This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2019.2908280, IEEE Communications Surveys & Tutorials

On Mobile Edge Caching

Jingjing Yao, *Student Member, IEEE*, Tao Han, *Member, IEEE*, Nirwan Ansari, *Fellow, IEEE*

***Abstract*—With the widespread adoption of various mobile applications, the amount of traffic in wireless networks is growing at an exponential rate, which exerts a great burden on mobile core networks and backhaul links. Mobile edge caching, which enables mobile edges with cache storages, is a promising solution to alleviate this problem. In this paper, we aim to review the state of the art of mobile edge caching. We first present an overview of mobile edge caching and its advantages. We then discuss the locations where mobile edge caching can be realized in the network. We also analyze different caching criteria and their respective effects on the caching performances. Moreover, we compare several caching schemes and discuss their pros and cons. We further present a detailed and in-depth discussion on the caching process, which can be delineated into four phases including content request, exploration, delivery and update. For each phase, we identify different issues and review related works in addressing these issues. Finally, we present a number of challenges faced by current mobile edge caching architectures and techniques for further studies.**

***Index Terms*—Mobile edge computing, mobile edge caching, 5G, content management, content delivery.**

I. Introduction

Nowadays, mobile data traffic is experiencing explosive growth due to pervasive mobile services, ubiquitous social networking, and resource-intensive applications. According to Cisco [1], mobile data traffic is predicted to increase 500-fold in the next decade. It is estimated that a mobile user will download around 1 terabyte of data in a year by 2020 [2]. The increasing mobile traffic is mainly incurred by the newly emerging applications of mobile devices, such as the Internet of Things (IoT), Internet of Vehicles (IoV), e-healthcare, machine to machine (M2M) communications and virtual/augmented reality, which require higher network throughput and stricter network latency [3]. These use cases accelerate the standardization process of the fifth generation (5G) wireless networks in aspects of the network capacity and latency, which cannot be fulfilled by the current fourth generation (4G) wireless networks [4]. 5G wireless networks are planned to be standardized by 2020, and are supposed to provide 1000 times more network capacity and less than 1 millisecond of network latency as compared to 4G networks [5].

The major standardization bodies of 5G consist of the 3rd Generation Partnership Project (3GPP) (which provides complete system specifications including core network, radio access network and service capabilities), Telecommunication Standardization Sector of the International Telecommunications Union International Mobile Communications 2020/Study Group 13(ITU-T IMT2020/SG13) (which is responsible for all 5G no-radio network segments including overall network architecture, softwarization and management), Internet Engineering Task Force (IETF) (which covers 5G no-radio network segments including network slicing), the European Telecommunications Standards Institute (ETSI) (which is responsible for network function virtualization, mobile edge computing, next generation protocols), and Institute of Electrical and Electronics Engineers (IEEE) (which provides the WiFi and WiMAX standards) [6]. In order to achieve lower network latency and larger network capacity, 5G solutions in radio access network (RAN) include modification of physical subframe, polar coding, turbo decoding, quick path-switching methods, control channel sparse encoding, outer-loop link adaptation, mmWave based air interface, location-aware communication, and cloud radio access networks [7]. The 5G core network solutions include software defined networks (SDN), network function virtualization (NFV) and mobile edge computing (MEC) [7].

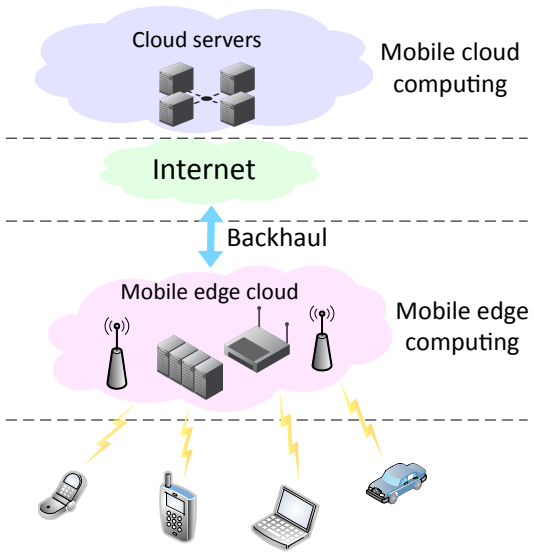


Fig. 1. Architecture of MCC and MEC

Mobile devices are now running more effective and powerful applications, which require more computational capacity, storage, bandwidth and energy. These applications generally include computationally and data intensive tasks, like computer vision, image processing, face recognition, optical character recognition, and augmented reality. However, the performance of mobile devices is often impoverished due to the limited computation, storage capacity and battery life. An

J. Yao and N. Ansari are with the Advanced Networking Laboratory, Helen and John C. Hartmann Department of Electrical and Computer Engineering, New Jersey Institute of Technology, Newark, NJ 07102 USA (e-mail: jy363@njit.edu; nirwan.ansari@njit.edu).

T. Han is with the Department of Electrical and Computer Engineering, University of North Carolina at Charlotte, Charlotte, NC 28223 USA (e-mail:than3@uncc.edu).

outstanding solution to address these limitations is to offload some computation to the cloud [8]. This solution is referred to as mobile cloud computing (MCC) [9], where the terminology "cloud" refers to a collection of servers usually located in a distant data center that provide adequate computing, storage and networking resources for mobile devices.

Since the delay of MCC is contributed by the core network, radio access network (RAN) and the backhaul links between them, the long latency between users and cloud becomes a challenge. New network architecture is required to satisfy the network latency required by 5G wireless networks for MCC. MEC, where cloud servers are placed in close proximity to users to provide computing and storage capacities, is proposed to address this challenge [12]. The architecture of MCC and MEC is illustrated in Fig. 1. Owing to the advantages of MEC, European 5G PPP (5G Infrastructure Public Private Partnership) has recognized MEC as one of the key technologies for 5G [13]. The ETSI Industry Specification Group (ISG) on MEC has produced normative group specifications with respect to the requirements, architecture and services of MEC and is developing the foundation to enable third party applications and content at the network edges [14]–[16]. By offloading the computing tasks to the edge servers instead of the remote cloud, the service response time can be greatly reduced and hence the user experience can be improved. The traffic through the backhaul links can also be alleviated. Furthermore, the edge servers can easily obtain the network status information (e.g., wireless channel conditions, traffic patterns and user mobility patterns), which can be analyzed and utilized to offer better services. For example, when the wireless channel conditions are poor, edge servers can allocate more computing resources to reduce the service processing time at the servers and hence compensate for the long wireless transmission latency. Hence, the total service response time (i.e., wireless transmission time plus service processing time) will not increase and the service performance will be guaranteed.

The MEC edge servers not only provides computing resources but also storage resources, and hence can be utilized as cache nodes to store popular contents requested by users. Caching at the mobile edge servers is referred to as mobile edge caching, which also helps alleviate the mobile traffic and reduce the content delivery latency [10]. Mobile edge caching is defined as the one of the applications of MEC by ETSI ISG in Reference [15]. In traditional cellular networks, user content requests are served by Internet content providers. When a user requests certain contents, the contents can only be obtained from the remote content servers. Some popular contents can be requested by users for multiple times. As a result, the content providers have to send the same contents repeatedly. This not only causes long service latency but also injects more traffic, thus potentially congesting the network. Mobile edge caching utilizes the cached-enabled edge servers to store popular contents so that these contents can be transmitted directly from the caches instead of from the remote cloud. Hence, the traffic load in backhaul links can be largely reduced. As filling the caches with popular contents from the remote cloud generates additional traffic into the network, it is usually realized during off-peak hours (e.g., nighttime) while serving requests with the cached contents during the peak-time (e.g., daytime).

The concept of caching has been well studied in web caching [17] and information centric networks (ICNs) [18], [19]. Ali *et al.* [17] discussed the existing web caching and prefetching approaches for ICN. To improve the content caching efficiency, many research efforts have been devoted to optimize the path selection [23], server placement [24] and content duplication strategy [25]. Scellato *et al.* [24] optimized the content placement problem by using geographic information extracted from social cascades. Borst *et al.* [25] presented algorithms to minimize bandwidth cost in content delivery networks with the assumption that the content popularity is given. However, the above caching schemes are only applicable to the wired networks. There are significant differences between caching in the wired and wireless networks. Wireless caching usually faces the challenge of dynamic traffic loads because of the user mobility. Therefore, network traffic is highly related to the user mobility pattern, and hence it is difficult to predict the network traffic. Furthermore, the wireless channels exhibit more uncertainties with limited spectrum resources and co-channel interference. Hence, designing caching schemes for wireless networks is more challenging.

Caching in wireless networks has been surveyed in several works. Han *at el.* [20] provided an overview of available solutions for content delivery acceleration in mobile networks. However, they only focused on content delivery, which is just one aspect of the caching process (including content placement, delivery and update). Liu *at el.* [21] discussed several issues in wireless edge caching including content placement and delivery. However, their work is not complete because they did not consider the content update problem of the caching process. They also did not survey the caching criteria and schemes. Wu *et al.* [22] discussed research challenges for mobile social device caching (MSDC), which is another aspect of mobile edge caching since it only solves the issues on device-to-device (D2D) caching. Wang *et al.* [10] discussed a broad topic about mobile edge networks including mobile edge computing, caching and communications. However, their work only covered general caching schemes and the criteria to be followed without discussing the specific phases of the caching process while our work focuses on mobile edge caching and hence is more specific. Li *et al.* [11] surveyed wireless caching techniques for cellular networks. However, they only focused on the content placement and delivery strategies in the physical and MAC layer. We additionally provide a survey according to caching locations, caching criteria and caching schemes. In particular, we also discuss the impacts of traffic patterns, user social relationships, user mobility patterns and content popularity on caching strategies. We summarize the overall related survey papers in Table I, where we list various aspects of caching in our work and compare ours against existing surveys. We also construe the structure of our work and detail the differences between our work and the two most related works [10] and [11] in Fig. 2. The flag marker represents the reference [10] and the star marker reflects the work [11].

