Cooperative Content Caching in 5G Networks with Mobile Edge Computing

Ke Zhang, Supeng Leng, Yejun He, Sabita Maharjan, and Yan Zhang

Abstract

Along with modern wireless networks being content-centric, the demand for rich multimedia services has been growing at a tremendous pace, which brings significant challenges to mobile networks in terms of the need for massive content delivery. Edge caching has emerged as a promising approach to alleviate the heavy burden on data transmission through caching and forwarding contents at the edge of networks. However, existing studies always treat storage and computing resources separately, and neglect the mobility characteristic of both the content caching nodes and end users. Driven by these issues, in this work, we propose a new cooperative edge caching architecture for 5G networks, where mobile edge computing resources are utilized for enhancing edge caching capability. In the architecture, we focus on mobility-aware hierarchical caching, where smart vehicles are taken as collaborative caching agents for sharing content cache tasks with base stations. To further utilize the caching resource of smart vehicles, we introduce a new vehicular caching cloud concept, and propose a vehicle-aided edge caching scheme, where the caching and computing resources at the wireless network edge are jointly scheduled. Numerical results indicate that the proposed scheme minimizes content access latency and improves caching resource utilization.

Introduction

In recent years, wireless communication has been witnessing explosive growth of smart devices and mobile users. Along with the development of mobile networks, the advancements in wireless technologies as well as the Internet of Things (IoT) bring us a wide variety of powerful mobile applications and multimedia services. These attractive applications and services heavily rely on high-speed data rates and low-latency transmission, which pose significant challenges to mobile networks.

To meet these unprecedented traffic demands and challenges, lots of efforts have been made. In fifth generation (5G) networks, the shrinking of cell sizes and dense deployment of wireless access points open up new opportunities for faster data delivery. However, the centralized nature of mobile network architectures as well as the limited transmission capacity entailed by the wireless backhaul links make this approach unable to keep pace with the explosively growing traffic [1]. Thus, a new paradigm that breaks the bottleneck of massive content delivery is required.

In this regard, the context-aware 5G network with edge caching has emerged as a promising approach, which has the potential to yield significant gains for both mobile users and operators. In edge caching networks, by utilizing the storage resources at the edge nodes, popular contents can be cached close to end users. Some key features of edge caching have been studied in the context of 5G mobile Internet. As effective cooperation of edge caching nodes needs to be strategically managed, a hybrid architecture that harnesses the benefits of edge caching and cloud access networks was proposed in [2]. To further utilize cloud-based radio access networks and provide a more flexible caching service in 5G mobile systems, a caching scheme with a virtualization-evolved packet core concept was introduced in [3], where third-party service providers can be adaptively empowered. As physical layer technologies play a vital role in the content delivery process, it is crucial that the physical layer also be considered while designing the caching control mechanism. Such an approach possesses great potential to enhance the spectral efficiency in caching-enabled 5G networks [4]. For mobile users, their social interest greatly influences traffic fluctuation and content distribution of base stations (BSs) [5]. Motivated by such considerations, the social characteristics of end users were exploited to support content sharing between mobile devices and cache deployment of BSs in [6, 7]. Moreover, several economic approaches have been developed to incentivize caching service providers to distribute popular contents in an efficient manner (e.g., [8]).

With software-defined networking (SDN) and network functions virtualization (NFV) in place, communication and computing functionalities are converging in 5G networks [9]. Following these developments, jointly optimizing caching and computing capabilities in mobile networks may provide higher efficiency for users' applications with extensive computation demands and continuous content delivery. However, the caching capacity improvement brought by computing resource consumption is always neglected. In augmented reality applications, it is common practice to withdraw some key features from the

Digital Object Identifier:
10.1109/MWC.2018.1700303

Ke Zhang and Supeng Leng are with the University of Electronic Science and Technology of China; Yejun He is with Shenzhen University; Sabita Maharjan is with Simula Metropolitan Center for Digital Engineering; Yan Zhang (corresponding author) is with University of Oslo.

1536-1284/18/$25.00 © 2018 IEEE

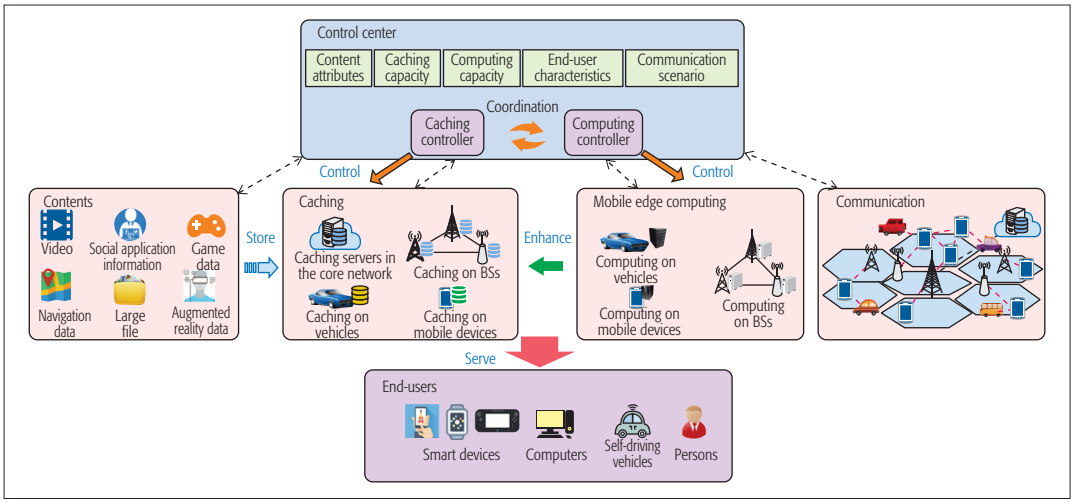

FIGURE 1. The proposed cooperative edge caching architecture.

originally captured videos in order to save caching and transmission resources [10]. It is therefore crucial to exploit computing resource to alleviate the strain on caching resources for the nodes with poor storage resources.

