A Dynamic Edge Caching Framework for Mobile 5G Networks

Open Call

Dinh Thai Hoang[1], Dusit Niyato[2], Diep Nguyen[1], Eryk Dutkiewicz[1],
Ping Wang[2], and Zhu Han[3]
[1] School of Electrical and Data Engineering, University of Technology Sydney, Australia
[2] School of Computer Engineering, Nanyang Technological University, Singapore
[3] University of Houston and Kyung Hee University.

Abstract

Mobile edge caching has emerged as a new paradigm to provide computing, networking resources, and storage for a variety of mobile applications. That helps achieve low latency, high reliability, and improve efficiency in handling a very large number of smart devices and emerging services (e.g., IoT, industry automation, virtual reality) in mobile 5G networks. Nonetheless, the development of mobile edge caching is challenged by the decentralized nature of edge nodes, their small coverage, limited computing, and storage resources. In this article, we first give an overview of mobile edge caching in 5G networks. After that, its key challenges and current approaches are discussed. We then propose a novel caching framework. Our framework allows an edge node to authorize the legitimate users and dynamically predict and update their content demands using the matrix factorization technique. Based on the prediction, the edge node can adopt advanced optimization methods to determine optimal contents to store so as to maximize its revenue and minimize the average delay of its mobile users. Through numerical results, we demonstrate that our proposed framework provides not only an effective caching approach, but also an efficient economic solution for the mobile service provider.

Index Terms

Edge caching, 5G networks, matrix factorization, machine learning, blockchain, recommendation system, and mixed integer linear programming.

Contact: **D. T. Hoang**, School of Electrical and Data Engineering, University of Technology Sydney, Building 11, Level 8, Broadway, Sydney, NSW 2007, Australia, Telephone: 61-29514-2402, E-mail: HOANG.DINH@uts.edu.au. Editor: Gilberto Berardinelli.

I. Introduction

Mobile Edge Caching (MEC) is an emerging architecture concept that provides storage and computing capabilities at the "edge" of 5G networks. The basic idea of the MEC network is to "move" the edge servers closer to the mobile users, thereby reducing the service delay/latency for customers while mitigating the network congestion. With over 5 million cell towers currently on earth, it is anticipated that MEC can save mobile network operators up to 35% on backhaul usage, and the latency in mobile services can be reduced by 50% [1]. However, the development of MEC networks faces many challenges especially in managing cache resources. Due to special characteristics of the MEC network, e.g., small coverage and limited computing and storage resources, conventional cache management approaches may become ineffective. For example, typical cache management schemes rely on global network information about content demand, which is rarely available in the distributed environment of the MEC network. Thus, there is an urgent need for innovative methods to deal with this problem.

In this article, we first present an overview of MEC in 5G networks including system architecture, benefits, intrinsic features, and challenges. We then discuss smart caching approaches in the literature for MEC networks. After that we introduce a dynamic caching framework which allows an MEC node to authenticate the legitimate mobile users and predict its mobile users' content demands by using an advanced machine learning approach, called, non-negative matrix factorization (NMF). The NMF technique allows to predict the demand of users with a high accuracy through using the users' access history. This technique has a wide range of applications in practice, especially for recommendation systems such as Youtube and Netflix. Based on this prediction, a mixed integer linear programming algorithm is proposed to find an optimal caching policy which minimizes the average delay for the mobile users under a limited cache storage capacity of the MEC node. Alternatively, based on the predicted users' content demand and the achieved optimal caching policy, we deliver the optimal business strategy for the mobile service provider to maximize its revenue. Through numerical results, we show that our proposed framework is an effective and reliable solution for future MEC networks.

The rest of paper is as follows. In Section II, we give an overview of MEC in 5G mobile networks. The existing caching strategies used in MEC and their limitations are discussed in Section III. After that, our proposed optimal caching approaches are presented in Section IV. Finally, we highlight the challenges for the development of MEC networks, and conclude the paper in Section V.

II. Overview of Mobile Edge Caching in 5G Networks

A. Fundamental Background and Benefits of Mobile Edge Caching

According to the Cisco's latest forecast [3], a massive number of connected devices as well as demands for mobile video, and emerging services (in IoT, industry automation) of 5G networks will drive mobile data traffic to increase as much as 800% in the next five years. To cope with this challenge, *mobile edge caching* (MEC) has been introduced as one of the most effective solutions in future 5G mobile Internet. MEC aims to distribute popular contents closer to mobile users by deploying servers at the "edge" of the 5G mobile networks. As such, MEC is expected to provide the following benefits to mobile users, mobile network operators, and proximity service providers.

- *Benefits to mobile users:* Firstly, MEC servers provide a high-speed and reliable data transfer for mobile users through WiFi connections and Gigabit Ethernet. Secondly, caching popular contents at the edge can reduce service delay and bring better user experiences. Thirdly, with local high-speed connections, energy consumption on mobile devices can be significantly reduced.
- *Benefits to mobile network operators:* The deployment of MEC networks can reduce operation costs for mobile network operators up to 67% by reducing the total throughput and peak backhaul bandwidth consumption [4]. Alternatively, the MEC networks allow the mobile network operators to deploy MEC nodes flexibly, depending on demands of the mobile users in distinct areas at different times.
- *Benefits to proximity service providers:* The architecture of MEC networks can bring a great advantage for proximity services. This is thanks to the edge servers that are now placed closer to the end-users and device-to-device communication technology can be exploited. For example, fast charging services can be deployed at MEC nodes to wirelessly charge mobile devices.

B. Architecture of a Mobile Edge Caching Network

Fig. 1 illustrates a general architecture of an MEC network. An MEC node is composed of four main components, i.e., a wireless interface, an edge server, an edge storage, and an Internet backhaul connection. The access point usually uses WiFi connections to provide low energy consumption and a high-speed wireless connection to the mobile users. The edge storage is used to store popular contents according to users' demands which can be location-dependent. The edge server is used to handle users' requests, implement the optimal caching strategy, and perform pre-processing tasks for some specific contents, e.g., video transcoding and mobile social data analytics. Alternatively, the Internet connection is used to provide a high-speed connection to the public content servers and other MEC nodes.