Although many works have attempted to address various issues in mobile edge caching, a comprehensive summary is

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2019.2908280, IEEE Communications Surveys & Tutorials

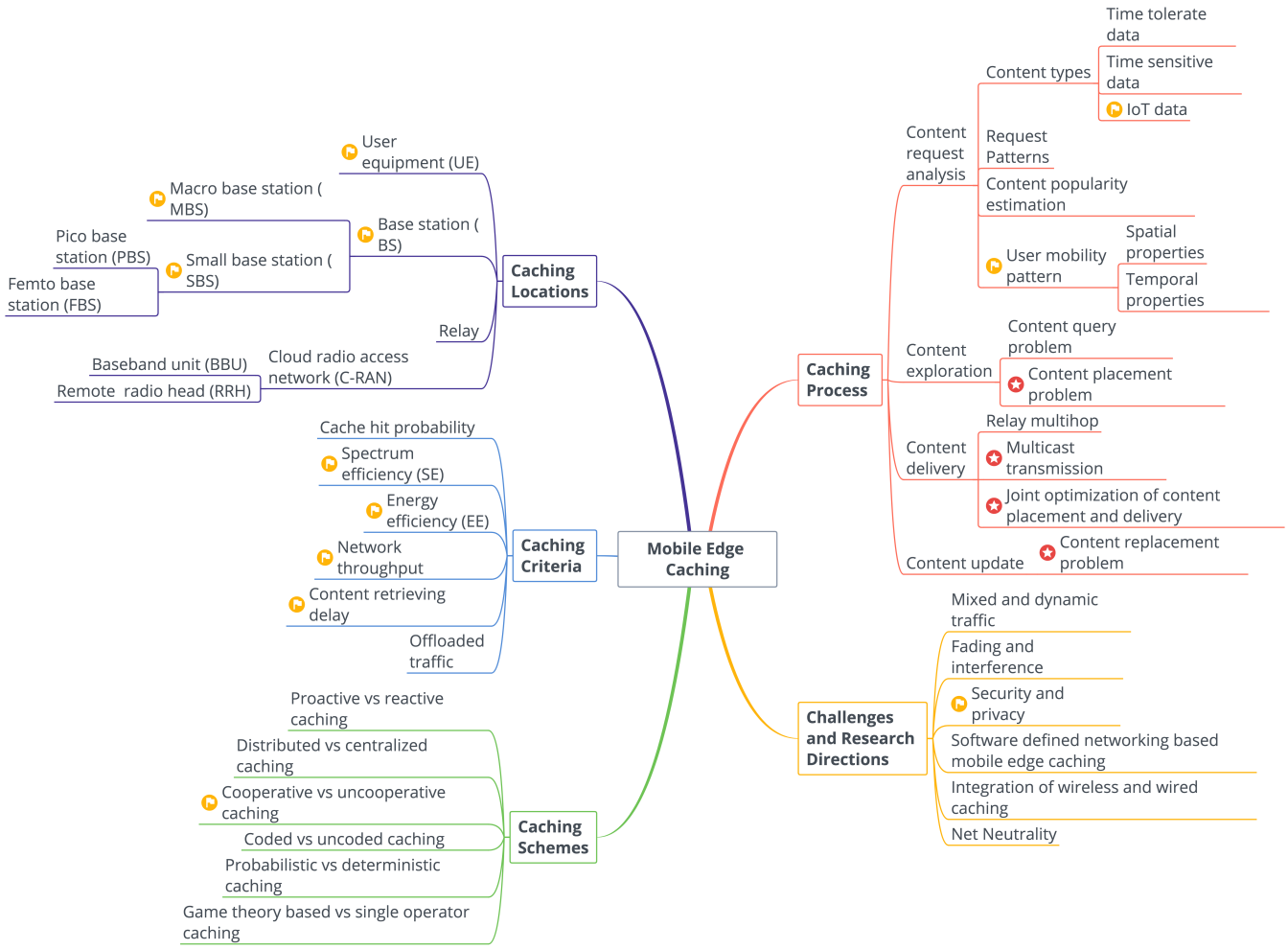


Fig. 2. Classification structure and differences from the two most related works (flag: [10], star: [11]).

still required. To fill this gap, we conduct a survey regarding different aspects of issues in mobile edge caching. For better readability, we summarize all abbreviations used in this paper in Table II. Hereafter, the "content", "file" and "data" are used interchangeably since all of them refer to the cacheable content. Note that MEC makes no assumption on the underlying mobile network infrastructure [26]. The MEC system can be deployed in the existing mobile networks and transited to future 5G (which is specified in [27]) by software upgrading network functions [26]. The deployment of MEC system in 4G and 5G networks is described in an ETSI whitepaper [26]. Specifically, in 4G and 5G networks, BSs do not read the user data packets, and instead transparently transfer packets to the users. Therefore, the analysis of content and caching functions are usually deployed in specialized servers co-located with BSs or packet gateways. However, our work does not focus on the physical implementation and deployment of MEC. Instead, we describe the caching strategies (i.e., content management strategies) on where, what and how to cache popular data to improve the performance of mobile edge caching systems. The caching strategies surveyed in our work, which are unaware of the physical functions of different mobile networks, should be applicable to current mobile networks and future 5G networks. We summarize our main contributions as follows.

1) We conduct a comprehensive survey of mobile edge caching and point out the advantages of mobile edge caching.
2) With regard to where to cache, we summarize the related works based on different caching locations.
3) To address the issue of how to cache, we classify publications based on different caching criteria. We also compare several caching schemes and analyze their advantages and disadvantages.
4) We summarize the caching process into four phases including the content request phase, content exploration phase, content delivery phase and content update phase. For each phase, we identify the related issues and review the corresponding publications.
5) We identify and analyze the challenges and research directions to motivate further research.

The rest of the survey is organized as follows. Section II provides an overview of mobile edge computing and mobile edge caching. Section III summarizes the possible locations where cache storages can be located. In Section V, several

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2019.2908280, IEEE Communications Surveys & Tutorials

Table I. Summary of related works.

Related works	Mobile network	Locations	Criteria	Schemes	Request analysis	Content exploration	Content delivery	Content update	Challenges & Directions
Web caching and prefetching [17]		✓	✓	✓				✓	
Caching in information centric networking [18]				✓				✓	✓
Caching mechanisms in information centric networking [19]			✓	✓		✓	✓		
Accelerating content delivery [20]	✓			✓			✓		
Caching at the wireless edge [21]	✓	✓				✓	✓		✓
Mobile Social Device Caching [22]	✓			✓		✓			✓
Mobile edge computing, caching and communications [10]	✓	✓	✓	✓					✓
Content placement and delivery strategies in cellular networks [11]	✓					✓	✓	✓	✓

caching schemes are introduced. The content request analysis is discussed in Section VI. Section VII presents several issues in the content exploration phase. The content delivery problem is described in Section VIII. In Section IX, content update problems are discussed. Section X discusses several challenges faced by current mobile edge caching. Section XI concludes the survey.

II. Mobile Edge Caching: An Overview

Since the data growth is a major challenge in today's mobile infrastructures, managing overloaded networks becomes an important issue. Moreover, user demands for mobile networks are becoming more stringent, e.g., the high data rate and low latency. At the same time, various applications have emerged such as virtual reality and IoT, which require better and faster services (i.e., high data rate and low latency). The traditional BS centric network architecture cannot fulfill these requirements anymore, and hence new network architectures are called for.

A. Mobile Edge Computing

By offloading the computation to the central cloud, MCC provides powerful computing and storage capability by leveraging the cloud platform [28]. However, it suffers from long latency between mobile devices and the cloud as well as high backhaul bandwidth consumption in backhaul links, and hence is not suitable for real-time applications. In order to solve this problem, MEC is proposed to move computing tasks and contents closer to end users [29].

MEC presents several advantages in various aspects. First, the network latency can be reduced significantly because computing and storage capabilities are provisioned in proximity to end users [30]. Second, the bandwidth consumption of backhaul links can be reduced due to the deployment of edge servers [31]. According to [32], the bandwidth of backhaul links can be saved up to 22% by caching. Third, since computing tasks are offloaded to the cloud, mobile and IoT devices do not have to consume their own computing resources to process these tasks. Hence, the energy consumption of mobile and IoT devices can be reduced [33]. Fourth, network status (e.g., wireless channel conditions, user preferences, etc.) can be collected by the cloud and analyzed to provide better services.

Table II. Summary of abbreviations.

Abbreviation	Full name
5G	5th generation
BS	base station
UE	user equipment
MIMO	multiple-input multiple-output
D2D	device-to-device
IoT	internet of things
IoV	internet of vehicles
M2M	machine to machine
MCC	mobile cloud computing
MEC	mobile edge computing
SE	spectrum efficiency
EE	energy efficiency
ICN	information centric network
MSDC	mobile social device caching
UE	user equipment
MBS	macro base station
SBS	small base station
PBS	pico base station
FBS	femto base station
RAN	radio access network
C-RAN	cloud radio access network
F-RAN	fog radio access network
BBU	baseband unit
RRH	radio remote head
QoE	quality of experience
QoS	quality of service
PPP	poison point process
TDMA	time division multiple access
SP	service provider
SINR	signal to interference plus noise ratio
ICI	inter-cell interference
HetNet	heterogeneous network
OFDMA	orthogonal frequency division multiple access
ILP	integer linear programming
BILP	binary integer linear programming
MIP	mixed integer programming
CSI	channel state information
BP	belief propagation
MNO	mobile network operator
SDN	software defined network

B. Mobile Edge Caching

Mobile edge caching, which utilizes the storages provided by mobile edge servers, is a use case of MEC [34]. A cache-enabled mobile edge server can be an independent server attached to a mobile edge node or the storage of the edge node. Without mobile edge caching, content requests of mobile users are usually served by remote Internet content servers provided by the content providers (e.g., web servers).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2019.2908280, IEEE Communications Surveys & Tutorials

When users retrieve the same popular contents from remote servers, remote servers have to send the same files repeatedly that may lead to duplicate traffic and network congestion. However, by caching popular contents closer to users, the latency for retrieving contents can be greatly reduced and the duplicate transmissions from content servers to the cache-attached network nodes can be avoided. Moreover, as the cost of storages is lowering, deploying caches at wireless edge becomes more cost effective [35].

In mobile edge caching, content requests, which are issued by user equipments (UEs), are responded by one of the nodes, which contain the requested content. Usually, the domain name system (DNS) redirects the request of the user to the nearest cache capable of satisfying the content. There are several advantages brought by mobile edge caching. First, as mobile edge caching is facilitated at the network edge which is closer to users than the remote Internet content servers, it reduces the latency of acquiring user requested contents. Second, mobile edge caching avoids the data transmissions through the backhaul links, and hence reduces the backhaul traffic. Third, mobile edge caching helps reduce energy consumption. For example, when the requested data are cached at the small cell base station, the energy consumption for transmitting data from the macro base station can be avoided. Fourth, the spectrum efficiency can be improved by mobile edge caching. For instance, when multiple users request the same content, the serving BS can transmit the cached file by multicasting, which shares the same spectrum [11]. Fifth, mobile edge caching can leverage the network information collected by the mobile edge servers (e.g., user preferences, file popularities, user mobility information, user social information and channel state information) to improve caching efficiency. For example, the user social relationships can be explored to cache and disseminate contents via D2D communications.

C. Issues regarding Mobile Edge Caching

The problems of where, how and what to cache are the key research issues in mobile edge caching. Where to cache refers to the selection of caching locations. Caching schemes can be implemented on UEs by utilizing their own storages. The cached content on UEs can be shared via D2D communications. Popular contents can also be cached at BSs, e.g., relays, femto base stations (FBSs), pico base stations (PBSs), small base stations (SBSs), and macro base stations (MBSs). In cloud radio access network (C-RAN) [36], the content can be cached at both remote radio heads (RRHs) and baseband unit (BBU) pools.

How to cache involves the problem of choosing caching criteria and designing caching schemes. To improve the performance of mobile edge caching, there are several basic criteria that should be followed. First, the cache hit probability should be high. Cache hit probability refers to the ratio of the number of cached files requested by the users over the total number of files in the caches. Second, SE and EE are major performance metrics in 5G, caching schemes should be designed to improve both of them. Third, minimizing content retrieving delay should be accounted for, as it directly relates to user quality of experience (QoE). Fourth, caching popular contents at the edge can offload traffic from backhaul links, and hence maximizing traffic offloading can be one of the caching criteria.

Several caching schemes have been studied recently. Reactive and proactive caching are proposed with regard to deciding whether to cache a content after or before it is requested. Based on where the caching decisions are made, caching schemes can be classified as centralized and distributed caching. Centralized caching uses a central controller to determine all content placement schemes while distributed caching is only aware of neighboring UEs' or neighboring BSs' information and makes decisions with respect to local caches. Since the caching space at the edge node is limited, designing caching policy for each node individually may cause insufficient utilization of caches. Cooperative caching copes with this problem because different caching nodes can share contents with each other. The under-utilized caches can be used by other nodes and hence the utilization rates of all caching nodes can be improved. Coded caching utilizes the network coding techniques where data messages are aggregated (encoding), forwarded to the same destination and then separated into different messages (decoding). This technique can increase network throughput and reduce delays by reducing the number of transmissions. Probabilistic caching deals with the problem of uncertainty and movement of user locations. Game theory based caching is used to analyze the cooperations and competitions among different service providers (SPs).