Although edge caching offers contents to end users in proximity, with the ever growing numbers of portable and handheld equipments, the unpredictable user mobility may heavily affect the caching strategies and complicate the content delivery process. Mobility-aware caching, however, has not witnessed much work yet. In dense BS deployment 5G networks, during a request for a content of large size, such as a video file, a user with high mobility may pass several small cells. Thus, the contents should be optimally cached at the edge nodes along the user's path such that it can be fetched by the user when he/she requires it [11].

Along with recent advances in wireless communication and IoT, vehicular networks have become an important 5G application. Enabled by LTE-V or IEEE 802.11p technologies, a vehicle can communicate with infrastructures, pedestrians, and other vehicles. Together with their computing and storage capability, communication-enabled smart vehicles are able to act as moving caching nodes, bringing contents to end users in wide areas.

In this article, we present a new cooperative content caching framework, and propose a hierarchical mobility-aware edge caching scheme that harnesses the synergies between mobile edge computing (MEC), multi-BS caching, and vehicular caching. The main contributions of this article are summarized as follows.

- We propose a new cooperative edge caching architecture. By leveraging MEC for content caching enhancement, this architecture brings an efficient and flexible content service, while also improving storage resource utilization.
- We design mobility-aware edge caching strategies that store popular contents in the BSs passed by mobile end users, consequently minimizing content access delay.
- To further improve the edge caching performance, we exploit content caching and delivery capabilities of moving vehicles. To this end, we present a mobility-aware cooperative edge caching scheme that jointly optimizes caching and computing resources of BSs and smart vehicles.

The remainder of this article is organized as follows. We introduce our proposed cooperative edge caching architecture in the following section, along with the main characteristics and key technologies of this architecture. Next, we focus on mobility-aware cooperative content caching and develop a vehicle-aided hierarchical edge caching scheme. Then we present numerical results for performance evaluation. Conclusions and future research challenges are presented in the final section.

Cooperative Caching Networks

In this section, we present a cooperative edge caching network that leverages MEC resources to improve caching capability. This network provides flexible resource utilization while facilitating content caching and delivery.

Cooperative Caching Architecture

Figure 1 shows the architecture of the proposed caching network. In this architecture, there are two types of resources. One is caching resources, and the other one is MEC resources. They serve end users under the management of the controllers.

Caching resources are composed of various types of content-storage-enabled entities. The first one is the caching servers located in the core network. Although these servers are far away from the end users, the servers still play a vital role in the content serving process. As edge nodes always have limited caching capabilities, they cannot cache all the contents themselves at the same time. Furthermore, the popularity of the contents is time-varying. Thus, the contents cached on the edge nodes should be updated adaptively.

In the Internet, contents may be generated from a large amount of providers. If all the newly updated contents are obtained directly from the providers to the edge nodes, high end-to-end latency may be caused by complex interactions between edge nodes and providers, and bandwidth limitation at the content providers. Caching servers can help address this issue by obtaining and caching the new contents according to the

Along with recent advances in wireless communication and IoT, vehicular networks have become an important 5G application. Enabled by LTE-V or IEEE 802.11p technologies, a vehicle can communicate with infrastructures, pedestrians and other vehicles. Together with their computing and storage capability, communication-enabled smart vehicles are able to act as moving caching nodes, bringing contents to end-users in wide areas.

The mobility characteristics of both the computing-enabled vehicles and handhold devices make them a mobile computing resource pool with variable capacity. Accurate prediction of vehicular traffic or flow of people may be greatly beneficial toward improving resource utilization and caching performance.

content popularity. Being intermediate content caching and forwarding devices in the core network, these servers can be accessed and utilized easily by the edge nodes. Furthermore, the high bandwidth of the core network is helpful to form the cooperation of the caching servers for sharing their cached contents, which can further improve the caching efficiency.

The second part of the caching resources is the cache-enabled BSs. Delivering contents directly to end users, BSs are considered as effective nodes to cache popular contents and reduce the duplicate content transmissions from the core network. Heterogeneous networks consisting of multiple types of BSs will constitute a major part of 5G architecture. As various types of BSs have different coverage areas and serve different numbers of users, the content caching strategies for each type of BS need to be carefully designed. For instance, compared to microcell BSs, a macrocell BS covers a wider area with more end users. To provide better caching service, the contents that meet the main requirements of the covered users should be stored on the caching of macrocell BSs. On the contrary, the microcell BSs need to follow and cache the particular content demands from the covered local area.

Furthermore, user mobility patterns are also required to be considered in the BS caching process. During the movement of users, several cells may be passed by. In the case where the size of the content is large, (e.g., video streaming and file sharing), the content caching and delivery tasks may be shared by several BSs. As the characteristics of the content as well as the moving speed and directions of the users may affect the caching process, how to effectively arrange the content segments to the cache of the BSs along the way forward, and ensure the content accessed by the users as needed, is a challenge.