Callout: Figure 1 A general architecture of a mobile edge caching network.

In an MEC network, MEC nodes are often deployed at locations with high network traffic demands to meet the users' requirements and offload for the backbone mobile network. When receiving a request to download and process a content from a mobile user, the MEC node first authenticates the user's information. Then, if the required content is stored in the edge storage, it will be transferred to the mobile user immediately. Otherwise, the MEC node will download the required content from the content server, and transfer the content to the user. The requesting users' identification and their downloaded contents may be stored at the MEC node to help the MEC node predict the demands of the mobile users in a certain location and make the optimal caching policy.

C. Development of Mobile Edge Caching Networks Worldwide and Commercial Products

For the benefits of MEC, many telecom companies worldwide have been developing their own MEC networks to reduce operational costs and improve the quality of service (QoS) for their customers. In particular, in the USA, AT&T has introduced a Content Delivery Network Service (CDNS) through deploying edge servers near to the customers, thereby reducing waiting time to access popular web pages. Verizon has selected Akamai, the world leader in CDNS, as a partner to provide CDNS to its customers. In Europe, Deutsche Telekom launched a New Edge Optimizer Service (EOS) fully optimized Webpages in 2014. This is the first implementation that executes PageSpeed logic directly on the edge server. Alternatively, many other telecom companies worldwide such as Vodacom Tanzania in Africa, China Telecom in Asia, and Optus in Australia, have developed their own MEC networks, making the MEC service market more active than ever.

Together with the development of the MEC networks worldwide, commercial MEC products including edge servers, platform, and software have received paramount attention recently. Specifically, Nuxeo has introduced a new MEC system, namely Nuxeo Edge Cache, to reduce the enterprise content download time up to 95% and bandwidth consumption up to 90% while keeping the digital assets safe at all times even when cached. Amazon has recently introduced Amazon Web Services (AWS) Snowball Edge, a 100TB data transfer device with on-board storage and computing capabilities. In addition, there are many other companies developing MEC products such as Edge Gateway 3000 Series by Dell, Edge Computing IoT Gateways AR500 series by Huawei, Edge Cloud Platform for mobile operators by Qwilt. Clearly, MEC has a promising market since it will be an indispensable part for the future 5G mobile Internet.

D. Key Features and Challenges of Edge Caching in mobile 5G Internet

In addition to common challenges of wireless caching networks, e.g., content popularity, dynamic demand, and prefetch, MEC networks possess some special features which may have significant influence to caching policies.

1) Small coverage: WiFi connections used at the MEC nodes have limited operation ranges, e.g., 30 meters for IEEE 802.11g. Hence, MEC nodes are often deployed at specific areas to serve hotspot demands of users. For example, an MEC node can be deployed at an enterprise to cache frequently used applications and software for its staff. An MEC node can be placed at a residential area to provide popular entertainment programmes for its residents. As a result, the optimal caching policy of an MEC node must account for the location-based demand of the mobile users.

2) Content diversity, pre-processing, and heterogeneous connections: Due to the composite architecture of MEC networks, edge caching policies in mobile 5G Internet are complicated due to great content diversity, long pre-processing time, and bandwidth allocation. For example, an MEC node must know the demand of users in its serving area, sizes of contents, pre-processing time, bandwidth allocated to the users, and Internet connection to the content server to find the optimal caching contents. Therefore, solutions for this problem need to be further investigated.

3) Limited computing and storage resources: To reduce the implementation and service costs, MEC nodes are usually equipped with much smaller computing and storage resources than those of the content servers. For example, a standard Redis caching package offered by Microsoft Azure costs approximately \$41 and \$103 monthly for 250MB and 1GB, respectively. Therefore, the MEC node needs to incorporate both its capacity and the users' demand to maximize its profit. Furthermore, due to a large number of users and a large amount of available contents in practice but a limited computing resource at the edge, effective optimization approaches need to be developed.

4) User privacy and security concern: Due to the pervasive deployment of MEC nodes, authentication and confidentiality are key concerns of MEC networks and mobile users. In particular, when there is a content request from a mobile user, the MEC node needs to verify whether the user is allowed to use the service or not. Furthermore, the mobile users always do not want to share their access histories which may reveal their private information. However, this information is essential for proactive caching solutions. As a result, security approaches need to be considered to resolve these problems. Alternatively, unlike the content server, MEC nodes have limited computing resources, and thus conventional privacy-preservation techniques, e.g., K-anonymity, L-diversity, and T-closeness, may not be effective. Moreover,

MEC infrastructure providers and mobile service operators can be separate entities, which makes privacy-preservation issue for the mobile users more complicated. Unfortunately, currently there is no security solution proposed for MEC, which can guarantee its success [6].

III. Smart Caching Approaches in Mobile Edge Caching Systems

Due to the aforementioned special characteristics and requirements of MEC networks, smart caching approaches were proposed to address these challenges.

A. Location-Based Caching Policies

In [5], the authors investigated the performance of an edge caching network for mobile video delivery using measurement studies and trace-driven experiments. In particular, the authors used a real-world dataset containing 50 million trace items of nearly 2 million users viewing more than 0.3 million unique videos on mobile devices in a metropolis in China. From this dataset, it is observed that the top 0.3% videos have relatively different popularity ranks at different locations. This implies that global popularity cannot be directly used to infer the local popularity of mobile video contents. Thus, the authors adopted an entropy analysis approach to analyze the video request entropy and the location request entropy. Then, it was concluded that the least frequently used method is better for locations with smaller location entropy, while the least recently used method is better for locations with larger location entropy. Then, based on the measurement insights, a caching strategy was also proposed. Trace-driven experiments then verified the effectiveness of the proposed solution. However, the security and video streaming pre-processing at the edge were not taken into considerations.