Deciding what to cache requires the awareness of content request patterns. Content types consists of multimedia files (e.g., videos and files) and IoT data, which exhibit more dimensions and shorter lifetimes. It is challenging to make decisions in an optimal manner without knowing the request patterns. In order to obtain the probability of a particular content being requested, we need to estimate the content popularity and user preference. User mobility pattern affects how users are associated with BSs and can further modify the request patterns received by mobile networks. The user association problem [37] determines which user is served by which BS and hence can affect the requests received by each BS.

The practical mobile edge caching process carries out the following phases.

1) Content request phase (Section VI): requests are initiated by users.
2) Content exploration phase (Section VII): the requested content is searched in mobile network to determine whether the content has been cached at one of the cache storages.
3) Content delivery phase (Section VIII): the requested content is delivered to users either from a cache node or from remote content service providers.
4) Content update phase (Section IX): caches are updated by evicting old ones and caching new ones according to the popularity information.

In the content exploration phase, when the mobile network receives a request, it searches for the content through the network. How to search for the content is referred to as

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2019.2908280, IEEE Communications Surveys & Tutorials

the content query problem. Where to find the content is determined by the content placement problem. The content placement problem addresses which content is cached at which caching storage. In the content delivery phase, the multicast transmission is usually utilized to improve the network throughput. Furthermore, the content delivery is closely related to the content placement results, thus justifying the joint optimization of them. As network traffic and user demands are dynamic, utilizing overly expired information may not guarantee the caching performance. Hence, caches should be updated every time interval, which is specified according to the variance of the traffic (e.g., one week for movies and two or three hours for news). In the content update phase, the content replacement problem should be well designed with regard to determining when to update the caches and which content should be removed from the caches.

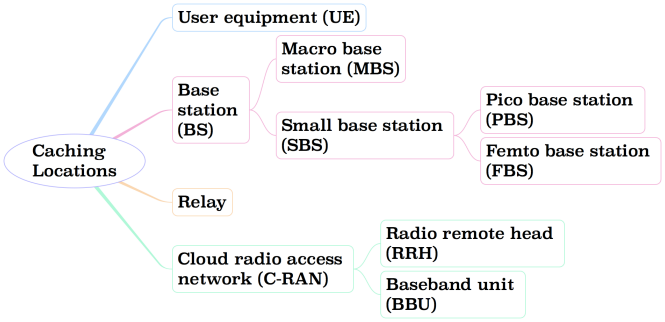


Fig. 3. Caching locations.

III. Caching Locations

We classify the existing research according to the caching locations in this section. As elicited in Fig. 3, the cache nodes can be deployed at UEs, MBSs, SBSs, PBSs, FBSs and relays in traditional cellular network, and RRHs and BBU pools in C-RAN.

A. Caching at UEs

Current smartphones are becoming more sophisticated with enhanced computing and storage capabilities. Therefore, mobile user devices can act as caches to store content locally and share content with other UEs directly via D2D links using licensed-band (e.g., LTE) or unlicensed-band protocols (e.g., Bluetooth and WiFi). This caching strategy is called D2D caching (Fig. 4) in which the BS usually keeps track of the caching status (e.g., availability of cached content and caching storage) in each UE and also directs requests from one UE to suitable nearby devices which have the requested content. If none of nearby devices possess the requested content, the BS then provides the content through downlink transmission. Different UEs can form different clusters and usually only the UEs within one cluster can communicate. Obviously, D2D caching can inherit all the benefits promised by D2D communications. These benefits include improving spectrum utilization, energy efficiency and throughput, and enabling new peer-to-peer and location-based applications and services [10]. Furthermore, UEs can cache contents according to their own preferences, thus providing higher caching flexibility as compared to caching at network entities [22].

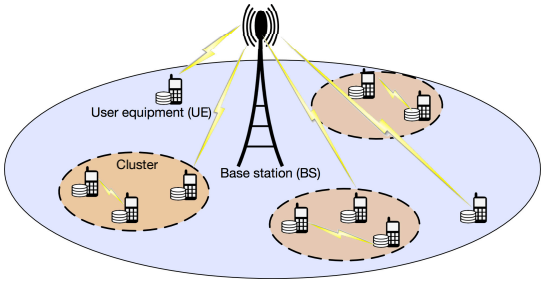


Fig. 4. Caching at UEs

Caching at UEs has long been applied as a technique to improve user QoE [38]. Pyattaev *et al.* [39] proposed an integrated two-tier wireless network including 3GPP LTE in licensed frequency bands and D2D system in unlicensed bands. In their network system, UEs can obtain content either by downloading from the BS or by establishing D2D links with other proximate UEs. They considered the characteristics of mass events (e.g., festivals, concerts and sport events) and focused on a particular isolated cellular cell. They also assumed that the frequency channels of the D2D and cellular systems are non-overlapping and hence there is no interference across two tiers. In the D2D system tier, however, transmissions from different neighboring UEs may still interfere with each other. Instead of a hierarchical two tier architecture, Ji *et al.* [40] proposed a wireless D2D grid network formed by UE nodes on the unit square without a central BS. Only the UE nodes, which are within a certain pre-defined distance, can communicate with each other. Each UE node arbitrarily caches several packets of a file from the file library and obtains the file by exchanging packets with one another. Their subsequent work [41] extended their previous grid network with clusters and considered hybrid D2D links including microwave links at 2.45 *GHz* carrier frequency and high capacity unlicensed millimeter-wave links at 38 *GHz*. They assumed that only UEs within one cluster can communicate. If a UE finds its requested file in its cluster, it first checks whether the millimeter-wave is available (because it suffers from strong path loss and is easily blocked by obstacles) and obtain the file by millimeter-wave links, and otherwise by microwave D2D links. If the requested file cannot be obtained by D2D communications, the BS will transmit the file by downlinks at 2.1 *GHz* carrier frequency.

One challenge of D2D caching is attributed to the relatively small storage capabilities and limited batteries of UEs that may degrade the caching performance. Sheng *et al.* [42] proposed a multilayer caching and delivery architecture consisting of both SBSs and UEs. They utilized multihop D2D communications to help reduce the energy consumption and battery life of UEs. Multihop D2D communications can reduce the transmission coverage for each hop and hence the transmission power of UEs, which results in decreasing the UE energy consumption and battery life. In order to improve the efficiency of caching storage usage, Ji *et al.* [43] investigated the coded caching scheme in D2D wireless networks. Different from uncoded

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2019.2908280, IEEE Communications Surveys & Tutorials

caching, coded caching does not require to store the whole complete file from the file library, thus provisioning UEs with more flexility and storage efficiency. Moreover, Zhang *et al.* [44] studied cooperative D2D caching. In their system model, if two UEs both need to cache two files, instead of downloading both two files for each UE, one UE downloads one file and the other UE downloads the other file, and then they share the files with each other. Therefore, cooperative caching can greatly reduce the storage consumption of both UEs.

Another challenge of D2D caching is that D2D transmissions are easily interfered by other D2D pairs within the collaboration distance. Hence, there is a tradeoff between cache hit probability and interference. A smaller collaboration distance introduces less interference and hence increases the frequency reuse and potential throughput. On the contrary, a larger collaboration distance introduces more interference but can increase the probability of finding the requested file cached at nearby UEs. This tradeoff was studied in [45], where a closed-form expression for the optimal collaboration distance was derived. In their work, they tried to maximize the frequency reuse for a given popularity distribution and storage capacity. However, their work is limited because they only consider a fixed number of uniformly distributed UEs in the grid-based D2D wireless network. Altieri *et al.* [46] proposed a stochastic geographic model to maximize the number of requests served by D2D caching. UEs are distributed as a Poisson point process (PPP) and can exchange contents through D2D links. In their model, interference can be avoided by the time division multiple access (TDMA) scheme, in which time is divided into equally-length time slots and only one request is served in each time slot.

Owing to the characteristic of D2D communications, the social relationships have a great impact on the D2D caching problem, especially on the content delivery routing path selection. Social characteristics determine the user mobility pattern and willingness of sharing resources, and hence they can help predict future requests. Wang *et al.* [47] proposed a traffic offloading framework for a 5G system by exploiting user social relationships by assuming that UEs can cache contents according to their own preferences or the group's demands. Their framework can measure and analyze user access patterns and delays, and then disseminate contents of interest. They concluded that user social behaviors have several properties: 1) a small number of users can significantly impact other users; 2) clusters of users, where they transfer and share contents, are usually formed by interests; 3) users, who are geographically close, have higher trends of sharing information. Wu *et al.* [22] proposed a two-layer cache-enabled social network model including the social network layer and the physical network layer. In the social network layer, the network links denote the social relationships between UEs. The network links in the physical network layer reflect the physical connections of network infrastructures (e.g., BS and UEs). Furthermore, the social network layer is divided into two sublayers including online sublayer (interest intimacy) and offline sublayer (user mobility). Wu *et al.* [48] then formulated a submodular function maximization problem to maximize the cache hit ratio in mobile social networks. The user interest similarity matrix and contact probability matrix are defined to characterize the social relationships. A semigradient-based algorithm is designed to obtain the optimal content placement strategy iteratively. They demonstrated that their semigradient-based algorithm provides lower computational complexity and faster convergence than the classical scheme proposed in [49].

B. Caching at BSs

As mobile users require higher throughput and lower network latency, caching at BSs, which brings contents closer to mobile users, becomes a promising solution. Caching at BSs faces the challenges of limited coverage, uncertainty of wireless connections and inter-cell interference (ICI) [50]. Caching at BSs becomes more complicated when considering overlapped densely deployed SBSs and heterogeneous networks.

Heterogeneous networks (HetNets), which augment macro base stations (MBSs) by deploying small base stations (SBSs) with small coverage areas, are introduced to increase the network coverage and throughput [51]. In HetNets, SBSs, e.g., pico base stations (PBSs) and femto base stations (FBSs), are densely deployed. We compare the configurations of different BSs in Table III and discuss caching at different types of BSs in the following subsections.

Table III. Comparison of BSs.

Characteristics	MBS	PBS	FBS
Indoor/Outdoor	Outdoor	Indoor or Outdoor	Indoor
Number of users	200 to 1000+	32 to 100	4 to 16
Maximum output power	40 to 100 W	250 W	20 to 100 mw
Maximum cell radius	10 to 40 km	200 m	10 to 50 m
Bandwidth	60 to 70 MHz	20 MHz	10 MHz

1) Caching at MBSs: Many works have studied the framework of caching at SBSs with the whole file library at the mobile core network. MBSs, as parts of the wireless network entities, can also serve as cache nodes in cache enabled networks. Zaidi *et al.* [52] presented a two-tier cellular network including the MBS tier and SBSs tier with the objective to maximize the cache hit probability. All MBS and SBSs are equipped with cache storages. Users can retrieve the requested contents via four schemes: 1) directly from the serving SBS; 2) directly from the serving MBS; 3) from the MBS via the relay of the serving SBS; 4) from other SBSs via the relay of the MBS. Chang *et al.* [53] proposed an algorithm based on the college admission matching to assign contents to cache entities including both MBSs and SBSs in heterogenous small cell networks. The drawback of this work is that they assumed that each content can only be cached at one cache entity. This assumption will limit the caching performance when users at different geolocations request the same content and hence several users have to suffer from great path loss if the content is only cached at a remote SBS or MBS.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2019.2908280, IEEE Communications Surveys & Tutorials

2) Caching at SBSs: Bringing small BSs closer to the users can potentially support the low transmission power and high data rate. In SBS caching, each SBS is equipped with limited storage for caching, and users can retrieve contents directly from the SBS instead of from the remote servers. This can offload some traffic from the backhaul and alleviate backhaul congestion. Golrezaei *et al.* [54] first proposed caching at SBSs, which only have low-bandwidth backhaul links but are equipped with large storage capacity. They studied the content placement problem at SBSs to minimize the average delay of all mobile users. Blasco *et al.* [55] also studied the content placement problem in the cache-enabled SBS network. Meanwhile, they considered the cache replacement phrase in which the cache content can be refreshed at each time interval according to the varying file popularity. They designed a learning-based algorithm to maximize the system reward, which was defined as the bandwidth alleviation of backhaul links. However, caching and transmission policies are considered separately among these works. Gregori *et al.* [56] jointly optimized caching and delivery strategies to minimize the MBS energy consumption. In their system model, SBSs and UEs are equipped with cache storages. A SBS can serve multiple users simultaneously and UEs can share data through D2D communications. However, these works do not consider the interference between different cells. Khreishah *et al.* [57] proposed a coordinated SBS cellular system where each SBS can use a set of secondary channels to communicate with other SBSs, and the MBS stores all files and can always serve users if a requested content is not found at any of the SBSs. They jointly optimized the channel allocation and the content placement problem to maximize the system throughput. In order to address the channel interference, they introduced a conflict graph and then formed several independent sets where channels do not interfere with each other in one independent set.