Besides caching servers and BSs, there are cache-enabled vehicles and mobile devices in caching resources, which can be categorized as mobile caching nodes. Due to the development of IoT, smart vehicles and devices are growing ubiquitously, and have been empowered with caching as well as computing and communication capabilities. Although the caching resource of one vehicle or mobile device is limited, the accumulative caching power gathered from a group of these mobile nodes is sufficient for storing contents. Hence, caching on mobile nodes is an important method of efficient content distribution. The characteristics of group aggregation and highly dynamic topology of the mobile nodes pose significant challenges on edge caching. The formed cluster of vehicles and mobile devices may be separated due to different directions or different moving speeds of these mobile nodes. Thus, the formation and active duration of a mobile node group play an important part in mobile caching utilization. Furthermore, the caching capability together with the communication capacity of various types of mobile nodes are also different. How to efficiently arrange them to cater to the content requirements of the users needs to be carefully investigated.

MEC, a key technology toward 5G, provides cloud computing capabilities and task offloading service at the edge of mobile networks [12]. Due to the proximity of MEC servers to end users, tasks can be offloaded and accomplished with low latency and high efficiency.

Similar to the composition of caching resources, in the proposed architecture, MEC resources consist of heterogeneous BSs, smart vehicles, and mobile devices equipped with computation capabilities. Although MEC resources seem different from caching resources, they are closely coupled. For instance, using MEC resources on file compressing, the size of a file may be reduced. Thus, some storage space can be saved. From another perspective, the caching capability of nodes is enhanced. Another example is augmented reality, where the key elements of the captured video can be extracted from the original data through information processing and computing. As the size of key elements is small, they can be cached and distributed easily. Based on the received key elements, end users may reconstruct the original image or video.

Although individual computing and caching resources of the edge nodes can be collaboratively used to improve the content caching performance, the cooperative approach also brings in several issues. The first one comes into effect because of the variability in computing and caching capabilities of different nodes, which make the joint resource scheduling a complicated task. Machine learning is a feasible approach to address the problem. For instance, a reinforcement learning algorithm can be adopted to adaptively arrange serving capabilities to satisfy the content requirements based on various factors and long-term outcome evaluation [13]. Another problem is related to content compression. Some contents can be significantly compressed by means of specific processing, for which computation resources are needed. To this end, the trade-off between computing cost and caching gains should be taken into account in the caching process. Furthermore, the mobility characteristics of both the computing-enabled vehicles and handheld devices make them a mobile computing resource pool with variable capacity. Accurate prediction of vehicular traffic or flow of people may be greatly beneficial toward improving resource utilization and caching performance.

In the proposed architecture, the control center manages the edge caching process in a logically centralized way. The main components in this center are caching and computing controllers, which manage various caching resources and MEC resources, respectively. In their control process, both controllers monitor corresponding resource utilization state, and update the caching capacity and MEC capacity records, which are key to the resource management problem in this context.

User applications or content can vary widely in terms of popularity, content size, caching, or computing demands. To accommodate their requirements and deliver the contents efficiently, the attributes of the contents should be learned and be incorporated into the control strategy design. As for end users, who are the content requesters, they have multiple factors that may affect content caching, including content preference, mobility pattern, application quality of service (QoS), and so on. These factors also need to be analyzed in

the control center. As the efficiency of content delivery also depends on data transmission, it is crucial that the control center has an overview of states of the communication scenarios, such as channel states and wireless interference. With a global view of the states and requirements of the network, caching and computing controllers cooperate with each other to efficiently schedule the content caching process.

Benefits for Content-Centric 5G Networks

The proposed cooperative edge caching architecture leverages MEC resources to alleviate the content caching burden on edge nodes with strained storage capability. Furthermore, in this architecture, potential advantages of heterogeneous cache-enabled IoT are fully explored and exploited for providing collaborative content caching and forwarding services. With centralized control of both computing and the caching process, a flexible and efficient caching mechanism can be achieved. The benefits of this architecture for content-centric 5G networks are summarized as follows.

Heterogeneous Capability Integration: Caching and computing capabilities, which are heterogeneous but ubiquitous at the edge nodes, are integrated in the content caching process. The MEC capability of the nodes acts as an enabler to improve caching capability through content information processing. In this regard, MEC can also contribute toward enhancing caching.

Efficient Resource Utilization: The storage resources of heterogeneous integration of a large number of things are fully utilized for content caching. In particular, the resources of moving smart vehicles as well as of handheld devices are exploited to cache and spread popular contents to a wide area. Hence, resource utilization can be improved.

Flexible Cache Scheduling: This architecture abstracts the caching and computing capabilities of the network, and manages content caching in a centralized manner. Various types of resources from heterogeneous edge nodes can be adaptively arranged in independent or cooperation modes according to the characteristics of the content, the requirements of end users, and the features of communication scenarios.

Mobility-Aware Cooperative Edge Caching Schemes

In this article, we propose mobility-aware cooperative content edge caching schemes, which can improve end-user experience by reducing their content acquisition time. In particular, we explore both storage and computing capabilities of caching nodes, and present an approach to raise their storage capabilities by utilizing the computing resources. Moreover, mobile characteristics of the vehicles and end users are considered in the caching scheme design. We propose mobility-aware edge caching strategies that optimally cache content at the heterogenous edge nodes along the end users' way. In this section, we study two edge caching scenarios: caching with and without vehicles. The system models as well as optimal caching problem formulation and solution approaches are described next.

