms into networks to motivate the UTs to cache contents for sharing and delivery. Nonetheless, D2D transmission and caching depend on the users in the vicinity willing to assist, which might change with time and specifically with the state of their batteries. Consequently, the design of incentive mechanisms should consider energy consumed by D2D transmission and caching, social relationships, and preference similarity of adjacent users. Game theory and auction are feasible approaches for incentive mechanism designs.

D. RADIO RESOURCE MANAGEMENT WITH EDGE CACHING

Edge caching is capable of reducing both backhaul link load and content access delay. Meanwhile, either BS or UT caching will induce resource allocation problems. The comprehensive consideration of caching optimization and efficient resource allocation is a promising solution to the network performance bottleneck suffered by insufficient radio resources. The optimal caching BS can be selected to optimize both content placement and user association, thereby alleviating the content access delay and backhaul traffic load. In the case of caching at the UTs, it is necessary to consider D2D power control and channel allocation to determine transmission data rates and coverage of the UTs. Although high transmission power increases data rates as well as enlarge coverage of the UTs, a trade-off needs to be carefully considered. Ideally more time frequency resources should be allocated to the UT that caches more popular contents to improve the transmission data rate and further reduce the content access delay. Since most of the optimization objectives, such as energy consumption minimization, spectrum efficiency maximization, and content access latency minimization, are conflicting with each other, designing reasonable utility functions to compromise the above optimization objectives can be a promising direction.

IV. EDGE CACHING AT UTS AND PERFORMANCE EVALUATION

Edge caching at the BSs in small cells has been discussed recently [8], [24], where performance gains of backhaul traffic offloading, content access delay, and cache hit ratio have been investigated. In this section, we focus on performance improvement of edge caching at UTs with D2D communications. Since the incentive is one of the main challenges of Edge caching at UTs, we propose an auction based caching placement in cache-enabled D2D networks.

A. AUCTION BASED CACHING PLACEMENT

In order to motivate UTs for content caching and sharing, the caching placement at UTs is modeled as an auction process. In our content auction model, the buyers (bidders) are the UTs, the seller is the BS, the auction products are the contents, and the auction is hosted by a third-party agent, such as, a functional entity deployed in BBU in the proposed framework. Since we assume that there is a centralized agent entity for cache management, it is natural to implement the cache consistency maintenance and named contents management in this centralized agent entity.

In this article, we consider the revenue of all the UTs as social welfare, so the social efficiency can be ensured by maximizing the sum revenue of UTs. Considering individual rationality and truthfulness, the auction model we use is the second-price sealed bid auction, which can guarantee the truthfulness of bidders' bids because the winner's payment is not determined by themselves but by the second highest loser. In the proposed auction model for content caching and sharing, the revenue of UT n for caching content m is $v_n^{(m)}$, the bid of UT n for content m is $b_n^{(m)}$, and the winner UT n needs to pay $p_n^{(m)}$ for caching the content m. If UT n caches the content m, $x_n^{(m)} = 1$; $x_n^{(m)} = 0$ otherwise. We consider that $b_n^{(m)} = v_n^{(m)}$ because the second-price sealed bid auction model can guarantee the truthfulness of the bidders. The UTs win the same content form a winner UT set and each set can be viewed as one bidder. There are M winner UT sets for M content chunks, which means the M content chunks are cached in the M winner UT sets, then the caching placement vectors are denoted as $x^{(1)}$, $x^{(2)}$, ..., $x^{(M)}$. Correspondingly, there are M payment price vectors $p^{(1)}$, $p^{(2)}$, ..., $p^{(M)}$.

There is a predefined "cache conflict" restriction relationship of UTs participating in the auction process. The "cache conflict" restriction can improve the cache utility by avoiding caching the same contents repeatedly on multiple proximal UTs. When the physical distance of two UTs is small, the content sharing can be achieved by D2D communications between these two UTs. In such a case, caching the same content in these two UTs leads a waste of the cache space and diminishes the content diversity in the networks. The UT "cache conflict" restriction matrix is defined as $E = \{E_{n,n'}\}$, where $E_{n,n'} = 1$ when the transmission signal power level between UT n and UT n' is smaller than a predefined threshold γ. Otherwise, $E_{n,n'} = 0$. When $E_{n,n'} = 1$, UT n' is regarded as a neighbor of UT n, in this case, the UT n and UT n' cannot cache the same content, i.e., $x_n^{(m)} + x_{n'}^{(m)} \leq 1$.