Although the content caching and spectrum sharing have been widely exploited recently, most works only consider them separately. Ul-Hassan *et al.* [58] characterized the outage probability in cached-enabled SBS networks, defined as the probability of not satisfying users' requests over a given coverage area, as a function of cache size and SBS density. However, they did not discuss the impact of spectrum on the outage probability. Their following work [59] jointly optimized the spectrum allocation and caching in cache-enabled SBS networks where the SBSs are distributed following a homogeneous PPP. Their objective is to minimize the cache miss probability (defined as the probability of requests not fulfilled in a given area) under the constraint of cache storage capacities.

3) Caching at PBSs: Caching at PBSs can help reduce network backhaul traffic load. Li *et al.* [60] proposed a weighted network traffic offloading problem with cache storage located at PBSs. In their system model, the MBS, acting as a central controller which determines the content placement strategies, is connected with geographically distributed cache-enabled PBSs. The users can obtain the requested contents from the caches at their serving PBSs or from their neighboring PBSs because PBSs are assumed to be able to share contents with each other. Meanwhile, Cui and Jiang [61] jointly optimized the caching and multicasting in a two-tier HetNet including the MBS tier and the PBS tier to maximize the successful transmission probability. The location distributions of MBSs and PBSs follow PPPs with different densities. They considered two caching schemes, i.e., identical caching in the MBS tier and random caching in the PBS tier. Identical caching means that each MBS stores the same set of files while random caching enables PBS to store different files randomly. The intuition of the hybrid caching schemes is that user requests are either served by its nearest MBS or multiple nearby PBSs. Therefore, identical caching in MBS tier does not lose much optimality when the density of MBSs is low and random caching in PBS tier can take advantage of multiple spatially distributed PBSs by multicasting content delivery.

4) Caching at FBSs: An FBS (also called helper node), which usually has low bandwidth backhaul links (e.g., wireless backhaul links) and large storage capacities, is one kind of SBSs. They are more flexible and cost efficient to be deployed than traditional BSs. Caching at FBSs is usually referred to as femtocaching. Golrezei *at el.* [49] presented a femtocaching architecture for video contents dissemination in wireless networks. Femtocaching equips FBSs with cache storages for storing popular video contents. They also demonstrated femtocaching improves the system throughput by one to two orders of magnitude as compared with the architecture without helper nodes. Shanmugam *et al.* [62] discussed the content placement problem in a femtocaching network with the objective to minimize the expected download time of all files. The matching theory was utilized to match UEs to helpers in a bipartite graph formed by UEs and helpers. They further extended their work in [63], which introduced the concept of dynamic femtocaching by considering the user mobility and changing topology.

C. Caching at Relays

Wireless relays are usually deployed to extend the wireless coverages and improve the spectrum efficiency. They can also be used as urban hot spots where contents can be cached [64]. Wang *et al.* [65] proposed the cache-enabled relay cellular network with one BS and multiple relays installed at the cell edge to serve users within their coverages. They provided insights on how and when to cache by a Markov decision process aiming at minimizing the energy consumption of all relays and the BS. In their work, the caches can be refreshed by dropping the least popular contents. However, they assumed that the transmit power at relays is adjustable, which is impractical when relays are serving as hot spots. For a similar network architecture, Erol-Kantarci *et al.* [66] deployed cache-enabled relays as hot spots at random grid points in a cell coverage area. They discussed three problems: 1) relay selection, which minimizes the uplink power of UEs (i.e., the energy consumption for UEs to access the BS); 2) content placement, which minimizes the sum of the uplink power and caching power; 3) relay placement, which jointly minimizes the number of relays and the uplink power. Liu and Lau *et al.* [67] introduced an opportunistic cooperative

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2019.2908280, IEEE Communications Surveys & Tutorials

MIMO (CoMP) framework for wireless video streaming. The cached-enabled relays can cache a portion of the video files to avoid the expensive backhaul traffic cost. Without cache at the relays, the MIMO relay channel is utilized to improve the wireless coverage. Contrarily, the opportunistic CoMP broadcast can be utilized to broadcast the cached video packets at the relays to users. They jointly optimized the cache control and power allocation problem to exploit the tradeoff between CoMP opportunities and relay cache sizes.

D. Joint Caching in HetNets

HetNets helps the wireless networks accommodate the dramatically growing mobile traffic at the expense of imposing a significant challenge on the backhaul links. To reduce the backhaul load, joint caching in HetNet has been regarded as a promising method [61]. Joint caching in HetNets, as illustrated in Fig. 5, consists of caching at MBS, FBS, PBS, relay and UE. Wang *at el.* [68] jointly optimized the user association and content placement problem in a heterogeneous cellular network where a single MBS and multiple SBSs are connected to the mobile core network by optical fiber and bandwidth-limited wired backhaul links, respectively. The interferences between SBSs and the MBS cannot be avoided because they share the MBS downlink resources. They demonstrated that the problem is NP-hard and then designed a distributed algorithm to address the problem. Yang *et al.* [69] proposed and analyzed a three-tier cache-enabled HetNet which is composed of BS layer, relay layer and UE layer. The locations of BSs, relays and UEs are modeled as mutually independent PPPs with different densities. They assumed that all cache-enabled UEs cache the same content and all relays store the same content as well. They then discussed four content access cases: 1) UEs (without cache abilities) obtain the requested contents from the closest cache-enabled UE, relay or BS; 2) UEs (with cache abilities but without requested contents in local caches) obtain requested contents from the closest relay or BS; 3) UEs (without cache abilities while the closest cache-enabled UE does not contain the contents) obtain requested content from the closest relay or BS; 4) UEs (with cache abilities with requested contents in local caches) obtain requested contents from local caches. They considered both inter-tier and intra-tier interference because both the communication links connecting relays to UEs and D2D links between UEs share the frequency resources with those between BSs and UEs. In their work, the outage probability was analyzed to improve the user quality of service (QoS).

E. Caching in C-RAN

C-RAN is proposed as a novel architecture for 5G cellular networks to reduce the CAPEX (capital expenditure) and OPEX (operating expenditure) by adopting cloud computing technology [36]. C-RAN addresses the capacity and coverage issues by deploying multiple RRHs at cell sites. The computational functionalities of BSs are centralized in a common cloud processing unit, i.e., BBU pool. These RRHs serve as distributed antennas to interact with various users. RRHs and BBUs are connected by high bandwidth and low

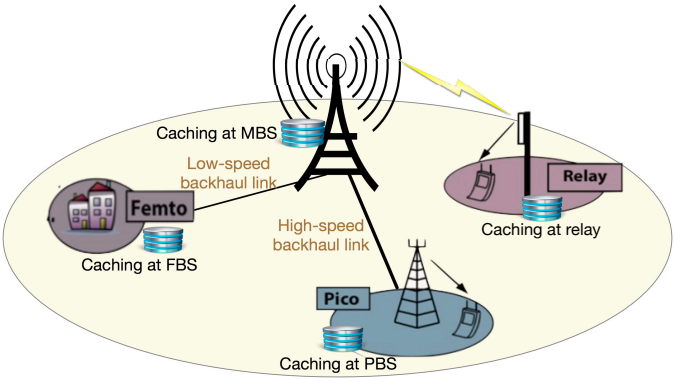


Fig. 5. Joint caching in HetNets (BSs and relays)

latency fronthaul links [70]. Although C-RAN can provide strong computing abilities by sharing computing and storage resources at the BBU pool, it still suffers from the performance limitation owing to the constrained capacities of fronthaul and backhaul links [35]. To overcome this shortage, caching techniques have been considered as an effective method to alleviate the network traffic in both fronthaul and backhaul links. In C-RAN architecture, cache storages can be deployed at both BBU or RRH level (illustrated in Fig. 6). The C-RAN architecture utilizing distributed edge caching techniques is also referred to as the fog radio access network (F-RAN) [71].

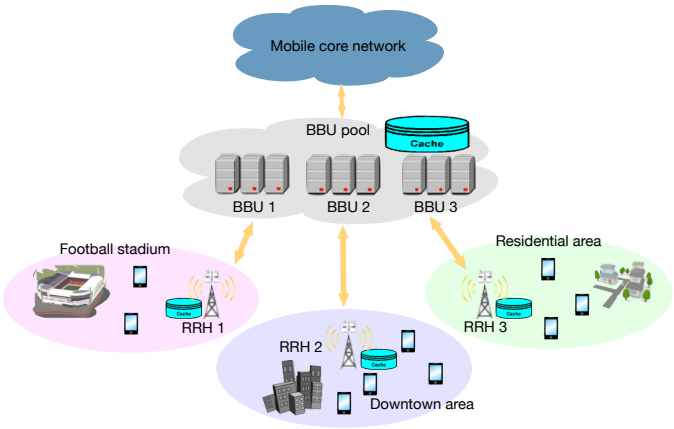


Fig. 6. Caching in C-RAN.

Effective caching strategies for F-RAN are discussed in [72]. In their F-RAN architecture, all RRHs are equipped with cache storages and the BBU pool has the whole file library. By considering the limited fronthaul capacity, cache capacities and wireless resources constraints, they discussed the content placement problem in RRH caches and content delivery problem in the downlink wireless channels. In a similar network architecture, Tao *at el.* [73] studied a content transmission design by integrating both multicasting and caching. In their work, RRHs can form multiple RRH clusters and each UE is served by all RRHs in one cluster. Meanwhile, UEs, with similar content preferences, can act as a multicast group in which UEs can receive data concurrently by multicast transmission. They addressed the problem of how to form RRH clusters and UE multicast groups with the objective to minimize the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2019.2908280, IEEE Communications Surveys & Tutorials

system cost incurred from fronthaul traffic cost and RRH energy consumption. Mosleh *et al.* [74] jointly optimized the content placement and cooperative transmit beamforming to minimize the system cost (i.e., fronthaul cost and transmission power) under the constraints of QoS, peak transmission power and cache capacity. They further separated this joint problem into two subproblems including the content placement problem and beamforming design problem, and then designed heuristic algorithms for both subproblems. Stephen and Zhang [75] tried to optimize the energy efficient transmission in C-RAN which uses the orthogonal frequency division multiple access (OFDMA) scheme. The constraints in their problem include the fronthaul capacity and required minimum user data rates.