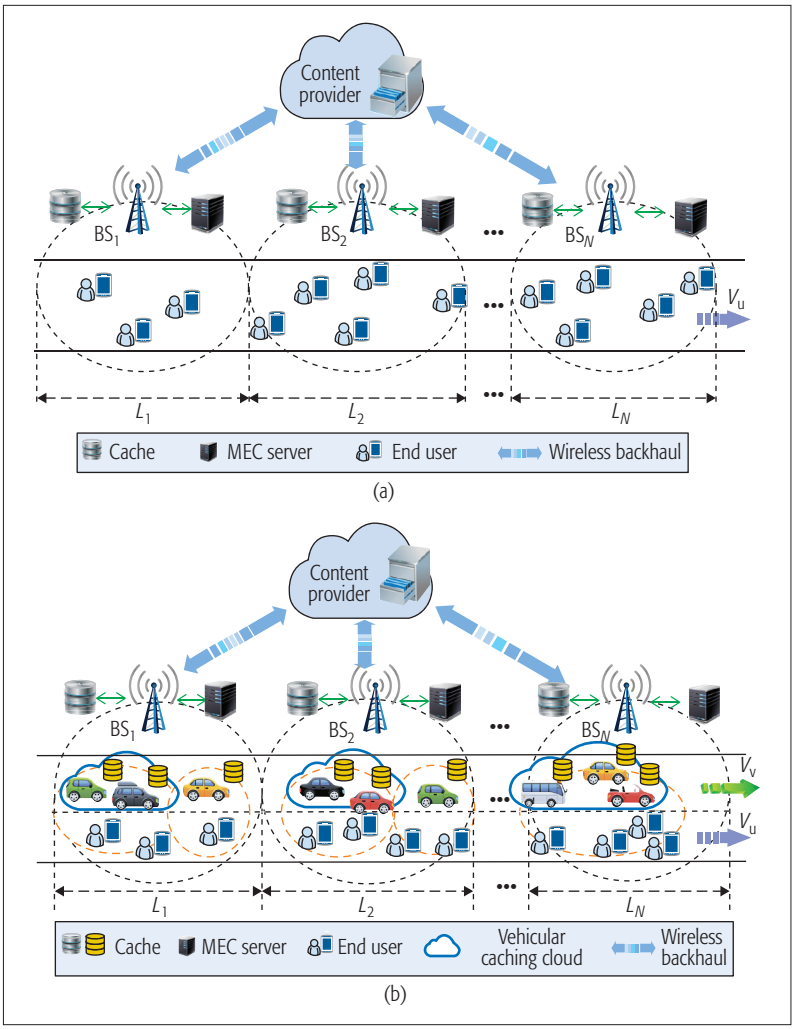

FIGURE 2. Two edge caching scenarios: a) cooperative edge caching without vehicles; b) vehicular cloud-aided cooperative edge caching.

Cooperative Edge Caching without Vehicles

The specification of the proposed cooperative caching model without vehicles is shown in Fig. 2a. Let $\mathcal{N}$ denote the set of N BSs, which are located along a unidirectional road. Each BS is equipped with an MEC server. The amount of cache resource in one BS and the computing resource of an MEC server are f_b and c_b, respectively. The end users that hold the mobile devices requesting contents are moving along the road at speed V_u. During the movement, the end users may pass through several wireless coverage areas of BSs. Let the length set of road sections covered by these BSs be $\{L_1, L_2, ..., L_N\}$.

To study the effects of the content characteristics on the design of caching schemes, the contents requested by the end users are classified into S types. Each content is identified with three components, and is presented as $f_i = \{d_i, \rho_i, e_i\}$, $i \in \mathcal{S}$, where d_i and ρ_i are the size of content i and the popularity of content i, respectively. e_i is the compressibility of the content with a unit computing resource.

Each end user randomly chooses one type of the contents to download when they arrive at the starting point of the road. As users move along the road, they may get part of the content from

Symbol	Description
d_i, ρ_i, e_i	Size, popularity, and compressibility of content i
$x_{i,j}$	Caching resources on BS j allocated to cache content i
$y_{i,j}$	Computing resources of MEC server j used to process content i
t_c, t_b	Time for getting a unit content from content provider and from the cache at a BS
t_e	Processing latency of compressing contents

TABLE 1. Main notations.

one BS and other parts from the upcoming ones. To provide continuous content delivery service, the contents should be located in the caches of the BSs efficiently. Let $x_{i,j}$ denote the amount of the caching resources on BS j allocated to cache content i, where $i \in \mathcal{S}$ and $j \in \mathcal{N}$. Considering the storage limitations of the BSs, to improve the content caching capacities, MEC servers can be utilized to reduce the size of the content files. The computing resources of MEC server j used to process content i is $y_{i,j}$.