Then, we define the revenue of UT n for content m considering the number of UTs in its D2D communication range, the average transmission rate between the UTs, and their preference for content m. The revenue of UT n for content m in its neighboring UT set $\mathcal{N}_n$ is defined as,

$$v_n^{(m)} = \sum_{n' \in \mathcal{N}_n} f_{n'}^{(m)} r_{n,n'} E_{n,n'}, \tag{1}$$

where $f_{n'}^{(m)}$ denotes the preference of UT n' for content m,

$r_{n,n'}$ denotes the average transmission rate between UT n and n'. It is assumed that each UT can cache one content in one auction. The caching placement indicator $x_n^{(m)}$ can be obtained by maximizing the revenue of all the UTs,

$$U^* = \max_{x^{(1)},x^{(2)},\ldots,x^{(M)}} \sum_{m=1}^{M}\sum_{n=1}^{N} v_n^{(m)} x_n^{(m)} \quad (2a)$$

$$s.t.\ x_n^{(m)} + x_{n'}^{(m)} \le 1,\ \forall n,n'\ if\ E_{n,n'} = 1,\ \forall m, \quad (2b)$$

$$\sum_{m=1}^{M} x_n^{(m)} s \le s_0,\ \forall n, \quad (2c)$$

$$x_n^{(m)} = \{0,1\},\ n \in \mathcal{N},\ m \in \mathcal{M}, \quad (2d)$$

where (2b) denotes the "cache conflict" restrict relationship of UT n and UT n', (2c) constrains that the cached content chunks in a UT cannot exceed its cache space s_0, and (2d) constrains that $x_n^{(m)}$ is a binary variable. By solving this problem to maximum total revenue of all UTs for content caching and sharing, we obtain the optimal caching placement $x^{(1)}$, $x^{(2)}$, ..., $x^{(M)}$ and winner sets for different content chunks $\mathcal{N}^{(1)}$, $\mathcal{N}^{(2)}$, ..., $\mathcal{N}^{(M)}$. In general, the cache space of a UT is limited, so it can only cache a few content chunks, as indicated in [25]. In this paper, for simplicity but without loss of generality, it is assumed that one UT can cache one content chunk with a size of s. If a UT has cache space more than one chunk, it is equivalent to multiple UTs with a unified cache space size s. Since this problem is a maximal weighted independent set (MWIS) problem, we use the semidefinite programming (SDP) relaxation method to obtain a suboptimal solution [26].

In the proposed auction model, the UTs in each set can be viewed as one bidder and the winner set is a set of UTs with the highest total bids. Hence, the pricing strategy of the proposed auction can be obtained using the same way as the pricing strategy of second-price sealed bid auction. According to the second-price sealed bid auction mechanism, the pricing process is as follows:

1) First, the problem (2) is solved to get the winner sets, $\mathcal{W} = \cup_{m=1}^{M}\mathcal{N}^{(m)}$, that includes all content chunks.

2) The winner sets, $\mathcal{W}$, are removed from the buyer set, and content chunk m is auctioned again in the remaining loser set, $\hat{\mathcal{N}} = \mathcal{N}\backslash\mathcal{W}$, from which, we can calculate the maximum total actual revenue from (2), denoted $U_{-\mathcal{W}}^{(m)*}$, for the winner set $\mathcal{N}^{(m)}$. In our auction model, $U_{-\mathcal{W}}^{(m)*}$ is set as the total price that the winners of content chunk m need to pay to the BBU.

Since we consider the revenue of all the UTs as social welfare, so the social efficiency can be ensured. At the same time, only when the UT's revenue for caching a chunk is greater than 0, it will participate in the auction. This assumption is in line with individual rationality of each UT. Finally, the second-price sealed bid auction can guarantee the auction's truthfulness.