The above works only consider caching at the RRH level, and there is no collaboration among these caches. Tran and Pompili [76] proposed a novel caching framework, Octopus, which equips RRHs with distributed edge-caches at RRHs and BBUs with cloud cache. Their overall system aims to provision optimized caching at multiple layers such that the total content access delay is minimized. Yao and Ansari [77] addressed the content placement problem in C-RAN at both the RRH and BBU level with the objective to minimize the average file download latency. Different from most works on C-RAN caching, multiple BBU pools are included in their network system. They further investigated the joint optimization of content placement and storage allocation problem in [78].

Most existing works in the context of caching in C-RAN neglect the effect of user's mobility and social relationships. With the similar architecture of joint caching at both BBU and RRH level, Chen *et al.* [79] addressed the content placement problem to minimize the network delay in C-RAN and introduced a framework of echo state networks to predict the user request and mobility patterns with the aid of machine learning technologies. By extracting the information (e.g., age, job and location) from user content requests, the echo state network can track the current network status and predict the future information. From the perspective of users' social relationships, Wang *et al.* [80] discussed the impact of mobile social networks on the performance of edge caching schemes in F-RANs. They aimed at minimizing the bandwidth consumptions of both fronthaul links and RANs by caching at RRHs and UEs. UEs can share contents according to their social ties and behaviors.

F. Summary and Discussion

In this section, we classify the existing research according to the caching locations at UEs, MBSs, SBSs, PBSs, FBSs and relays in traditional cellular networks, and RRHs and BBU pools in C-RAN [81].

1) Caching at UEs: Caching at UEs is also referred to as D2D caching. D2D networks allow direct communications between UEs using license-band (e.g., LTE) or unlicensed-band protocols (e.g., bluetooth and WiFi). The devices are often organized into clusters, which are controlled by the BS. A user's content request can only be satisfied by other users within the same cluster. The physical locations of users play an important role in designing D2D caching strategies, so that a user can easily find its requested content at one of the neighboring users. Designing caching strategies for D2D networks faces four major challenges:

- Each user may find its requested content at multiple neighboring users and each content may also be requested by multiple neighboring users. How to establish the D2D communication links to transmit content in order to maximize the active D2D links is very complicated and challenging.
- UEs usually have limited storage and battery capacity, which may degrade the caching performance. Multihop D2D communications can be adopted to reduce the energy consumption of each UE.
- The D2D transmission is easily interfered by those from different neighboring UEs within the collaboration distance. Larger collaboration distance introduces more interference while increasing the probability of finding the requested file; this leads to a tradeoff between the interference and content finding probability. An optimal collaboration distance should be explored to address this tradeoff.
- User social relationship plays an important role in D2D caching because it determines the user willingness to share contents. How to exploit the user social relationships is also a challenging issue.

2) Caching at BSs and Relays: Caching at BSs can reduce backhaul traffic by installing caching storages at MBSs, SBSs, PBSs and FBSs, because users can obtain the requested contents directly from the BSs rather than from the remote Internet content servers. As MBSs usually have larger cache storages and coverage areas than SBSs, the objective of caching at MBSs usually involves minimizing backhaul traffic, network latency and successful transmission probability. However, the caching storages of SBSs are relatively small. Hence, cooperative caching (i.e., different SBSs can share contents with each other) is usually adopted. In addition, in order to make the most use of the limited caching storages of SBSs, cached contents should be updated more frequently according to the dynamic content popularities. PBSs and FBSs are special SBSs. PBSs require high-speed backhaul links connected to the MBS that may incur high traffic cost. Caching at PBSs reduces the traffic through the backhaul links and hence reduces the traffic cost. In contrast, FBSs usually have smaller backhaul links than those of PBSs, and may even have wireless backhaul links. Hence, FBSs are flexible and cost efficient to be deployed but within smaller coverages. Relays are usually deployed to extend the wireless coverages or act as urban hot spots. When designing caching schemes for relays, the energy consumption of relays along with their locations should be taken into consideration.

HetNet improves area spectral efficiency by densely deploying SBSs. It consists of MBSs, SBSs (e.g., FBSs and PBSs), and relay nodes, where the coverage of macrocells overlaps with those of the small cells and relays and so they can share the same spectrum. Hence, the interferences between SBSs and the MBS cannot be avoided. In a typical cache-enabled HetNet model, the MBS can act as the central controller

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2019.2908280, IEEE Communications Surveys & Tutorials

to determine the caching strategies. The users can obtain the requested contents from their serving SBSs (or relays) or their neighboring ones if cooperative caching is adopted, in which different SBSs and relays can share contents with each other. In the HetNet, users may frequently pass through different small cells. Hence, user mobility should be taken into consideration in designing caching schemes.

3) Caching in C-RAN: C-RAN aggregates the computing capabilities to the BBU pools by adopting cloud computing technology, which introduces flexibility and agility for wireless networks. The densely deployed RRHs at cell sites help improve the network capacity and coverage. The fronthaul links, connecting RRHs to the BBU pools, may be congested due to the growing mobile traffic. Caching at RRHs can reduce the traffic in the fronthaul links and also the backhaul links which connect BBU pools to the mobile core network. Caching at BBU pools help reduce the traffic in the backhaul links. The caching strategies in C-RANs are usually designed to minimize the traffic in fronthaul and backhaul links, and the energy consumption of RRHs. In C-RANs, users are usually served by a group of RRHs by using the coordinated multipoint transmission technique to increase the network capacity. Hence, how to form groups of RRHs to serve users, which is referred to as the RRH clustering problem, is critical. Owing to the coordinated multi-point transmission technique, each RRH is equipped with multiple antennas and several RRHs in a multicast group can cooperatively transmit contents to users using multicast beamforming techniques. Hence, how to design the beamforming vectors to minimize the backhaul cost is a critical issue.

IV. Caching Criteria

When designing a caching scheme, we should try to improve the caching performance in terms of several caching criteria. We summarize these criteria in Table IV and discuss them in this section.

Table IV. Summary of caching criteria.

Criteria	Definition	Works
Cache hit probability	Ratio of the number of cached files requested by the users over the total number of files in the caches.	[82]–[85]
Spectrum efficiency	Supported data rate that can be transmitted over a given frequency bandwidth	[86]
Energy efficiency	Supported data rate that can be transmitted over per unit energy	[50], [87], [88]
Network throughput	Maximum rate the network can provide	[57], [89], [90]
Content retrieving delay	Round-trip-time of obtaining contents by users	[91]–[93]
Offloaded traffic	The traffic difference in backhaul links between uncached systems and cached systems	[60], [94]

A. Cache Hit Probability

The cache hit probability refers to the ratio of the number of cached files requested by the users over the total number of files in the caches. A higher cache hit probability means that more user requests are satisfied by the cached contents. Increasing the cache size can improve the cache hit probability, and hence lower the required backhaul capacity. Therefore, there is a tradeoff between the cache size and the required backhaul capacity. Pantisano *et al.* [82] proposed a a collaborative framework, where SBSs can form a collation and share contents with each other in order to improve the cache hit probability. They discussed the content placement problem by designing a decentralized algorithm. By further considering the scarcity of bandwidth resource, the D2D communications technique is introduced in the caching strategy design [83] to improve the cache hit probability. Chen and Kountouris [84] compared the performances of D2D caching and SBS caching in terms of cache hit probability. They concluded that D2D caching performs better when user density is high because more user requests can be served simultaneously through short-distance D2D communications. Otherwise, SBS caching is more beneficial because cache storages of SBSs are usually larger and hence the cache hit probability can be improved. Blaszczyszyn and Giovanidis [85] studied the content placement policy in cellular networks with the objective to maximize the cache hit ratio. In their network system model, a user can be covered by multiple cache-enabled BSs and connect to any of the BSs. They considered several BS coverage models. The first model is the signal-to-interference plus noise ratio (SINR) based model where a user can only connect to the BS with the received SINR larger than a predefined threshold. The second model defines that each BS has a coverage radius and can only connect to the users within its radius. In the third model, when multiple network operators (e.g., 3G/4G BS and WiFi hotspots) coexist over an area, the user connects to different operators following predefined probabilities.

B. Spectrum Efficiency

SE is referred to as the supported data rate over a given frequency bandwidth. To improve the area SE, network densification by deploying more SBSs in a macro cell has been utilized. Caching can also improve SE by reducing network traffic and improve network throughput. Liu *et al.* [86] studied the performances of the area spectral efficiency in a two-tier HetNet including helper nodes (with high capacity caches but without backhaul connections) and PBSs (without caches but with limited backhaul links). The HetNet consists of the MBS tier and the helper node/PBS tier, where the MBS can serve multiple users at each time slot while both the PBS and helper node can only associate with one user. The essential difference between a PBS and a helper node is that the downlink transmission rate for the PBS can be limited by the PBS's weak backhaul links while it only depends on the wireless channel between the helper node and the user. They assumed that each helper caches the most popularity files until it reaches its cache storage capacity and then derived the closed forms of area spectrum efficiency for the two kinds of HetNets. They concluded that 1) deploying PBSs and helper nodes can both double the spectrum efficiency as compared with only MBS; 2) deploying helper nodes achieves more spectrum

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2019.2908280, IEEE Communications Surveys & Tutorials

efficiency improvement and requires less cost of deployment and management as compared with PBSs; 3) increasing cache capacities of helper nodes can achieve comparable spectrum efficiency improvement as deploying more helper nodes.

C. Energy Efficiency

EE, which is defined as the supported data rate over a given energy consumption, is an important performance metric for 5G cellular networks. In traditional cellular networks, energy can be saved by turning BSs into the sleep mode when there is no traffic load. The energy consumption can also be decreased by caching at mobile edges because caching helps eliminate duplicated transmissions. Based on the Shannon equation $C = B\log_2(1 + \frac{P_t G}{I+N_0})$, where P_t, G, I, N_0 indicate the transmission power, channel gain, interference and noise power, respectively, we can deduce that if interference I is reduced, transmission power P_t can decrease accordingly to maintain a certain wireless capacity C. The energy efficiency, which equals to $\frac{C}{P_t T}$, will increase. Hence, controlling the interference can help improve the EE. Liu and Yang [50] studied the impact of interference on EE and explored the EE gain brought by caching in downlink cellular networks. The energy consumption for a BS is incurred by both the BS transmission and the backhaul traffic. They derived the closed-form expression of EE and then concluded that the EE is higher when caching is enabled because it helps reduce the backhaul traffic which contributes to the energy consumption. They explained that EE gain from caching can be much higher if the interference is well controlled because well controlled interference helps increase the wireless throughput. They also demonstrated that caching at PBSs increases EE gain more than caching at MBSs because the weak backhaul capacity limits the throughput of PBS while caching at PBSs helps alleviate the traffic in weak backhaul links and hence the throughput can be improved. Perabathini *et al.* [87] discussed how to optimize the BS transmit power to minimize the energy efficiency in the cache-enabled cellular networks, where they modeled locations of BSs and users as PPPs with different densities. They imposed the QoS constraint which is characterized as the BS coverage probability. They also assumed that the content placement strategy adopts the global popularity strategy where each BS caches the most popular contents up to its cache storage.

EE is also critical for UEs because smart phones today are usually battery constrained. AlMonmani *et al.* [88] proposed a heuristic scheme for the overlay small cell content-caching network to determine the SBS to be connected with each UE so that all UEs' energy consumption can be minimized. Their proposed scheme is based on particle swarm optimization (PSO), where a number of particles are utilized to iteratively calculate the optimal solution. In the first iteration, each particle calculates its optimal solution (i.e., local optimal solution) in its search space, and then the global optimal solution is the maximum (or minimum) local optimal solution among all particles. In the next iterations, all particles move to other positions to attain another local optimal solutions, and the global optimal solution can be obtained by taking the maximum (or minimum) local optimal solution. After several iterations, the overall optimal solution of the problem is determined by the maximum (or minimum) of all global optimal solutions among all iterations.