In order to minimize the average latency of the contents downloading process, the problem of optimal cooperative content caching can be formulated as

$$\min_{\{x_{i,j},y_{i,j}\}} \sum_{i=1}^{S} \rho i \sum_{j=1}^{J_i^{\max}} \{(L_j / V_u t_b - x_{i,j}) t_c + y_{i,j} t_e + x_{i,j} t_b\} / S$$

$$\text{s.t.} \quad \text{C1}: \sum_{i=1}^{S} x_{i,j} / (1 + e_i y_{i,j}) \le f_b, \quad j \in \mathcal{N}$$

$$\text{C2}: \sum_{i=1}^{S} y_{i,j} \le c_b, \quad j \in \mathcal{N} \qquad (1)$$

$$\text{C3}: \quad x_{i,j} \le L_j / V_u t_b, \quad i \in \mathcal{S}, j \in \mathcal{N}$$

where t_c and t_b are the time spent by the users to get a unit content from the content provider and the cache at a BS, respectively. Due to the long transmission distance between the content provider in the core network and the end users, $t_c > t_b$. t_e is the processing latency of compressing the contents with a unit computing resource. $J_i^{\max}$ is the index of the farthest BS where type i content is cached in. $J_i^{\max}$ can be defined as

$$\sum_{j=1}^{J_i^{\max}-1} L_j / V_u t_b < d_i \le \sum_{j=1}^{J_i^{\max}} L_j / V_u t_b.$$

In Eq. 1, constraints C1 and C2 ensure the allocated caching resource and computing resource within the storage capacity of each BS and computing capacity of each MEC server, respectively. In addition, constraint C3 states that the size of the content cached in BS j should not exceed the maximum amount required by the end users.

To solve Eq. 1, we use a game theoretic approach to achieve the optimal cooperative caching and computing strategies. In this game, the players are the N types of contents. The strategy set of content i is $\{x_{i,j}, y_{i,j}\}$, which is nonempty, convex, and compact. Choosing a caching and computing joint strategy, the utility of each player is the waiting time to receive the content. Clearly, the utility function of content i is continuous and quasi-concave in terms of $x_{i,j}$ and $y_{i,j}$. A Nash equilibrium (NE) of the game is a solution, in which no player can further reduce its waiting time by changing the strategy unilaterally, given the joint strategies of the other players. According to the Nash existence theorem, this game possesses at least one pure strategy NE [14]. Thus, we can obtain an NE, which is the solution of Eq. 1, in a heuristic manner, where each type of content iteratively updates its joint caching strategy based on the strategies of the other content types.

Cooperative Edge Caching Aided by Vehicular Caching Cloud

Recent advances in IoT and intelligent vehicle technologies have greatly increased the information processing capability of vehicles and have helped them to provide new applications. In particular, cache-enabled vehicles can be considered as a new approach to store and spread data. However, the characteristics of high mobility of vehicles and dynamically changing topology of vehicular networks pose significant challenges on the design of efficient vehicular content caching schemes. Therefore, it is imperative to design a cooperative vehicle-aided content edge caching scheme to minimize the content delivery latency for mobile end users while alleviating the caching pressure of the BSs in 5G networks.

Our proposed vehicle-aided hierarchical caching scheme is shown in Fig. 2b. In our model, the cache-enabled vehicles arrive at the road following a Poisson process. Let λ be the traffic density in terms of vehicles per unit distance. We consider each vehicle having a homogeneous caching resource. Let f_v be the maximum amount of data that can be stored in the cache of each vehicle. As the speed of the vehicles and that of the users are different, during the movement of a user on road section L_j, the average number of vehicles passing by the user is $Q = \lfloor \lambda L_j (V_v - V_u)/V_u^2 \rfloor$. Then the average amount of data that is delivered by a vehicle to an end user during the passing period will be $w = \lambda l/(V_v - V_u)t_v$. Here, l and t_v are the length of the road section covered by the vehicle's wireless signal and the time cost for transmitting a data unit from the vehicle to the end user, respectively. We assume that the time spent by the user to get a unit data from the vehicle is longer than getting it from the BSs but shorter than getting it from the content provider (i.e., $t_b < t_v < t_c$). Considering the caching capacity of the vehicles, the active information service capability for one vehicle to an end user can be denoted as $q = \min\{w, f_v\}$. To fully exploit the caching capability of the vehicles and make collaboration between the BSs and the vehicles efficient, we propose a vehicular cloud-aided caching scheme. The cloud is formed with several cache-enabled vehicles, where the contents are well segmented and stored in these vehicular caches. These cached contents are delivered directly from the running vehicles to the end users. In this way, the content receiving latency of the end users can be greatly reduced, especially for BSs with poor storage capacity. We design a heuristic vehicular caching cloud formation algorithm for content processing and storing. The main steps of the algorithm are described next.

Step 1: Based on the solution of Eq. 1, for each road section, if there are contents that need to be downloaded from the content provider in the core network, these contents are first divided

into blocks with the same size q. Then the content blocks may be stored in the cache of the vehicles. One vehicle caches one block.

Step 2: Searching for each road section, for section j, $j \in \mathcal{N}$, if content type i needs processing, compare the time cost of the processing process with that of the data transmission from vehicles to end users. If $t_v < y_{i,j}t_e + t_b/(1 + e_i y_{i,j})$, divide the content into blocks and cache it into vehicles.

Step 3: Caculate the total number of content blocks Z. In the traffic, choose $\{Z, Q\}$ consecutive arriving vehicles to form the cloud, which store and deliver the content blocks to the end users while passing through the road.