B. PERFORMANCE EVALUATION

The proposed caching placement method in cache-enabled D2D networks is simulated by MATLAB. In the simulation, N UTs are randomly distributed in the macrocell and can communicate with any neighbor UTs within the coverage. The popularity of M contents follows a Zipf-like distribution [8] with parameter $\alpha = 1$. The size of each content is set to one M bytes. We use $f_n^{(m)}$ to model the data request of UT n for content m. For UT n, we assume that $\sum_{m=1}^{M} f_n^{(m)} = 1$. The BBU makes the caching placement decision with the statistical knowledge of content preference of UTs. In the simulation, the parameter settings and channel model used for D2D communications follow the Technical Report of 3GPP [27]. The main parameters are listed in Table I.

TABLE 1: Simulation Parameters

Parameter	Value
Carrier frequency	2 GHz
Radio bandwidth	20 MHz
Backhaul data rate	1.5 Mbps
D2D transmit power	23 dBm
BS transmit power	43 dBm
Pathloss from BS to UT	37.6log10(d[km])+128.1 dB
Pathloss of D2D channel	40log10(d[km])+148 dB
Noise power spectral density	-174 dBm/Hz

We compare the caching performance of the proposed auction based caching (ABC) with the social-aware caching game (SAGG) [28] and random caching (RC). In RC, each UT randomly selects a content for caching. The criteria of the caching performance are average content access delay and the BBU traffic offloading ratio [26].

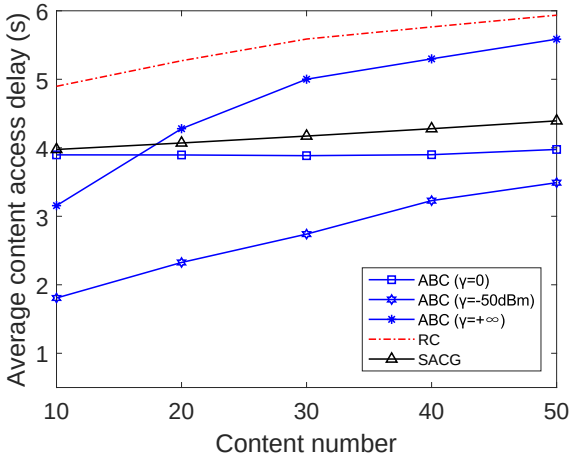


FIGURE 5: Content access delay comparison with varying content numbers.

The caching performance comparison in term of varying content numbers is shown in Fig. 5 and Fig. 6. We simulate the performance of ABC when the transmission signal power level threshold γ is equal to 0, ∞ and 50 dB, respectively. When $\gamma = 0$, there is no "cache conflict" between the UTs.

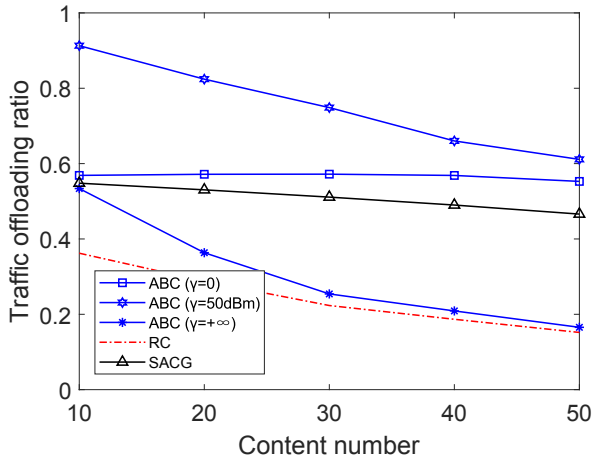


FIGURE 6: Traffic offloading ratio comparison with varying content numbers.