B. Delay Tolerance Caching Policies

The main aim of deploying MEC nodes is to improve the QoS for mobile users by reducing the delay of mobile users' requests. The delay of a content request is composed of two major components, i.e., time to pre-process the content and time to download the content. First, the pre-processing time depends on the type of data. Many recent advanced mobile applications require complicated computations before the applications can be executed, e.g., image processing, video transcoding, and game rendering. The pre-processing time at the edge can thus be long due to its limited computing resource. Additionally, the downloading time depends on the content size and the allocated bandwidth between the mobile user and the MEC node. Therefore, to minimize the delay for a request, both components must be taken into account. In [6], the authors introduced a joint caching and pre-processing solution which trade-offs edge

caching storage and computing resources with backhaul bandwidth consumption. In this approach, popular videos are distributed to the edge nodes until their cache storage are full. Then, when a user requests a video with a different version in the cache, the transcoding task will be assigned to a free edge node or the public content server. However, this approach does not consider the location-based popularity of contents as well as the security issue for mobile users.

C. Dynamic Pricing Caching Model

In [7], the authors introduced a dynamic pricing caching model for layered video contents in a mobile social network. In this model, instead of storing entire videos, some layers/descriptions of the videos can be cached at the MEC nodes. When a mobile user requests for a certain video, the cached layers can be immediately sent to the user from the corresponding MEC node, while the rest of video layers will be fetched from the original video server afterward. To find optimal caching layers at the MEC node given its caching capacity and offered price, the authors modeled the interaction between the MEC node and the mobile users as a Stackelberg game, while the competition among the mobile users was formulated as a non-cooperative game. The results showed that the proposed solution can achieve a higher hit ratio and a lower delay than those of uniform and random cache allocation schemes. However, this approach is limited to video streaming applications since videos can be layered and transmitted sequentially, and thus it is applicable to general contents.

D. Proactive Caching

The aim of proactive caching is to predict the users' demands and their content popularity, thereby making an optimal caching decision. Thus, the proactive caching solutions maximize the cache hit rate, thereby improving the QoS for mobile users and minimizing service costs for mobile network providers. In [8], the authors introduced a centralized network architecture in which a big data platform is deployed at the core to track and predict content demands of all users in the MEC network. In this approach, all users' information collected from the MEC nodes will be sent to the platform to analyze and extract useful knowledge. Nevertheless, this is a centralized architecture which has many limitations such as scalability and reliability. In [9], a decentralized and scalable proactive caching approach was proposed. Specifically, through analyzing YouTube requests observed at the MEC node over a 20-month period, the authors calculated statistics for the number of requests per videos and derived the distribution of the total number of times that a video will be requested. Although the proposed solution can address the ephemeral content popularity problem through a discrete power-law distribution function, its performance mainly relies on

the accuracy of this function. Nevertheless, this function is not guaranteed to have the same accuracy for all datasets with different mobile users' behaviors in different places.

IV. Dynamic Mobile Edge Caching Approach

In this section, we introduce a framework which addresses the aforementioned problems. Specifically, we propose a novel security method which can be deployed effectively at an MEC node to authenticate the legitimate users without a need of a centralized node and/or a security third-party. Alternatively, this method also allows the MEC node to use the users' content access history without revealing their private information. Then, a proactive caching approach based on matrix factorization technique is adopted to help the MEC node dynamically update and predict the users' content demands. Based on this prediction, a mixed integer linear programming algorithm is proposed to find an optimal caching policy which minimizes the average delay for the mobile users. Furthermore, we deliver the optimal business strategy for the mobile service provider to maximize its revenue. In Table I, we summarize and compare features of our proposed framework and other related works in the literature.

Callout: Table I Comparisons.

A. The Proposed Dynamic Mobile Edge Caching System

Fig. 2 illustrates the proposed smart edge caching framework with four main features.

Callout: Figure 2 An illustration of the proposed smart caching system.

1) Verification: In this paper, we adopt the public key cryptography which has been implemented successfully in blockchain networks [10], e.g., Bitcoin, ColoredCoins, and Bitcoincash, for mobile users's request authentication process to overcome limitations of conventional authentication methods, e.g., through SMS or emails. In particular, in our proposed authentication method, when a mobile user subscribes to the edge caching service provided by the mobile network operator, the operator will allocate a digital wallet to the user. This wallet is actually a mobile software used to generate a pair of public and private keys for the user. The private key of the user is kept secretly and only the user knows it, while the public key of the user can be sent to edge nodes for verification. Then, when the mobile user sends his access request with his public key to an edge node. The edge node can verify his public key, generate a smart contract including its public key, and send the contract back to the user. The mobile user then verifies the edge's

public key, signs the contract using his signature generated by his private key, and sends the contract to the edge node, the edge node verifies the contract and transmits the requested content to the user if the contract is authenticated. With the proposed verification method, mobile users can be authenticated without revealing their private information, so we can reduce the risk of exposure and spoofing by access points. Furthermore, this method can be performed in a decentralized way without a need of a centralized node and/or a security third-party as the method used by Nuxeo Edge Cache and Microsoft Azure Redis Cache. Thus, this method can be implemented effectively in mobile edge caching networks where edge nodes are deployed in a decentralized way, and they may not belong to the mobile network operator.

2) Dynamic demand prediction: After the request is verified, it will be stored in the *request log file*. This file contains a table with two parameters, i.e., users' public keys and contents. Each cell in this table represents the number of accesses of a user to a content. If the user has never accessed a content before, the corresponding cell is empty. This table is especially important in dealing with big data and dynamic demand problems of the mobile users. In particular, when a new user joins the network or a new content is made available, we just need to add one more column or row, respectively, to the table. Furthermore, we apply the one-year storage rule for the request log file. It was pointed out that users' demands and the popularity of contents are very dynamic and change over time [11]. Thus, if a user does not access any content in a sufficient long period, e.g., 6 months, this user information will be removed from the table. Similarly, if any content has not been accessed by any user in the table in a sufficient long period, e.g., 6 months, the information of this content will be removed from the table. This policy is not only to address the dynamic demand of users and the popularity of contents, but also to deal with the big data problem when the number of records quickly grows over the time.