Illustrative Results

In this section, we show illustrative results to demonstrate the performance of our proposed cloud-aided caching scheme. We consider five BSs located along a unidirectional road. The caching capacity f_b and MEC capacity c_b of each BS are set as 1 GB and 50 units, respectively [15]. The end users taking on normal buses move at the speed V_u = 80 km/h, while the smart vehicles run at V_v = 120 km/h. In this network, the contents required by the end users have large size, which is randomly taken from [500, 1000] MB.

Figure 3 shows the comparison of average content downloading latency of different caching schemes. Our proposed cloud-aided caching scheme has the shortest latency compared to the other two schemes in the scenarios with various numbers of content types. In the cloud-aided caching scheme, both the MEC resources on BSs and the caching capacity of smart vehicles are fully utilized. Compared to the other two schemes, more contents are stored in the edge nodes near the end users. Thus, less contents need to be transmitted from the core network, and downloading time is saved. It is noteworthy that when the number of content types is less than 10, the latency reduction of the vehicular cloud-aided caching scheme is not significant. As a small number of types means few content caching requirements, the caching capacity of BSs may meet almost all of the caching demands. Therefore, the caching performance improvement of our cloud-caching scheme is not pronounced. However, when the number of types is large, our cloud-aided scheme yields a significant reduction in content receiving latency.

In addition, we can see from Fig. 3 that through utilizing computing resources to enhance the caching capacity of BSs, the scheme with MEC servers but without smart vehicle-aided caching outperforms the one with only caching resources in the BSs. The difference between these two schemes is significant, when the number of content types are 11 and 12. As the number increases, the difference becomes smaller. The reason is that the computing resources of an individual MEC server is limited. With the continued increase of the content types, the computing capacity of MEC servers is exhausted. However, in our vehicular cloud-aided scheme, besides the MEC servers, the storage resources provided by smart vehicles are also utilized, which greatly improves the caching capacity of the network. Thus, our proposed scheme is more effective in downloading latency reduction.

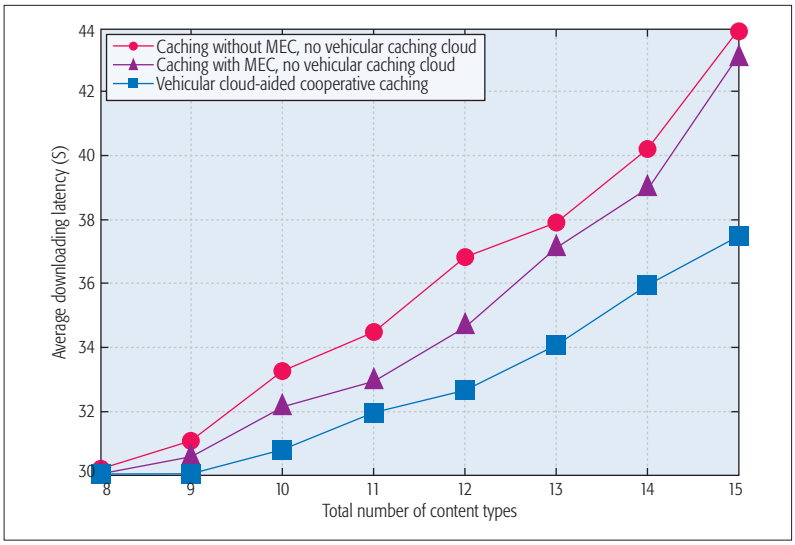

FIGURE 3. Comparison of content downloading latency of different caching schemes.

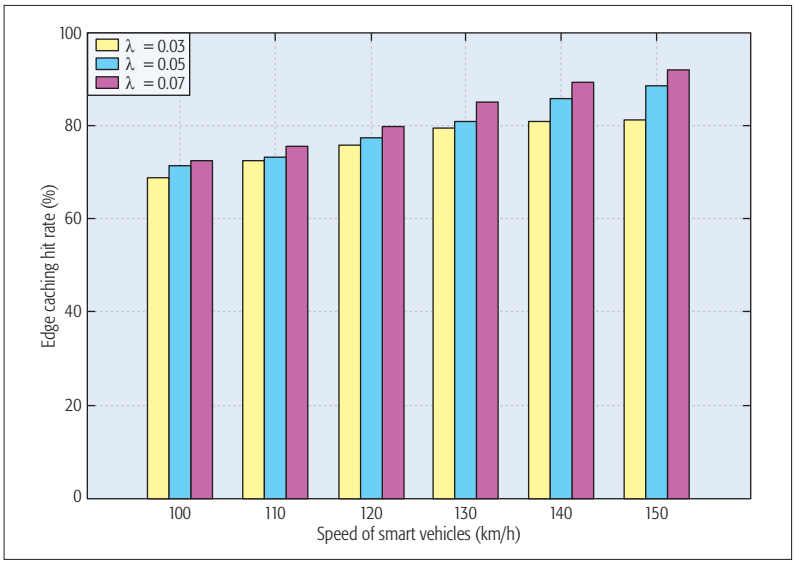

FIGURE 4. Comparison of edge caching hit rate with different vehicle speeds.