D. Network Throughput

The network throughput is defined as the maximum data rate the network can provide and affects the network performance. A larger network throughput leads to lower content download delay. Caching schemes should try to maximize the network throughput by caching more data closer to the users. The network throughput is also highly related to the caching storage capacities. Yang *et al.* [89] modeled and evaluated the performance of coexistence of RAN caching and D2D caching aiming at maximizing the network throughput. We demonstrated that the network throughput can be increased by 57% compared with traditional non-caching methods. Khreishah *et al.* [57] considered joint caching, routing, and channel assignment problem for the video delivery over coordinated small-cell cellular systems with the objective to maximize the throughput of the system. An approximation algorithm based on column generation was designed to solve this problem. Nguyen *et al.* [90] investigated the joint optimization problem of beamforming and power allocation in a cache-enabled wireless small cell HetNet to maximize the small cell throughput.

E. Content Retrieving Delay

The content retrieving delay, often defined as the round-trip-time for obtaining contents by users, is directly related to the user QoS [95]. It usually consists of the wireless transmission delay from BSs to UEs and the backhaul delay from BSs to the mobile core network. The wireless transmission delay depends on the bandwidth and SINR, while the backhaul delay depends on the link length, traffic load and the number of BSs connected to the mobile core network. Users may fetch contents from either local UEs, BSs or the mobile core network with different delays. Hsu and Chen [91] combined the content placement problem with the wireless radio spectrum allocation problem to minimize the network latency in RANs. In their system model, there are several UEs and a BS, both of which are cache-enabled. Users can acquire contents from the BS by reliable connection or from other UEs by establishing D2D connections. A fraction of the total available license spectrum is assigned for D2D communications while the other fraction is used for cellular downlink. They transformed the objective of minimizing the network latency into minimizing the backhaul and downlink traffic. They then developed a branch and bound algorithm to find the optimum spectrum allocation and caching distribution. Without a central coordinator, Liu *et al.* [92] proposed a distributed algorithm to address the joint content placement and transmission problem to minimize the average download delay. In their system model, BSs are equipped with cache storages. Upon receiving a user request, the serving BS decides to either transmit the requested content directly, or cooperate with other BSs and transmit the content to the user with cooperative beamforming. If none of the BSs cache the file, the serving BS will retrieve the content from the content

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2019.2908280, IEEE Communications Surveys & Tutorials

server via backhaul links and then transmit the content to the user. They assumed that the channel gains are identically and independently distributed over the time slots. Hence, the content delivery rate, which is related to the channel gain, is a random variable. In each time slot, they chose the transmission scheme with less transmission delay between direct transmission scheme and cooperative beamforming transmission scheme. They then calculated the expectation of delays from all time slots as the average download delay.

Most works do not take the request forwarding problem into consideration. Dehghan *et al.* [93] jointly optimized the content placement and user request forwarding problem to determine which content should be cached at each cache and how user requests are forwarded to minimize the average content access delay. In their network model, there are several caches and an one-hop backend server which can always serve the users. User's requests can be forwarded either to the backend server by costly, congested uncached paths or to the caches by cheaper and faster cached paths. However, if the cache, to which a user routed, does not cache the requested content, it will first download the content from the backend server and then transmit the content to the user, resulting in additional content access delay. They discussed two delay models: 1) the congestion-insensitive model in which the delay is independent of the traffic load and can be considered as a constant; 2) the congestion-sensitive delay model in which the delay is related to the traffic congestions and the paths are modeled as M/M/1 queues with fixed request arrival rates and service rates. They also demonstrated that for both delay models, the joint optimization problem is NP-complete and hence cannot be solved in polynomial time and approximation algorithms were designed to address the problem.

F. Traffic Offloading

Mobile edge caching can offload the traffic from backhaul links. Maximizing traffic offloading leads to better caching performance. Li *et al.* [60] tried to maximize the traffic offloading by caching in HetNets. They formulated their problem as a minimization of expected sum of traffic loads subject to cache storage capacities and backhaul link limitations. The traffic loads are incurred between PBS and user, PBS and PBS, MBS and user, as well as mobile core network and MBS. A suboptimal greedy algorithm was designed to obtain the optimum content placement decisions. By incorporating user mobility, Rao *et al.* [94] explored the content placement problem in a two-tier wireless caching network to maximize the traffic offloading probability. The locations of cache-enabled helpers and users are modeled as two independent PPPs. The users can acquire contents from helpers, other UEs or MBS. Helpers can transmit data to users within their coverage ranges and only UEs within a certain distance can establish D2D connections. They derived the expressions of the cache hit ratios at helpers and UEs as P_h and P_u, respectively. Then, the traffic offloading probability can be calculated as $1-(1-P_h)(1-P_u)$.

G. Summary and Discussion

In this section, we classify the existing research according to different caching criteria including cache hit probability, spectrum efficiency, energy efficiency, network throughput, content retrieving delay and offloaded traffic.

The cache hit probability reflects the percentage of cached files that are used and is an important metric to evaluate the performance of caching placement algorithms. Hence, most existing works try to design a caching placement policy, which determines the content to be cached at different caches, to maximize the cache hit probability. Another important factor that affects the cache hit probability is the caching storage size. A larger cache size implies that more contents can be cached, hence increasing the cache hit probability. How to allocate caching storages to different locations is still under investigation. Intuitively, the locations with more users should be allocated with a larger cache size. Furthermore, the cache hit probability can also be increased by cooperative caching, which allows multiple caches share contents with each other. For example, different BSs can cache different contents and share with each other to serve users, thus improving the cache hit probability.

SE reflects how much network capacity can be provided by a unit of spectrum resource. The area spectrum efficiency (ASE) is also measured to indicate how many users in a certain area can be accommodated. In wireless networks, HetNet is usually deployed to improve spectrum efficiency by deploying several SBSs in each macro cell. Hence, most caching policies, aiming at maximizing the SE, is in the context of HetNet. As the MBS and SBSs share spectrum resources, how to allocate the spectrum resources is a challenging issue. Caching helps increase the network capacity to accommodate more users, thus increasing the SE. On the other hand, increasing the density of SBSs also helps improve the SE. Therefore, the SBS density can be traded off by increasing the cache size to achieve a targeted SE. In addition, D2D caching with unlicensed-bands increases the SE by establishing direct communications among UEs instead of obtaining contents from BSs via downlinks. In that case, maximizing the SE is equivalent to maximizing the number of D2D communication links; this has been widely investigated in D2D networks.

EE reflects the supported data rate per given energy consumption. The energy consumption of a BS is attributed to both the BS transmission and the backhaul traffic. Caching at BSs helps reduce the backhaul traffic and hence reduces the energy consumption. The energy consumption of BS downlink transmission can be reduced by optimizing the BS transmission power. Note that, in HetNet, caching at SBSs can achieve higher EE than caching at MBSs because the limited backhaul capacity of SBSs (i.e., links between MBSs and SBSs) may limit the network throughput when caching at MBSs and may hence limit the EE. However, caching at SBSs suffers from the limited cache storages of SBSs. Therefore, the content placement problem, which places the contents at both SBSs and MBSs should be well designed, is a crucial issue to be investigated in order to maximize the EE. Caching policies should also consider the impact of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2019.2908280, IEEE Communications Surveys & Tutorials

interference because too many interferences may degrade the network throughput and hence reduce the EE. In addition, the EE is very critical for D2D caching because UEs are usually battery constrained. How to minimize the energy consumption of UEs while satisfying all user's content requests is still an ongoing research.

The network throughput reflects the maximum data rate that the network provides. It affects the user QoS especially in the video streaming application, where users usually require the minimum network throughput. In order to improve the network throughput, the contents should be cached as close to the users as possible. Otherwise, any hop of the multiple network hops between the content and users could be the bottleneck and limit the network throughput. The network throughput is also highly related to the caching storage size. A larger cache size implies more contents can be cached and hence higher network throughput. The cache storage allocation problem to maximize the network throughput is still under investigation.

Content retrieval delay directly affects the user QoS. Most delay-sensitive services (e.g., disaster response and virtual reality) usually define strict deadlines for content retrievals. In wireless networks, content retrieval delay usually considers the downlink delay (i.e., the delay between the caches or content servers and users) and neglect the uplink delay (i.e., the duration of forwarding the requests). The one-hop network delay usually consists of three parts including the data transmission delay, queueing delay and propagation delay. Most works only consider the data transmission delay. The content retrieval paths may include several network hops, which depend on where the content is located. If the content is obtained from the mobile core network, the content retrieval delay includes the backhaul delay and the wireless transmission delay. The backhaul delay is depends on the link length, backhaul capacity and traffic load while the wireless transmission delay depends on the wireless channel conditions, bandwidth, SINR and interferences. For cooperative caching, the content transmission latency between two cache nodes should also be considered. The content retrieval delay is affected by both the content placement strategy (which determines where the contents to be placed) and the content delivery strategy (which decides how the contents are delivered). Hence, minimizing the content retrieval delay is usually the objective of the content placement and delivery problems.

Offloaded traffic reflects how much traffic can be reduced by caching in the network. The traffic in the backhaul links is usually measured to evaluate the offloaded traffic. More offloaded traffic implies better caching performance. Offloaded traffic is highly related to the cache storage sizes. A larger cache size implies more offloaded traffic. The offloaded traffic is mainly evaluated in two ways. The first way is to transform the offloaded traffic maximization problem into the traffic load minimization problem. The second way is to evaluate the traffic offloading probability which is defined as the number of requests served by one of the caches over the total number of requests. Hence, a higher cache hit probability usually leads to a higher traffic offloading probability.

V. Caching Schemes

We will next discuss different caching schemes, delineating their advantages and disadvantages as summarized in Table V.

Table V. Comparison of caching schemes.

Schemes	Features	Advantages	Disadvantages
Proactive caching [32], [49], [96]–[99]	Cache contents before request based on predicted patterns	Alleviate peak-hour traffic; reduce network latency	Caching performance relies on prediction accuracy
Distributed caching [49], [92], [100]–[102]	Cache nodes make decisions only with information from local and few adjacent nodes	Reduce burden on the central controller; save signaling overhead without central coordinator	May not get the global optimal solutions
Cooperative caching [99], [103], [104]	Caches share contents with each other	high utilization of caches	Induce signal overhead by sharing caching status
Coded caching [43], [105]–[109]	Merge several blocks of different files to the same destination using network coding techniques	Improve network efficiency, throughput; lower scheduling complexity	Requires more processing at intermediary and terminal nodes
Probabilistic caching [43], [85]	Content is cached with a certain probability	Adapt to the uncertain network status	Hard to obtain the optimal solutions
Game theory based caching [12], [110]–[113]	Analyze the competition and cooperation among MNO, SP and user	Cope with the competition and cooperation among different parties	Hard to characterize selfishness and willingness to cooperate

rephrase for chapter 2 state of the art

A. Proactive Caching

The reactive caching policy determines whether to cache a particular content after it has been requested. It typically happens when the network is at peak-traffic hour and cannot effectively cope with the peak traffic. On the other hand, a proactive caching policy determines which contents should be cached before they are requested based on the prediction of user demands [96]. Proactive caching usually utilizes the estimations of request patterns (e.g., user mobility patterns, user preferences and social relationships) to improve caching performance and guarantee QoS requirements. As machine learning and big data analytics advance, it is advantageous to cache popular contents locally before the requests truly arrive [32], [49]. Proactive caching improves the caching efficiency by pre-downloading popular contents during off-peak times and serving predictable peak-hour demands. Bastug *et al.* [32] proposed a proactive networking paradigm which leverages social networks and content popularity distributions to improve the caching performance in terms of the number of satisfied requests and the offloaded traffic. They demonstrated that proactive caching performs better than reactive

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2019.2908280, IEEE Communications Surveys & Tutorials

caching. Tadrous *et al.* [97] considered the system in which the popularity of services can be predicted. Cache nodes can proactively cache services during off-peak hours according to their popularities. They explored the proactive caching scheme by considering the resource allocation to maximize the cost reduction which is related to the offloaded traffic incurred by proactive caching.