After the information of users and contents is collected in the request log file, the MEC node needs to predict the real demand of users over contents in the future to make appropriate caching strategies. In this paper, we adopt non-negative matrix factorization (NMF) [2], an advanced machine learning technique for the demand prediction. There are many advantages of using NMF compared with Singular Value Decomposition (SVD) which was used in [8] for proactive caching. First, the request log file contains information about the number of accesses to the contents of the users, and the numbers are non-negative numbers which perfectly fit with the NFM technique. Second, the standard SVD assumes all empty entries are zeros. This leads to a poor prediction accuracy, especially when a dataset is extremely sparse. Third, unlike the SVD, the NMF provides an intuitive interpretation (as illustrated in Fig. 2) which allows to analyze and predict the demands of users and the popularity of contents correctly.

Consider a data matrix $\mathbf{A}$ of M rows and N columns with all elements greater than or equal to zeros. Its NMF seeks matrices $\mathbf{P}$ and $\mathbf{Q}$ of size M rows and K columns, and K rows and N columns, respectively, such that $\mathbf{PQ} = \mathbf{C} \approx \mathbf{A}$, and every element of matrices $\mathbf{P}$ and $\mathbf{Q}$ is either zero or positive. Here, each row of $\mathbf{P}$ would represent the strength of the associations between a user and its contents. Similarly, each row of $\mathbf{Q}$ would represent the strength of the associations between a content and its users [2]. To find $\mathbf{P}$ and $\mathbf{Q}$, the NMF uses an iterative procedure to modify the initial values of $\mathbf{P}$ and $\mathbf{Q}$ so that their product approaches $\mathbf{A}$. At each iteration the gradient descent technique is used to minimize the approximation error defined by the Frobenius norm $||\mathbf{A} - \mathbf{PQ}||_{\mathbf{F}}^2$. This procedure terminates when the approximation error converges or the specified number of iterations is reached. After that, we can derive the factorized matrix $\mathbf{C}$ from $\mathbf{P}$ and $\mathbf{Q}$.

It is worth noting that for a MovieLens 1M Dataset [12] with 1 million ratings from 6,000 users on 4,000 movies, when implementing and executing the NMF algorithm on our machine equipped with Intel(R) Xeon(R) CPU E5-1650 @3.2 GHz 16GB RAM, it takes around 137 seconds. Thus, the algorithm can be run offline at the MEC node, and it can be performed frequently to update the demand prediction of the users.

3) Optimal caching policy: After obtaining the prediction access-demand matrix by the NMF technique, we need to find an optimal caching policy for the MEC node. In particular, we denote $\mathbf{C}$ as the factorized matrix with N rows and M columns corresponding to N contents and M users. Different contents may have dissimilar data sizes denoted by $\mathbf{s} = \{s_1, \ldots, s_N\}$, and they may require disparate pre-processing time before the users can download them. For example, mobile devices are usually equipped with small screens with different aspect ratios (e.g., 4:3, 3:2, and 16:10), and thus videos and images should be pre-processed to fit with users' devices before the users download them. Due to the hardware limitation at the edge node, the pre-processing time at the edge for a content may be different from that of the public content servers. Hence, we denote $\mathbf{t}^e = \{t_1^e, \ldots, t_N^e\}$ and $\mathbf{t}^c = \{t_1^c, \ldots, t_N^c\}$ as the pre-processing time for the contents at the edge node and at the content server, respectively.

If we denote b_m as the bandwidth allocated to user m by the edge node, and I as the bandwidth from the edge node to the public cloud, then the delay when a user m requests to download a content n from the edge node is $\left(\frac{s_n}{b_m} + t_n^e\right)$ and from the public cloud is $\left(\frac{s_n}{b_m} + \frac{s_n}{I} + t_n^c\right)$. If we denote S as the storage

capacity of the edge node, the caching optimization problem can be formulated as follows:

$$\begin{aligned} \min_{\mathbf{X}} f(\mathbf{X}) = & \sum_{n=1}^{N}\sum_{m=1}^{M}\left(c_{nm}\Big(\frac{s_n}{b_m}+x_n t_n^e\Big)+c_{nm}\Big(1-x_n\Big)\Big(\frac{s_n}{I}+t_n^c\Big)\right) \\ & \text{s.t.} \quad \sum_{n=1}^{N} x_n s_n \leq S, \\ & \text{and} \quad x_n \in \{0,1\}, \ \ \forall n \in \{1,2,\ldots,N\}, \end{aligned} \tag{1}$$

where $\mathbf{X} = [x_1, \ldots, x_N]^\top$ is the decision variables and c_{nm} is the element of matrix $\mathbf{C}$ at row n and column m. In (1), we aim to minimize the total delay for all users at this edge node. The first and second terms of the objective function represent the total delay of contents downloaded from the edge node and the content server, respectively. The first constraint in (1) is to guarantee that the total size of all cached contents does not exceed the storage capacity of the edge. The second constraint is to express that the decision variables are binary variable in which values 1 and 0 correspond to decisions "cache" and "do not cache", respectively. Since $\mathbf{X}$ are binary variables, we adopt the Mixed-Integer Linear Programming (MILP) algorithm to find the caching policy for the edge node.

4) Business strategy of the mobile network operator: To minimize the implementation and deployment costs, the mobile network operator often finds a partner in content delivery, e.g., Akamai and Microsoft Azure, to deploy the MEC network. However, deploying the MEC network is not free. In practice, MEC infrastructure providers offer MEC packages with different prices for various caching capacities [13]. We denote $\mathcal{P} = \{P_1, \ldots, P_k, \ldots, P_K\}$ as the set of MEC packages offered to the mobile network operator. Each package P_k has a caching storage capacity S_k with a corresponding renting cost C_k in a fixed time period. Given a caching storage capacity S_k, the mobile network operator can find the optimal caching policy which minimizes the average delay for the mobile users by solving (1). Clearly, the larger the caching capacity is, the lower the average delay is. However, the renting cost is higher. Therefore, the mobile network operator has to determine which MEC package should be purchased such that its profit is maximized.