Figure 4 compares the performance of edge caching hit rate using the cloud-aided caching scheme, which indicates the percentage of the content that can be accessed directly from the edge nodes, including BSs and smart vehicles. It is clear that with the increase of vehicles running on the road, more vehicular caching resources can be utilized, and higher hit rates are obtained. This proves the caching effectiveness of our proposed vehicular cloud-aided caching scheme. We note that the difference between the rates of the three traffic densities at low vehicle speed is much less than that at high speed. For each smart vehicle, when it runs slowly, only a few end users can be passed by during its movement along the road. Since the vehicular caching capacity is limited, the contents cached on the slow moving vehicles can only be shared among a small amount of users. In other words, many duplicated contents may be stored on these vehicles serving for small and separate user groups. Hence, in the high traffic density case, although there are plenty of smart vehicles running on the road, the vehicular cach-

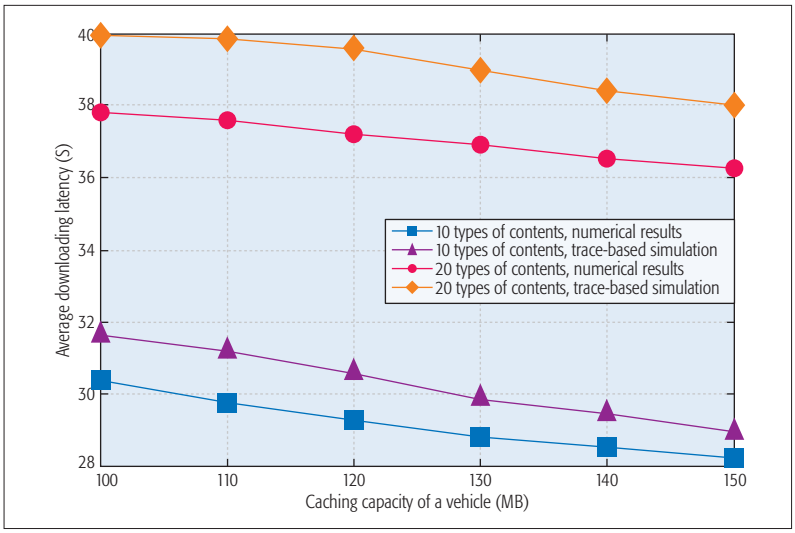

FIGURE 5. Performance comparison between numerical results and trace-based simulation results.

ing capacity is not fully utilized. However, as the speed increases, the contents cached on a vehicle can be directly obtained by more users. Thus, different contents are separately cached in several vehicles, and an efficient vehicular caching cloud can be formed and utilized.

Figure 5 shows the performance comparison between numerical results and trace-based simulation results applying our proposed vehicular cloud-aided edge caching scheme. It can be found that the average latency of trace-based simulation results is higher than that of numerical results. The average differences are 1.105 s and 2.073 s in the two cases with 10 and 20 types of contents, respectively. In our scheme, we arrange content caching according to the average vehicle speed on a road. However, in reality, both the traffic volume and vehicle speed are time-varying. These factors may dynamically change the size of the formulated vehicular cloud as well as its caching capacity. When some vehicles move out of a cloud, the contents stored on these vehicles are not available for direct vehicle-to-user transmission. Consequently, content download requirements from end users to providers are incurred, and the average latency increases. From Fig. 5, we can also see that the differences between the two results in both cases become smaller as vehicular caching capacity increases. Higher vehicular caching capacity means that more contents can be stored in the vehicles, and more vehicular clouds can be formulated on the road. When some vehicles move out of a cloud or leave the road, users may get contents from other vehicles of the cloud or from other vehicular clouds. Thus, the average content download latency is reduced.

Conclusion and Open Issues

In this article, we present a cooperative edge caching architecture for content-centric 5G networks. By leveraging MEC resources, the storage capability of the nodes is enhanced; consequently, a mobile network with flexible and efficient content caching and delivery is obtained. We further investigate a mobility-aware caching framework for mobile end users, where cache-enabled vehicles are utilized to share the content caching tasks with BSs. We propose an efficient vehicular cloud-aided edge caching scheme. Numerical results indicate that our proposed scheme greatly reduces the content downloading latency and improves caching resource utilization.

Edge caching is an important way to improve the content distribution efficiency in 5G networks. However, how to effectively utilize caching, computing, and communication resources of edge nodes in content storage and delivery is still a fundamental but unexplored question. In addition, the popularity of contents is time-varying. While the vehicles are running along the road, the decision of what to cache on the vehicles and when to update the cached contents are crucial for content delivery performance. Due to the dynamic mobility pattern of end users as well as time varying traffic volume of vehicles, the coming challenge is how to design a smart and adaptive caching mechanism. Furthermore, the way to incentivize a large amount of heterogeneous caching nodes to follow effective caching strategies while ensuring security of contents and networks also requires future study.

Acknowledgment

Work in this article was supported by the National Natural Science Foundation of China under Grant 61374189; the joint fund of the Ministry of Education of China and China Mobile under Grant MCM 20160304; the Shenzhen Science and Technology Programs under Grants JCYJ20170302150411789, JCYJ20170302142515949, and GCZX2017040715180580; the Guangzhou Science and Technology Program under Grant 201707010490; and the Fundamental Research Funds for the Central Universities, China, under Grant ZYGX2016J001.