To further improve the performance of proactive caching, it is desirable to jointly optimize the caches among multiple nodes. Hou *et al.* [98] exploited a learning-based approach for proactive caching to maximize the cache hit ratio. In their system model, different caches can share information and contents. They first estimated the content popularity by a learning method and then designed a greedy algorithm to obtain the suboptimal content distribution solutions. However, the caching performance highly depending on the prediction accuracy is the major drawback of proactive caching. Prediction errors can gravely degrade the caching performance [99].

B. Distributed Caching

Centralized caching uses a central controller, which possesses a global view of all network status, to determine caching schemes. The central controller usually tracks the information of user mobility patterns and the channel state information (CSI) by extracting and analyzing the received requests. Hence, the centralized caching is able to achieve the optimum caching performance with optimum caching decisions (e.g., content placement). However, obtaining full network information is challenging especially in the context of dynamic 5G wireless networks, which are expected to serve an increasing number of mobile users [100]. Furthermore, the central controller has to process a large amount of traffic, which incurs a great burden on the controller as well the links between the controller and network entities. In that case, the central controller can be the bottleneck of the mobile caching system. In distributed caching, which is also referred to as decentralized caching, cache nodes make decisions (e.g., content placement and update) only based on their local information and the information from adjacent nodes. Distributed caching is applied in [49] where adjacent BSs are jointly optimized to increase the cache hit probability. By fetching contents from multiple neighboring caches, the total cache size seen from the user can be increased.

The believe propagation (BP) method has been proposed as an efficient way to distributively solve the resource allocation problems in wireless networks. In BP, the complex global optimization problem is usually decomposed into multiple subproblems, which can be effectively addressed in the parallel and distributed manner. A tutorial of BP can be found in [101]. Li *et al.* [102] discussed the file placement problem to minimize the average file downloading delay. Their network architecture consists of a MBS and cache-enabled SBSs to which UEs' requests are preferentially forwarded. They divided the files in the file library into several file groups and assumed that each SBS can only cache one file group. A distributed BP algorithm was proposed with the aid of a factor graph, which is a bipartite graph consisting of factor nodes and variable nodes. A factor node refers to the utility function of a user and is related to the average file download delay. Each variable node indicates a cache status vector of each SBS. Only if a UE is under the coverage of a SBS, there can be an edge connecting the UE (factor node) to the SBS (variable node). The BP algorithm is then implemented by iteratively passing messages between the factor nodes and variable nodes. In each iteration, the message is represented by a probability mass function based on the UE's utility function; each variable node updates its message to be sent to connected factor nodes and each factor node updates its message to be sent to connected variable nodes. The BP algorithm terminates when the messages do not change. Different from [102] which assumed that each UE can only served by one BS, Liu *et al.* [92] proposed a distributed BP algorithm to minimize the average download delay in cellular networks where each user can be served by multiple cache-enabled BSs. The data transmission scheme depends on the cache placement. If only one BS caches the requested file, the BS will transmit the file to the user directly; otherwise, multiple BSs can transmit the file via cooperative beamforming. In their BP model, each BS iteratively collects local information (e.g., user requests and CSIs), runs computations, and exchanges messages with the neighboring BSs until convergence. They demonstrated that the distributed BP algorithm requires less calculations than the centralized one.

C. Cooperative Caching

Since the caching space in a BS is relatively small, designing a caching policy for each BS independently may result in an insufficient utilization of caches. This happens when some of the caches are overly used while others have many vacant spaces. In order to address this issue, cooperative caching policies have been proposed to improve the caching efficiency. In the cooperative caching, BSs are able to share cached contents with each other [99]. However, the delay of searching and retrieving contents from other caches may also be significant and hence should be taken into consideration. In order to actualize cooperative caching, network nodes should be aware of the caching status of other nodes by information exchanges that may induce significant signaling overheads. Hence, we need to find a solution to share the caching status with the minimum overhead. Jiang *et al.* [103] developed the cooperative caching policy for HetNets where users can fetch contents from FBSs, D2D communications or MBS. They formulated the cooperative content placement and delivery problem as an integer linear programming (ILP) problem to minimize the average downloading latency. A Lagrangian relaxation algorithm was then designed to decouple the original problem into two smaller subproblems which can be solved more efficiently. Additionally, the content delivery problem was also formulated and solved by the Hungarian algorithm.

Most researches on cooperative caching assume the static popularity; the joint consideration of the cooperation and learning of the time-varying popularity still requires further investigation. Song *at el.* [104] explored the content caching problem with an unknown popularity distribution. They incorporated the learning of the popularity distribution, and then

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2019.2908280, IEEE Communications Surveys & Tutorials

jointly optimized the content caching, content sharing and cost of content retrieving.

D. Coded Caching

In a traditional switching network, the network node forwards packets one after another: two packets are present in the node at the same time; one of the two packets is forwarded while the other one is queued even if both are headed for the same destination. This traditional packet forwarding mechanism requires separate transmissions and hence decreases the network efficiency. Network coding is a technique which merges two separate messages into one coded message and forwards them to the destination. After receiving the coded message, the network node separates them into two original messages. To enable the network coding technique, transmitted data are encoded at network nodes and then decoded at the destinations. Hence, the network coding technique requires fewer transmissions to transmit all the data. However, this scheme requires coding and decoding processes, and hence incurs more processing overheads to the network nodes. The complexity of network coding can be lowered by efficient packet transmissions [105].

In network coded caching, files in the file library are usually divided into coded packets and then any linear combination of these code packets can reconstruct the entire original object [43]. For example, the file library has the file C, which is divided into $C_1 \oplus C_2$. Owing to the cache storage limitation, a user, who requests file C for the first time, only caches packet C_1 after having received file C. When the user requests the same file C for the second time, the BS only needs to transmit C_2 to the user. On the contrary, in uncoded caching, file C has to be transmitted for both the first and second time. Therefore, coded caching helps reduce network traffic ($C + C_2 < C + C$). Maddah-Ali and Niesen [106] jointly optimized the caching and coded multicast delivery and demonstrated that the joint optimization problem can improve the caching gain when the demand for the cached content is uniformly distributed. They further showed the near-optimal performance of coded caching achieved by a random caching scheme [107]. They also presented that caching gain can be exploited from coded multicast transmissions in [108]. They proposed in [108] a decentralized coded caching scheme and discussed how to handle scenarios with asynchronous user demands, nonuniform content popularity, and online cache updating. Most works only consider the single layer coded caching, Karamchandani *et al.* [109] proposed a hierarchical coded caching scheme by considering a two-layer hierarchical cache. They first utilized the coded caching schemes in each layer and then combined the two layers by providing coded multicasting opportunities across different layers.

E. Probabilistic Caching

Different from wired networks with fixed and known topologies, wireless networks face the uncertainty about which user will connect to which BS due to undetermined user locations and the variance of user requests. Caching in wireless networks becomes more complex when a user moves from one cell to another during the content delivery. An approach to solve this problem is to employ a probabilistic caching policy in which the content can be placed in the caches according to some random distributions. To reflect the uncertainty, Blaszczyszyn and Giovanidis [85] modeled the user locations as a spatial random process. They optimized the probability of each content being cached at each BS with the aim to maximize the cache hit probability. They also demonstrated that the widely used greedy algorithm, which caches the most popular files, cannot always guarantee optimization in a general network unless no BS coverage overlaps exist. Ji *et al.* [43] discussed the random caching strategy with the aid of coded multicasting in D2D networks where UEs are uniformly distributed in a grid network and can share contents with each other. They pointed out that the drawback of deterministic caching is that the optimal cache placement cannot always be implemented without errors especially when D2D caching is considered. They demonstrated that their random caching strategy, where users make arbitrary requests for files, performs better as the network size grows.

F. Game Theory based Caching

In wireless networks, multiple parties coexist, including the service providers (SPs) who provide contents, mobile network operators (MNOs) who manage the radio access networks (RANs), and mobile users who consume different contents. When applying a specific caching strategy, the benefits of different parties could conflict with each other. For example, bringing more contents to BSs is beneficial to users while increasing the cost of MNOs due to the additional storages and power consumption. Since each party only cares about its own profit, competitions among them are unavoidable. To effectively cope with the competition and guarantee high overall user experience, game theory is adopted to analyze the interactions among these parties. An auction game is suitable to characterize the competition among SPs. In this setting, the cache storages are considered as objects to be auctioned and the price should be paid to MNOs by SPs. MNO should be the one who is in charge of the auction process.

Hu *et al.* [110] applied game theory to analyze how the selfishness of different parties may impact the overall caching performance by considering the relations and interactions among different parties. They considered two scenarios including the SBS caching and D2D caching. In the former one, multiple SPs aim to cache their own contents into SBSs with limited cache storages, and an auction game is proposed to solve the problem. For the latter one, they adopted a coalition game to analyze how a cooperative group can be formed to download contents together. They extended their work by introducing the concept of caching as a service in [12], where they utilized the wireless network virtualization technology and each SP has to pay for the SBS cache storages owned by the MNO. A multi-object auction mechanism was proposed to characterize the competition among SPs. Since all SPs tend to cache more contents to improve the service performance, they intented to act as the bidders and compete for limited cache storages. The utility function is related to the average content download

file. Their mechanism was carried out by a series of auctions, which are solved by the market matching algorithm [111]. Hamidouche *et al.* [112] assumed that all SBSs in a cache-enabled small cell network could choose their backhaul link types among wired links, mmW and sub6 GHz bands. They formulated a backhaul management minority game where the SBSs are the players and independently decide their backhaul link types and the numbers of files to download and cache from the MBS without sacrificing the current requests' QoS. The characteristic of a minority game is that players prefer the action selected by the minority group. The existence of a unique Nash equilibrium was then proved.

By considering the social ties among UEs, Hamidouche *et al.* [113] utilized the game theoretic approach to determine the content placement strategies to SPs. A many to many matching game was formulated between SPs and SBSs, where each file in SPs can be matched to a set of SBSs. SPs specify their preferences based on the average file download delay while SBSs prefer to store more popular files. The stable solution can be obtained by iteratively update the matching solution according to SPs' and SBSs' preferences, until neither of them can find a better preference.

G. Summary and Discussion

In this section, we survey several caching schemes and compare their pros and cons, including proactive caching, distributed caching, cooperative caching, coded caching, probabilistic caching and game theory based caching.

Proactive caching, contrary to reactive caching, caches the contents prior to receiving the requests. It helps improve the caching efficiency by pre-downloading popular contents during off-peak hours and serving users during peak hours. Hence, accurate prediction of user demands plays an important role in proactive caching. Most works characterize the user demands by estimating user mobility patterns, content popularity distributions and user social relationships via machine learning and big data analytics. Further research is still required to provide higher estimation accuracy.

Distributed caching, contrary to centralized caching, does not rely on the central controller to make caching decisions. Hence, it avoids the great burden on the single control node. In distributed caching, the caching strategies are usually made based on the local information (e.g., user requests and CSIs) and that from the neighboring ones, and hence can be addressed in the parallel and distributed manner. Most works investigate how to utilize the local and neighboring information to solve the content placement problem. However, unlike centralized caching, which owns a global view of all network status, distributed caching usually cannot obtain the optimal solutions. Hence, designing distributed caching strategies with performance guarantee still requires further research.