If we denote p_k as the price which the mobile network operator offers to the mobile users to download a unit of data (e.g., $\$0.01$ per 1MB) with average delay d_k (e.g., 0.1 second per 1MB), we can determine the average revenue at this edge by $p_k \times \overline{\mathcal{U}}$, where $\overline{\mathcal{U}}$ is the average usage demand in this location over the time period. The value of $\overline{\mathcal{U}}$ can be determined straightly from the factorized matrix by $\overline{\mathcal{U}} = \sum_{n=1}^{N}\sum_{m=1}^{M} c_{nm} s_n$. Note that p_k is a function of caching storage capacity which can be determined by $p_k = p^\dagger + \delta(d^\dagger - d_k^*)$. Here, $p^\dagger$ and $d^\dagger$ are base price and delay without caching, respectively. $\delta > 0$ is the conversion parameter and d_k^* is the optimal average delay obtained by (1). The higher the caching storage is, the lower the

average delay is, and thus the higher the offered price is. Then, the average revenue over the time period is $\overline{\mathcal{R}}(P_k) = p_k \times \overline{\mathcal{U}} - C_k$. Given a service package P_k, we can find the corresponding renting cost C_k and offered price p_k. Thus, for the MEC node with the average demand $\overline{\mathcal{U}}$, we can find the optimal service package P^* to maximize the average revenue for the mobile network provider. Note that considering the impact of service prices offered by the mobile network operator to the mobile users' decisions, contract theory and/or matching theory can be adopted to address this problem. For example, we can adopt the monopolist-dominated quality-price contract strategy proposed in [14] for the mobile network provider to offer a set of quality-price combinations to the mobile customers.

B. Numerical Results

To simplify the presentation of the numerical analysis, we use the following setting. The number of contents is 10 with different data size $\{70, 50, 5, 10, 40, 75, 20, 85, 47, 30\}$, e.g., Megabytes (MB), corresponding to different access frequency, i.e., the number of times requested, $\{38, 32, 41, 42, 39, 28, 37, 40, 22, 12\}$. This access frequency is derived from the matrix $\mathbf{C}$ after the normalization process using the NMF technique. The bandwidth allocated to each mobile user is 5Mbps and the bandwidth between the edge node and the public cloud is 500Mbps. The storage capacity is varied to evaluate the caching policy as well as the business strategy of the caching service provider. To evaluate the performance of the proposed solution, we compare with two other schemes, i.e., most frequently used (MFU) and without caching policies. For the MFU policy, contents with the highest access frequency will be stored in the edge storage.

We first vary the storage capacity of the edge node to evaluate the average delay of mobile users. In Fig. 3, we consider two cases, i.e., without and with pre-processing time, corresponding to Fig. 3(a) and Fig. 3(b), respectively. The pre-processing time for the contents is set at $\{1.5, 0, 0, 0, 0.5, 0, 0, 2, 0, 0\}$ on the edge node and at $\{0.5, 0, 0, 0, 0.1, 0, 0, 0.5, 0, 0\}$ on the content server. In Fig. 3(a), as the storage capacity increases, the average delay reduces gradually. Here, the average delay obtained by the proposed solution is always lower than that of the MFU policy. When the caching capacity is over 450MB, all the contents can be cached at the edge, and thus the average delays of the optimal policy and the MFU policy are the same. When we consider the pre-processing time, the delay obtained by the proposed solution is much lower than that of the MFU policy as shown in Fig. 3(b). The delay of the MFU policy is even higher than that in the case without caching. The reason is that the MFU does not consider the delay caused by pre-processing time of the contents. This is to show the importance of jointly considering the delays caused by pre-processing, content sizes, and access frequency.

Callout: Figure 3 The average delay of the mobile users when the storage capacity is varied (a) without considering the pre-processing time and (b) with pre-processing time.

We then study the business strategy of the mobile service provider given the demand of users at the MEC node, the offered mobile service price, and the caching price. We set the base price (i.e., price without caching) for the mobile users at 0.01 monetary unit (MU) per 1MB. The conversion parameter is set at 0.05. The caching prices are set at 30MUs for the first 50MB and 7.5MUs for the next 50MB. As shown in Fig. 4(a), given the demand of users at this location and the caching prices, the mobile service provider will choose to rent 400MB to maximize its revenue. However, if the demand of users at this location reduces by 30%, the service provider will rent only 350MB as shown in Fig. 4(b). Again, the average revenue obtained by the optimal policy always achieves the best performance compared with the MFU and without caching policies.

Callout: Figure 4 The average revenue when the storage capacity is varied (a) with 100% demand and (b) with 70% demand of the mobile users at this location.

C. Challenges and Open Research Issues

1) Mobile edge caching network deployment: A fundamental question for a mobile network operator when implementing an MEC network is how to efficiently deploy MEC nodes. With more MEC nodes being deployed, the delay experienced by the mobile users decreases, but the implementation cost increases. Thus, the mobile network operator needs to tradeoff between the delay requirement and the implementation cost to maximize its revenue. Alternatively, the mobile network operator needs to analyze its mobile users' content demands at different locations to determine where MEC nodes should be deployed to mitigate the mobile backhaul network.

2) Collaborative caching in mobile edge caching network: In this article, we consider the scenario when the MEC node is connected to the content server and downloads contents from the server when the requested contents are not stored locally. However, when an MEC network is deployed, an MEC node can download contents from other MEC nodes as well. Thus the MEC node has to decide whether the requested contents should be cached or not, and if the requested contents are not cached, where they can be downloaded from. However, the challenge is that cached contents at the MEC nodes and the connections from an MEC node to other MEC nodes can be changed over time. Therefore, how to design a dynamic

protocol which allows an MEC node to update information from other nodes in the network and make an optimal caching scheme accordingly in a decentralized manner remains an open question.