References

[1] X. Wang *et al.*, "Cache in the Air: Exploiting Content Caching and Delivery Techniques for 5G Systems," *IEEE Commun. Mag.*, vol. 52, no. 2, Feb. 2014, pp. 131–39.
[2] R. Tandon and O. Simeone, "Harnessing Cloud and Edge Synergies: Toward an Information Theory of Fog Radio Access Networks," *IEEE Commun. Mag.*, vol. 54, no. 8, Aug. 2016, pp. 44–50.
[3] X. Li *et al.*, "CaaS: Caching as a Service for 5G Networks," *IEEE Access*, vol. 5, Mar. 2017, pp. 5982–93.
[4] W. Han, A. Liu, and V. K. N. Lau, "PHY-Caching in 5G Wireless Networks: Design and Analysis," *IEEE Commun. Mag.*, vol. 54, no. 8, Aug. 2016, pp. 30–36.
[5] X. Wang *et al.*, "Tag-Assisted Social-Aware Opportunistic Device-to-Device Sharing for Traffic Offloading in Mobile Social Networks," *IEEE Wireless Commun.*, vol. 23, no. 4, Aug. 2016, pp. 60–67.
[6] L. Jiang *et al.*, " Social-Aware Energy Harvesting Device-to-Device Communications in 5G Networks," *IEEE Wireless Commun.*, vol. 23, no. 4, Aug. 2016, pp. 20–27.
[7] X. Zhang *et al.*, "Information Caching Strategy for Cyber Social Computing Based Wireless Networks," accepted for publication, *IEEE Trans. Emerging Topics in Computing*.
[8] M. Hajimirsadeghi, N. B. Mandayam, and Alex Reznik, "Joint Caching and Pricing Strategies for Popular Content in Information Centric Networks," *IEEE JSAC*, vol. 35, no. 3, Mar. 2017, pp. 654–67.
[9] S. Andreev *et al.*, "Exploring Synergy between Communications, Caching, and Computing in 5G-Grade Deployments," *IEEE Commun. Mag.*, vol. 54, no. 8, Aug. 2016, pp. 60–69.
[10] X. Wang *et al.*, "Cloud-Assisted Adaptive Video Streaming and Social-Aware Video Prefetching for Mobile Users," *IEEE Wireless Commun.*, vol. 20, no. 3, June, 2013, pp. 72–79.
[11] A. Mahmood *et al.*, "Mobility-Aware Edge Caching for Connected Cars," *Proc. Wireless On-demand Network Systems and Services*, Jan. 2016.
[12] P. Mach and Z. Becvar, "Mobile Edge Computing: A Survey on Architecture and Computation Offloading," *IEEE Commun. Surveys & Tutorials*, vol. 19, no. 3, Mar. 2017, pp. 1628–56.

[13] C. Wu *et al.*, "A Reinforcement Learning-Based Data Storage Scheme for Vehicular Ad Hoc Networks," *IEEE Trans. Vehic. Tech.*, vol. 66, no. 7, July 2017, pp. 6336–48.
[14] G. Debreu, "A Social Equilibrium Existence Theorem," *Proc. National Academy of Sciences of the United States of America*, vol. 38, no. 10, Oct. 1952, pp. 886–93.
[15] W. Jiang, G. Feng and S. Qin, "Optimal Cooperative Content Caching and Delivery Policy for Heterogeneous Cellular Networks," *IEEE Trans. Mobile Computing*, vol. 16, no. 5, May 2017, pp. 1382–93.

Biographies

Ke Zhang (zhangke@uestc.edu.cn) received his Ph.D. degree from the University of Electronic Science and Technology of China (UESTC) in 2017. He is currently a lecturer in the School of Information and Communication Engineering, UESTC. His research interests include scheduling of mobile edge computing, design and optimization of next-generation wireless networks, and the Internet of Things.

Supeng Leng [M'06] (spleng@uestc.edu.cn) is a professor with the School of Information and Communication Engineering, UESTC. He has published more than 100 research papers. His research interests include resources, energy, and networking in broadband wireless access networks. He serves as an Organizing Committee Chair and a Technical Program Committee member for many international conferences as well as a reviewer for more than 10 international research journals.

Yejun He [SM'09] (heyejun@126.com) received his Ph.D. degree from Huazhong University of Science and Technology, China. Since 2011, he has been a full professor with the College of Information Engineering, Shenzhen University. He has authored or coauthored more than 100 research papers and books or book chapters and holds 13 patents. He currently serves as an Associate Editor of leading journals such as *IEEE Network*, *IEEE Access*, and the *International Journal of Communication Systems*. He is a Fellow of IET and the IEEE Antennas and Propagation Shenzhen Chapter Chair.

Sabita Maharjan [M'09] (sabita@simula.no) received her Ph.D. degree in networks and distributed systems from the University of Oslo, and Simula Research Laboratory, Norway, in 2013. She is currently a senior research scientist at Simula Metropolitan Center for Digital Engineering, Norway, and an associate professor (adjunct position) at the University of Oslo. Her current research interests include wireless networks, network security and resilience, smart grid communications, cyber-physical systems, machine-to-machine communications, and software defined wireless networking.

Yan Zhang [SM'10] (yanzhang@ieee.org) is a full professor with the University of Oslo. He is an Editor of several IEEE publications, including *IEEE Communications Magazine*, *IEEE Network*, *IEEE Transactions on Green Communications and Networking*, *IEEE Communications Surveys & Tutorials*, and *IEEE Internet of Things*. His current research interests include next-generation wireless networks leading to 5G and cyberphysical systems. He is an IEEE VTS Distinguished Lecturer and a Fellow of IET.