When $\gamma = \infty$, the "cache conflict" condition between UTs is very strict, that is, only one content can be cached in the entire macrocell region regardless of whether there is D2D communications capability between UTs. The simulation results of Fig. 5 demonstrate that:

- No matter which algorithm is used for UT edge caching at off-peak hours, the average content access delay for content delivery during peak hours can be reduced since the average content access delay by the RRH transmission without caching during the peak period is 6.3 s in our simulation;
- Compared with SACG and RC, the proposed ABC can effectively reduce the traffic load of the BBU during peak hours and the average content access delay of UTs; when $\gamma = 0$ and $\gamma = \infty$, the BS traffic offloading performance of ABC is close to RC and SACG, respectively;
- It was demonstrated that the proposed caching placement algorithm solving the problem of user selfishness through the natural social efficiency and individual rationality of the proposed auction model;
- Based on the users' preference and the "cache conflict" restrict relationship of the UTs, the proposed algorithm have selected different independent UT sets to cache different contents, which effectively avoids the waste of cache space.

V. OPEN ISSUES AND CONCLUSION

In this article, we have proposed a content-centric mobile network framework for edge caching in 5G networks. Several implementation issues as well as the corresponding potential solutions of CCN based edge caching in 5G have been discussed. We have shown the edge caching performance gain via the proposed auction based caching strategy. Furthermore, it is also desirable to study the CCN based edge caching in some specific scenarios, such as wireless powered cellular networks, drone-aided cellular networks. Meanwhile, security of content-centric mobile networks and user privacy protection of content sharing are significant topics. Besides that the following interesting open issues are worthy of further study.

- **Cache store allocation**: The cache storage space allocation among the BBU and the UT greatly impacts the edge caching performance of mobile networks. It is worthy to be investigated carefully, especially under different optimization criteria, such as backhaul traffic offloading maximization, content access delay minimization, and energy efficiency maximization.
- **Edge caching and computing**: Edge caching and computing combination is an emerging topic in recent research. Joint management of caching and computing resources is an effective way to improve the network performance and enhance the mobile network resource utilization.
- **Big data-driven caching**: Most of the existing works focus on the proactive caching in static environments. However, considering the dynamic of content popularity, user preference and user position, a big data-driven learning based caching placement may be more effective and flexible in practical scenarios.

REFERENCES

[1] J.-C. Guey, P.-K. Liao, Y.-S. Chen, A. Hsu, C.-H. Hwang, and G. Lin, "On 5G radio access architecture and technology [industry perspectives]," IEEE Wireless Commun., vol. 22, no. 5, pp. 2–5, Oct. 2015.

[2] "Cisco visual networking index: Global mobile data traffic forecast update, 2016 to 2021 white paper," https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html, Mar. 2017.

[3] 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.

[4] 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.

[5] D. Liu, B. Chen, C. Yang, and A. F. Molisch, "Caching at the wireless edge: design aspects, challenges, and future directions," IEEE Commun. Mag., vol. 54, no. 9, pp. 22–28, Sep. 2016.

[6] C. Yang, Y. Yao, Z. Chen, and B. Xia, "Analysis on cache-enabled wireless heterogeneous networks," IEEE Trans. Wireless Commun., vol. 15, no. 1, pp. 131–145, Jan. 2016.

[7] G. Paschos, E. Bastug, I. Land, G. Caire, and M. Debbah, "Wireless caching: technical misconceptions and business barriers," IEEE Commun. Mag., vol. 54, no. 8, pp. 16–22, Aug. 2016.

[8] 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.

[9] M. Ji, G. Caire, and A. F. Molisch, "Wireless device-to-device caching networks: Basic principles and system performance," IEEE J. Select. Areas Commun., vol. 34, no. 1, pp. 176–189, Jan. 2016.

[10] D. Malak, M. Al-Shalash, and J. G. Andrews, "Optimizing content caching to maximize the density of successful receptions in device-to-device networking," IEEE Trans. Wireless Commun., vol. 64, no. 10, pp. 4365–4380, Oct. 2016.

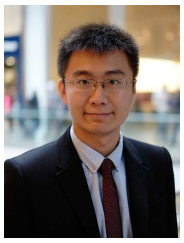
[11] T. Zhang, H. Fan, J. Loo, and D. Liu, "User preference aware caching deployment for device-to-device caching networks," IEEE Syst. J., vol. 13, no. 1, pp. 226–237, Mar. 2019.