3) Payment management in mobile edge caching network: When a mobile user downloads a content from an MEC node, the mobile user has to pay for its content request. The payment information will be then sent to the mobile network operator to charge the mobile user. However, due to the decentralized architecture of MEC networks, managing payments in a centralized manner as current conventional methods, i.e., all payments will be sent to the mobile network provider for further processing, may not be efficient. Recently, the blockchain technology [15] has been introduced as an effective way to manage data for decentralized systems. With blockchain, when a mobile user performs a transaction, i.e., a payment, this transaction will be verified by other MEC nodes in the network parallelly and transparently through mining processes. In this way, we can prevent data manipulation and fraud of transactions, and the network does not need a central authority. However, how to identify miners and how to encourage MEC nodes to perform mining processes in MEC networks are still an open question.

V. Conclusion

In this paper, we provided an overview of recent development of mobile edge caching and identified its challenges and open research problems. We then introduced a dynamic edge caching framework for mobile 5G networks, leveraging the blockchain transactions and non-negative matrix factorization. This framework allows the MEC node to authenticate the legitimate users and use the users' access history to predict the content demand without disclosing the users' private information. To predict the demand of users and find the optimal caching policy for the MEC node, non-negative matrix factorization and mixed integer linear programming have been adopted. The optimal caching policy jointly considers content sizes, access frequency, bandwidth allocation, and pre-processing time of contents. Furthermore, we have studied the business strategy for the mobile network operator to maximize its revenue given the optimal caching policy, the demand of the mobile users, and the service prices. The numerical results then have clearly shown the efficiency of the proposed solution.

Acknowledgements

This work was supported in part by Singapore MOE Tier 1 under Grant 2017-T1-002-007 RG122/17 and RG 33/16, MOE Tier 2 under Grant MOE2014-T2-2-015 ARC4/15, NRF2015-NRF-ISF001-2277, MoE Tier 1 RG 33/16, US NSF CNS-1717454, CNS-1731424, CNS-1702850, CNS-1646607, and ECCS-1547201.

REFERENCES

[1] Mobile Edge Computing Platforms for Outdoor Telecom Application, ADLINK Technology Inc. Available Online: http://www.adlinktech.com/PD/marketing/OtherDocument/SETO-1000/SETO-1000_OtherDocument_en_1.pdf. Last Accessed on Oct 2017.

[2] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, Aug. 2009, pp. 42-49.

[3] Cisco visual networking index: Global mobile data traffic forecast update 2015-2020, Cisco, White paper, Feb. 2016.

[4] S. Wang, X. Zhang, Y. Zhang, L. Wang, J. Yang, and W. Wang, "A survey on mobile edge networks: Convergence of computing, caching and communications," *IEEE Access*, vol. 5, Mar. 2017, pp. 6757-6779.

[5] G. Ma, Z. Wang, M. Zhang, J. Ye, M. Chen, and W. Zhu, "Understanding performance of edge content caching for mobile video streaming," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 5, Mar. 2017, pp. 1076-1089.

[6] T. X. Tran, A. Hajisami, P. Pandey, and D. Pompili, "Collaborative mobile edge computing in 5G networks: New paradigms, scenarios, and challenges," *IEEE Communications Magazine*, vol. 55, no. 4, Apr. 2017, pp. 54-61.

[7] Z. Su, Q. Xu, F. Hou, Q. Yang, and Q. Qi, "Edge caching for layered video contents in mobile social networks, *IEEE Transactions on Multimedia*, vol. 19, no. 10, Jul. 2017, pp. 2210-2221.

[8] 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 Communications Magazine*, vol. 54, no. 9, Sept. 2016, pp. 36-42.

[9] N. Carlsson and D. Eager, "Ephemeral content popularity at the edge and implications for on-demand caching," *IEEE Transactions on Parallel and Distributed Systems*, vol. 28, no. 6, Jun. 2017, pp. 1621-1634.

[10] K. Christidis and M. Devetsikiotis, "Blockchains and smart contracts for the internet of things," *IEEE Access*, vol. 4, May 2016, pp. 2292-2303.

[11] M. Cha, H. Kwak, R. Rodriguez, Y. Y. Ahn and S. Moon, "Analyzing the video popularity characteristics of large-scale user generated content systems," *IEEE/ACM Transactions on Networking*, vol. 17, no. 5, Oct. 2009, pp. 1357-1370.

[12] MovieLens 1M Dataset. Available Online: https://grouplens.org/datasets/movielens/1m/. Last Accessed on Oct 2017.

[13] Microsoft Azure. Available Online: https://azure.microsoft.com/en-us/pricing/details/managed-cache/. Last Accessed on Oct 2017.

[14] L. Gao, X. Wang, Y. Xu, and Q. Zhang, "Spectrum trading in cognitive radio networks: A contract-theoretic modeling approach," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 4, Apr. 2011, pp. 843-855.

[15] F. Tschorsch and B. Scheuermann, "Bitcoin and beyond: A technical survey on decentralized digital currencies," *IEEE Communications Surveys & Tutorials*, vol. 18, no.3, Jan. 2016, pp. 2084-2123.

PLACE PHOTO HERE

Dinh Thai Hoang (M'16) is currently a Lecturer (Assistant Professor) in the School of Electrical and Data Engineering, University of Technology Sydney, Australia. He received Ph.D. in School of Computer Science and Engineering from Nanyang Technological University, Singapore in 2016. His research interests include emerging topics in wireless communications and networking such as ambient backscatter, energy harvesting, IoT, mobile edge, blockchain, and 5G networks.