[12] E. Zeydan, E. Bastug, M. Bennis, M. A. Kader, I. A. Karatepe, A. S. Er, and M. Debbah, "Big data caching for networking: moving from cloud to edge," IEEE Commun. Mag., vol. 54, no. 9, pp. 36–42, Sep. 2016.

[13] D. T. Hoang, D. Niyato, D. N. Nguyen, E. Dutkiewicz, P. Wang, and Z. Han, "A dynamic edge caching framework for mobile 5G networks," IEEE Wireless Commun., vol. 25, no. 5, pp. 95–103, Oct. 2018.
[14] T. X. Tran, A. Hajisami, and D. Pompili, "Cooperative hierarchical caching in 5G cloud radio access networks," IEEE Network, vol. 31, no. 4, pp. 35–41, Jul. 2017.
[15] X. Zhang and Q. Zhu, "Hierarchical caching for statistical qos guaranteed multimedia transmissions over 5g edge computing mobile wireless networks," IEEE Wireless Commun., vol. 25, no. 3, pp. 12–20, Jun. 2018.
[16] L. Lei, X. Xiong, L. Hou, and K. Zheng, "Collaborative edge caching through service function chaining: Architecture and challenges," IEEE Wireless Commun., vol. 25, no. 3, pp. 94–102, Jun. 2018.
[17] B. Ahlgren, C. Dannewitz, C. Imbrenda, D. Kutscher, and B. Ohlman, "A survey of information-centric networking," IEEE Commun. Mag., vol. 50, no. 7, pp. 26–36, Jul. 2012.
[18] S. Shan, C. Feng, T. Zhang, and J. Loo, "Proactive caching placement for arbitrary topology with multi-hop forwarding in ICN," IEEE Access, vol. 7, pp. 149 117–149 131, 2019.
[19] T. Zhang, X. Xu, Le Zhou, X. Jiang, and J. Loo, "Cache space efficient caching scheme for content-centric mobile ad hoc networks," IEEE Systems Journal, vol. 13, no. 1, pp. 530–541, Mar. 2019.
[20] C. Fang, H. Yao, Z. Wang, W. Wu, X. Jin, and F. R. Yu, "A survey of mobile information-centric networking: Research issues and challenges," IEEE Commun. Surveys Tuts., vol. 20, no. 3, pp. 2353–2371, thirdquarter 2018.
[21] C. Xu, M. Wang, X. Chen, L. Zhong, and L. A. Grieco, "Optimal information centric caching in 5G device-to-Device communications," IEEE Trans. on Mobile Comput., vol. 17, no. 9, pp. 2114–2126, Sep. 2018.
[22] K. Wang, F. R. Yu, and H. Li, "Information-centric virtualized cellular networks with device-to-Device communications," IEEE Trans. Veh. Tech., vol. 65, no. 11, pp. 9319–9329, Nov. 2016.
[23] G. Chandrasekaran, N. Wang, M. Hassanpour, M. Xu, and R. Tafazolli, "Mobility as a service (MaaS): A D2D-based information centric network architecture for edge-controlled content distribution," IEEE Access, vol. 6, pp. 2110–2129, 2018.
[24] Z. Zheng, L. Song, Z. Han, G. Y. Li, and H. V. Poor, "A stackelberg game approach to proactive caching in large-scale mobile edge networks," IEEE Trans. Wireless Commun., vol. 17, no. 8, pp. 5198–5211, Aug. 2018.
[25] S. Umrao, A. Roy, N. Saxena, S. Singh, and J.-i. Jung, "Mobile network operator and mobile user cooperation for customized D2D data services," Journal of Network and Systems Management, vol. 26, no. 4, pp. 878–903, Oct. 2018.
[26] X. Fang, T. Zhang, Y. Liu, and Y. G. Li, "Multi-winner auction based mobile user edge caching in D2D cellular networks," in 2019 IEEE ICC, May 2019.
[27] 3GPP, "Technical specification group radio access network; study on LTE device to device proximity services," TR 36.843, Release 12, pp. 33–46, 2014.
[28] K. Zhu, W. Zhi, X. Chen, and L. Zhang, "Socially motivated data caching in ultra-dense small cell networks," IEEE Network, vol. 31, no. 4, pp. 42–48, Jul. 2017.

TIANKUI ZHANG (M'10-SM'15) received the Ph.D. degree in Information and Communication Engineering and B.S. degree in Communication Engineering from Beijing University of Posts and Telecommunications (BUPT), China, in 2008 and 2003, respectively. Currently, he is a Professor in School of Information and Communication Engineering at BUPT. His research interests include wireless communication networks, mobile edge computing and caching, signal processing for wireless communications, content centric wireless networks. He had published more than 100 papers including journal papers on IEEE Journal on Selected Areas in Communications, IEEE Transaction on Communications, etc., and conference papers, such as IEEE GLOBECOM and IEEE ICC. He is an Editor for IEEE Access.

XINYUAN FANG received the B.S. degree in communications engineering from Beijing University of Posts and Telecommunications, China, in 2017. She is currently pursuing the M.E. degree in electronics and communication engineering from Beijing University of Posts and Telecommunications. Her current research focuses on caching placement in D2D-enabled cellular networks.



YUANWEI LIU (S'13-M'16-SM'19) received the B.S. and M.S. degrees from the Beijing University of Posts and Telecommunications in 2011 and 2014, respectively, and the Ph.D. degree in electrical engineering from the Queen Mary University of London, U.K., in 2016. He was with the Department of Informatics, King's College London, from 2016 to 2017, where he was a Post-Doctoral Research Fellow. He has been a Lecturer (Assistant Professor) with the School of Electronic Engineering and Computer Science, Queen Mary University of London, since 2017.

His research interests include 5G and beyond wireless networks, the Internet of Things, machine learning, and stochastic geometry. He has served as a TPC Member for many IEEE conferences, such as GLOBECOM and ICC. He received the Exemplary Reviewer Certificate of IEEE WIRELESS COMMUNICATIONS LETTERS in 2015, IEEE TRANSACTIONS ON COMMUNICATIONS in 2016 and 2017, and IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS in 2017 and 2018. He has served as the Publicity Co-Chair for VTC 2019-Fall. He is currently an Editor on the Editorial Board of the IEEE TRANSACTIONS ON COMMUNICATIONS, IEEE COMMUNICATIONS LETTERS, and IEEE ACCESS. He also serves as a Guest Editor for IEEE JSTSP special issue on Signal Processing Advances for Non-Orthogonal Multiple Access in Next Generation Wireless Networks.

ARUMUGAM NALLANATHAN (S'97-M'00-SM'05-F'17) is Professor of Wireless Communications and Head of the Communication Systems Research (CSR) group in the School of Electronic Engineering and Computer Science at Queen Mary University of London since September 2017. He was with the Department of Informatics at King's College London from December 2007 to August 2017, where he was Professor of Wireless Communications from April 2013 to August 2017 and a Visiting Professor from September 2017. He was an Assistant Professor in the Department of Electrical and Computer Engineering, National University of Singapore from August 2000 to December 2007. His research interests include 5G Wireless Networks, Internet of Things (IoT) and Molecular Communications. He published nearly 400 technical papers in scientific journals and international conferences. He is a co-recipient of the Best Paper Awards presented at the IEEE International Conference on Communications 2016 (ICC'2016) and IEEE Global Communications Conference 2017 (GLOBECOM'2017). He is an IEEE Distinguished Lecturer. He has been selected as a Web of Science Highly Cited Researcher in 2016.

He is an Editor for IEEE Transactions on Communications. He was an Editor for IEEE Transactions on Wireless Communications (2006-2011), IEEE Transactions on Vehicular Technology (2006-2017), IEEE Wireless Communications Letters and IEEE Signal Processing Letters. He served as the Chair for the Signal Processing and Communication Electronics Technical Committee of IEEE Communications Society and Technical Program Chair and member of Technical Program Committees in numerous IEEE conferences. He received the IEEE Communications Society SPCE outstanding service award 2012 and IEEE Communications Society RCC outstanding service award 2014.