PLACE PHOTO HERE

Dusit Niyato (M'09-SM'15-F'17) is currently a professor in the School of Computer Science and Engineering, at Nanyang Technological University, Singapore. He received B.Eng. from King Mongkuts Institute of Technology Ladkrabang (KMITL), Thailand in 1999 and Ph.D. in Electrical and Computer Engineering from the University of Manitoba, Canada in 2008. His research interests are in the area of energy harvesting for wireless communication, Internet of Things (IoT) and sensor networks.

PLACE PHOTO HERE

Diep Nguyen (M'16) is a faculty member of the University of Technology Sydney (UTS). He received M.E. and Ph.D. in Electrical and Computer Engineering from the University of California, San Diego and The University of Arizona, respectively. Before joining UTS, he was a DECRA Research Fellow at Macquarie University, a technical staff at Broadcom (California), ARCON Corporation (Boston), consulting the Federal Administration of Aviation on turning detection of UAVs, Air Force Research Lab on anti-jamming.

PLACE PHOTO HERE

Eryk Dutkiewicz received the B.E. degree in electrical and electronic engineering and the M.Sc. degree in applied mathematics from the University of Adelaide in1988 and 1992, respectively, and the Ph.D. degree in telecommunications from the University of Wollongong in 1996. His industry experience includes management of the Wireless Research Laboratory, Motorola in 2000. He is currently the Head of the School of Electrical and Data Engineering, University of Technology Sydney, Australia. He has held visiting professorial appointments at several institutions including the Chinese Academy of Sciences, Shanghai Jiao Tong University, and Macquarie University. His current research interests cover 5G networks and medical body area networks.

PLACE PHOTO HERE

Ping Wang (M'08-SM'15) received the PhD degree in electrical engineering from University of Waterloo, Canada, in 2008. Currently she is an Associate Professor in the School of Computer Science and Engineering, Nanyang Technological University, Singapore. Her current research interests include resource allocation in multimedia wireless networks, cloud computing, and smart grid. She was a corecipient of the Best Paper Award from IEEE Wireless Communications and Networking Conference (WCNC) 2012 and IEEE International Conference on Communications (ICC) 2007. She served as an Editor of IEEE Transactions on Wireless Communications, EURASIP Journal on Wireless Communications and Networking, and International Journal of Ultra-Wideband Communications and Systems.

PLACE PHOTO HERE

Zhu Han (S'01-M'04-SM'09-F'14) received the B.S. degree in electronic engineering from Tsinghua University, in 1997, and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Maryland, College Park, in 1999 and 2003, respectively. From 2000 to 2002, he was an R&D Engineer of JDSU, Germantown, Maryland. From 2003 to 2006, he was a Research Associate at the University of Maryland. From 2006 to 2008, he was an assistant professor at Boise State University, Idaho. Currently, he is a Professor in the Electrical and Computer Engineering Department as well as in the Computer Science Department at the University of Houston, Texas. His research interests include wireless resource allocation and management, wireless communications and networking, game theory, big data analysis, security, and smart grid. Dr. Han received an NSF Career Award in 2010, the Fred W. Ellersick Prize of the IEEE Communication Society in 2011, the EURASIP Best Paper Award for the Journal on Advances in Signal Processing in 2015, IEEE Leonard G. Abraham Prize in the field of Communications Systems (best paper award in IEEE JSAC) in 2016, and several best paper awards in IEEE conferences. Currently, Dr. Han is an IEEE Communications Society Distinguished Lecturer.

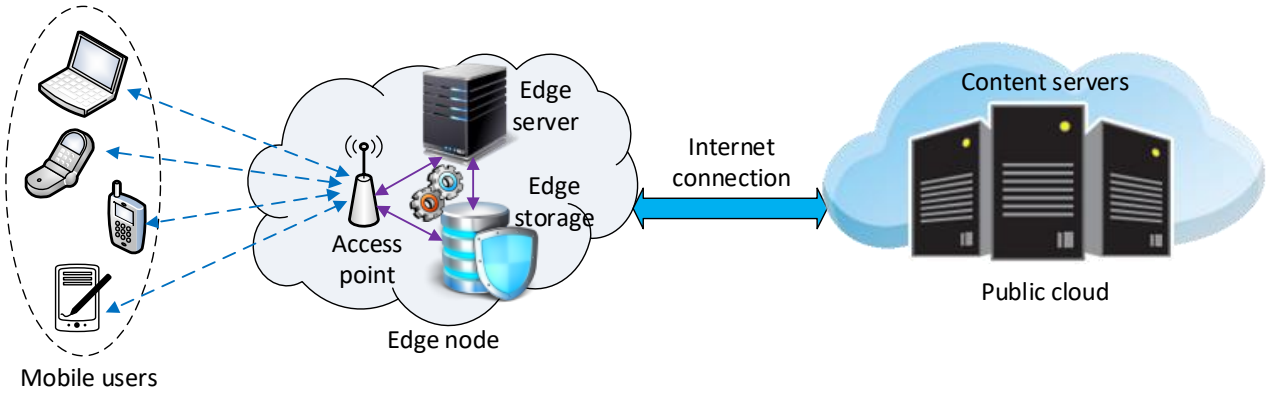


Fig. 1. A general architecture of a mobile edge caching network.

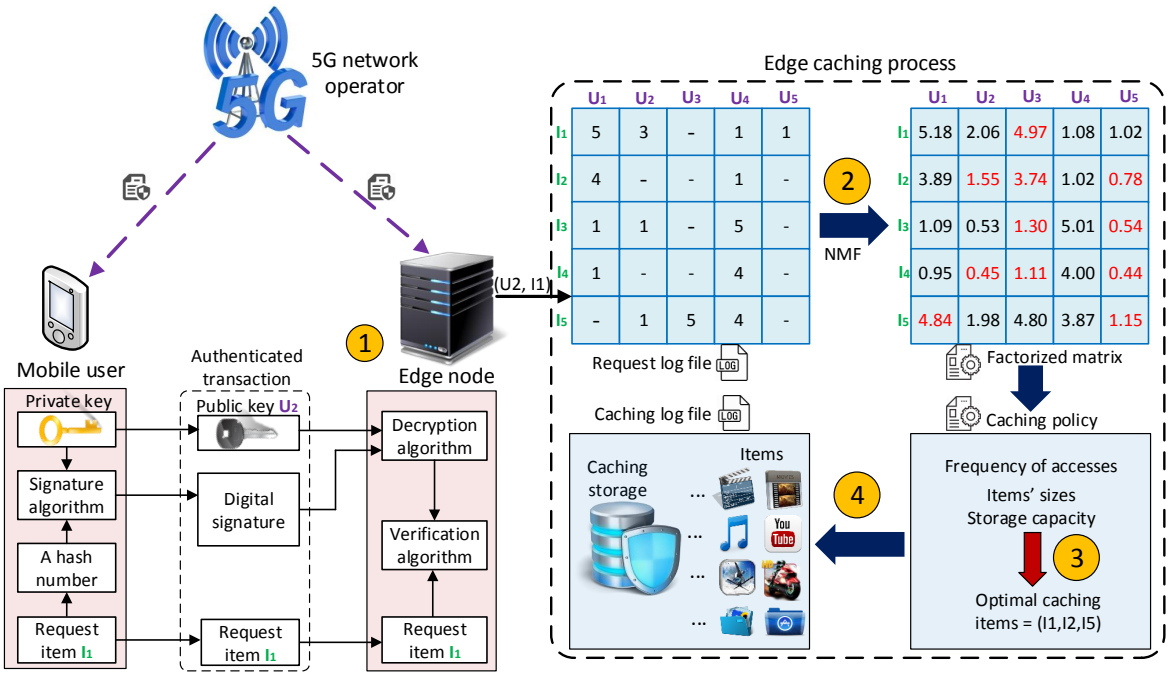
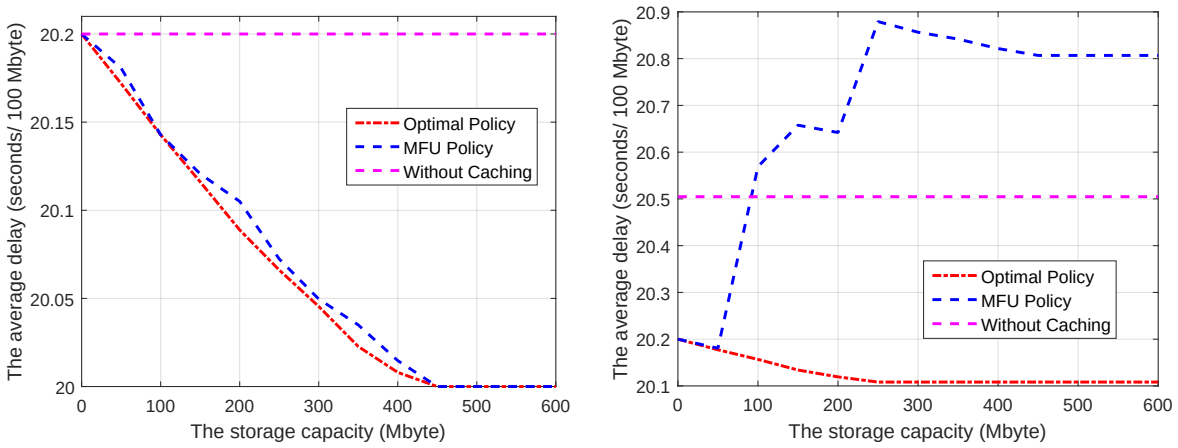
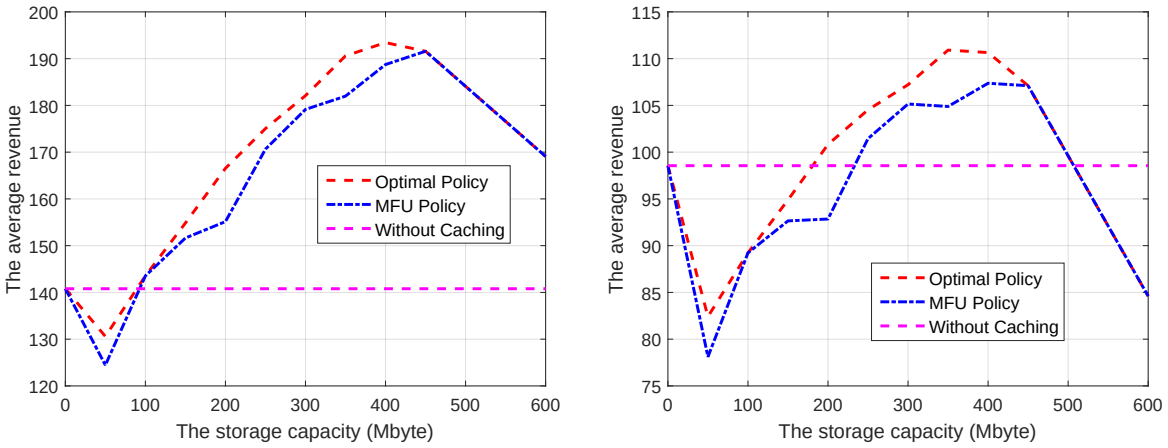


Fig. 2. An illustration of the proposed smart caching system.



(a) (b)

Fig. 3. The average delay of the mobile users when the storage capacity is varied (a) without considering the pre-processing time and (b) with pre-processing time.



(a) (b)

Fig. 4. The average revenue when the storage capacity is varied (a) with 100% demand and (b) with 70% demand of the mobile users at this location.

TABLE I

COMPARISONS

Features	[5]	[6]	[7]	[8]	[9]	The proposed framework
Location-based caching	✓	✓	✓		✓	✓
Proactive caching	✓			✓	✓	✓
Pre-processing time		✓				✓
Dynamic pricing			✓			✓
Authentication & Confidentiality						✓