Cooperative caching allows multiple caches to share contents with each other, and hence it can alleviate the shortage of caching storages. In cooperative caching, a cache is usually aware of the caching status of its neighboring caches by exchanging information; this may incur significant signal overheads. Most works on cooperative caching do not consider these overheads. Hence, further research is needed to minimize the content retrieval latency while minimizing the overhead.

Coded caching allows files to be divided into coded chunks with coding, which are then cached in different cache nodes. Users obtain different coded chunks from different cache nodes and then decode these chunks into a complete requested file. Coded caching is usually coupled with a multicast technique to provision content delivery. However, coded caching aggravates the network system complexity and introduces more processing at network intermediary and terminal nodes. This drawback of coded caching is neglected by most works and hence needs further investigation.

Probabilistic caching allows contents to be cached at different caches with different probabilities. It is proposed to address the uncertainty of wireless networks caused by varying wireless channel conditions and user mobilities. The objective of probabilistic caching is usually to maximize the cache hit probability by optimizing the probabilities with which contents are cached at different locations. Most works assume the deterministic network status and so probabilistic caching is still an ongoing research.

Game theory based caching investigates the interactions of multiple coexisted parties (e.g., service providers and mobile network operators). Each party selfishly optimizes its own benefits which may conflict among different parties. A typical case is the auction game where the service providers act as the bidders and compete for the limited caching storages in order to improve their own caching performances. Most works only consider the non-cooperative game, and so the cooperative game requires further investigation.

VI. Content Request Analysis

In the next four sections, we will discuss the specific caching processes as categorized in Fig. 7. The benefits and solutions of caching highly depend on the traffic characteristics (e.g., user demand profiles). In this section, we discuss the request types and patterns.

A. Content Types

Contents such as files and videos are the most commonly cached ones. File downloading (e.g., software or data library updates, music or video downloads) is applicable for delay tolerant data because files can only be used after they have been delivered. On the other hand, video streaming requires a low initial delay, high video quality and few stalls during playback, and users prefer to start playing the video immediately after sending the request; video streamed data are thus considered as time sensitive data.

Next generation 5G networks will be enabling and empowering various emerging IoT applications [114], [115]. IoT applications generate a large amount of monitoring, measurement, and automation data. Requesting for these data may incur enormous traffic to the network. Furthermore, frequently activating IoT devices to fetch data drains their batteries [116]; this is a major challenge. Therefore, it is beneficial to cache IoT data to reduce the frequency of activating IoT devices

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2019.2908280, IEEE Communications Surveys & Tutorials

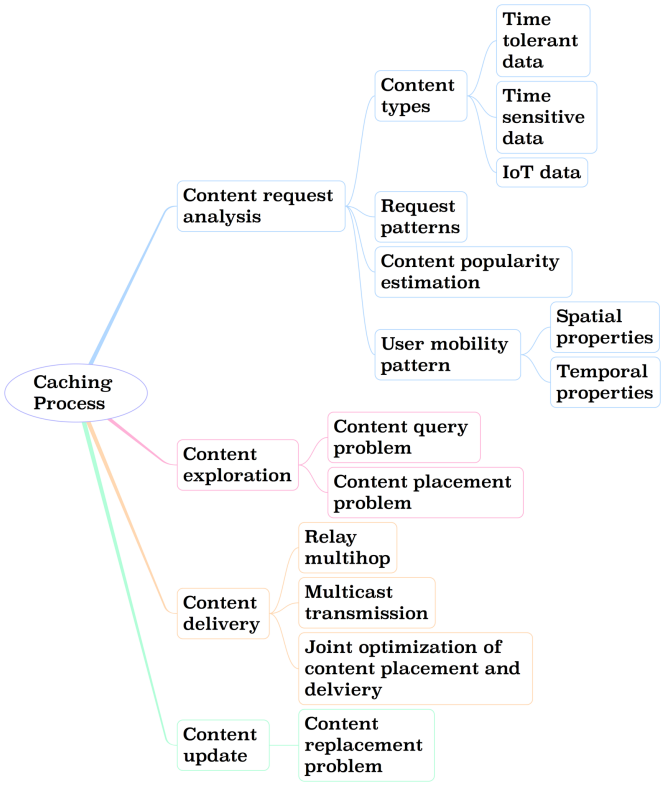


Fig. 7. Caching process.

and alleviating traffic loads in the network [80]. IoT data have much shorter lifetimes. Therefore, more intelligent caching strategies are called for to address this issue. Vural *et al.* [117] first proposed caching IoT data at network routers according to the data lifetime and hops between the respective router and the data source. A shorter data freshness indicates that caches should be updated more frequently and more requests should be sent to the IoT devices, thus resulting in more traffic load in the network. They further exploited the tradeoff between the traffic load and data freshness. Niyato *et al.* [118] discussed a novel caching scheme for IoT sensing service where IoT devices sense the ambient environment and send the data to the users. Considering the rapidly changing environment, a timer was set up to determine the freshness of IoT contents. The caches should be removed when they exceed the timer threshold. In order to maximize the cache hit ratio, a threshold adaptation algorithm was then designed to find the optimal threshold. Yao and Ansari [119] investigated caching schemes in a cache-enabled energy harvesting aided IoT network. A Stackelberg game was formulated to study the interactions between caching and energy harvesting.

B. Request Patterns

It is challenging to make optimum decisions for caching at the mobile edge without the knowledge of user request patterns. Request patterns can be extracted and analyzed to predict future requests and provide insights on the proactive caching. Observation of past request arrivals can be a feasible solution to obtain request patterns. However, it is challenging to either predict or model the request arrivals in the real world. Hence, intelligent algorithms based on predictions and stochastic models are required. In order to characterize the request patterns, most works assumed that the request arrivals follow the Poisson process [120]. Urgaonkar *et al.* [121] discussed the service migration problem from the cloud to the mobile edge clouds. They provided insights on where and when to migrate the services by incorporating the user mobility and request variation. They modeled request arrivals as a Markov process and proposed algorithms to minimize operational costs.

C. Content Popularity

The content popularity is defined as the ratio of the number of requests for a particular content over the total number of requests from users. It is usually obtained for a certain region during a given period of time. The key feature of the content popularity is that most people are interested in a few popular contents within a certain time period and hence these few contents account for major traffic loads. Li *et al.* [122] illustrated that top 5% videos in Youku contribute over 80% of contents inside mobile networks in China. In general, the content popularity distribution changes at a relatively slow speed [49]. Hence, the content popularity distribution is usually considered as a constant over a long time (e.g., one week for movies, and two or three hours for news) [49]. In addition, the global popularity in a large region like in a city or a country is often different from local popularity in a small region like in a campus [96]. The popularity prediction has become an active research field recently because it can be incorporated into many applications such as caching, online marketing, recommendation system, and media advertising. Many prediction methods have been proposed such as the cumulative statistics from the popularity correlation over time [123].

Several time series models have been proposed, e.g., autoregressive integrated moving average [124], regression models [125] and classification models [126]. These model-based forecasting schemes are usually carried out by machine learning methods. Leconte *et al.* [127] utilized a dynamic request model, the shot noise model (SNM) for mobile edge caching, to maximize the cache hit ratio. In SNM, each shot is associated with one content and is characterized by four dimensions including shape, duration, shot arrivals, and volume. For the shape, they utilized the rectangular pulse whose height represents the content popularity; the duration reflects the content life span; shot arrivals follow the Poisson process; volume is determined by a power-law distribution [123]. Famaey *et al.* [128] designed a content placement strategy based on the estimated content popularity. They proposed a general popularity prediction algorithm which learned from the historical request patterns and chose the best fit functions from constant function, power-law distribution, exponential distribution and Gaussian distribution. Then, future request patterns can be estimated based on the chosen best fit function.

The content popularity is reported to follow a Zipf distribution which belongs to the power law distribution [123]. The

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2019.2908280, IEEE Communications Surveys & Tutorials

Zipf distribution [129] defines the probability of a user in requesting the f-th file as

$$p_f = \frac{f^{-\delta}}{\sum_{j=1}^{N_f} j^{-\delta}}, \quad (1)$$

where δ is the skewness parameter. δ reflects the different levels of skewness of the Zipf distribution. A larger skewness implies larger deviations among different content popularities, i.e., most users request a small number of popular contents. Particularly, all the contents have the same popularity if δ equals 0, i.e., users request all contents with equal probability. However, the content popularity is time-varying and the fresh view of the system is required to know the content popularity. The massive data collection and processing are needed and hence the content popularity prediction is a complex task to handle.

Liu *et al.* [130] proposed a Hadoop-based distributed computing platform for monitoring large-scale network traffic data and demonstrated that their platform is efficient and cost-effective for analyzing user behaviors. Zeydan *et al.* [131] also focused on the deployment of a Hadoop-based big data processing platform inside a mobile core network in order to monitor the performance gains of caching with real data trials. They utilized machine learning algorithms to predict the content popularity and demonstrate improvements of QoE by the proactive caching at the edge.

D. User Mobility Pattern

Mobility is an important factor to be considered in mobile networks because it impacts mobile network topologies (e.g., the varying user-BS association) over time [132]. The latency of mobile networks can be attributed to unpredictable topology changes owing to mobility [133]. User mobility also contains many helpful information (e.g., social relationships and traffic patterns), which can be utilized to improve the caching performance. Poularakis and Tassiulas [134] discussed the storage allocation problem in Femtocaching networks. As the future position after the movement is highly related to the current position, user mobility can be modeled by a Markov chain. As users with more similar mobility patterns tend to have stronger social relationships, the user mobility pattern relies on social relationships [135]. Musolesi *et al.* [136] proposed a user mobility model based on user social relationships. They first created a social graph containing all social groups and then mapped these social groups to physical connected groups.

The user mobility pattern is generally classified into two categories including the spatial and temporal properties, which reflect the location-based and time-related characteristics, respectively [137]. We discuss these two categories in the following subsections.

1) Spatial Property: The spatial property refers to features regarding the physical location. The commonly used model for user mobility in mobile networks is the random waypoint model [138]. In this model, a user randomly moves to a destination point (i.e., waypoint) following a straight line towards the waypoint at a constant speed. After some waiting time in the waypoint, the user moves to another waypoint. Hence, the transition time and length between two waypoints are important parameters to measure. The Markov model can also be used to model the user mobility [139]. In the Markov model, a state represents the serving BS for each user and a transition probability indicates the probability of the network status moving from one state to another. Lv *et al.* [140] investigated the effect of living habits on the models of spatio-temporal prediction and next-place prediction based on the hidden Markov model.

2) Temporal Property: Temporal property characterizes the time-related features. The parameters to describe the temporal property include the contact time and inter-contact time. Content time refers to the time duration where users are within the transmission range and can be connected with each other. The inter-contact time refers to the time intervals between two contact times. Wang *et al.* [38] proposed a caching placement strategy with the aim to maximize the data offloading with the consideration of user mobility. They modeled the user mobility as an inter-contact model which collects the user connectivity information. Users within the transmission range are defined to be in contact and can share contents with each other by D2D communications. Rao *et al.* [94] jointly optimized caching in D2D caching and caching at helpers to maximize the traffic offloading probability. They considered the impact of user mobility which was modeled as the contact time and inter-contact time. They assumed that user locations remain static during the contact time and move to another location after the contact time.

E. Summary and Discussion

In this section, we discuss the request types, patterns, content popularity estimation and user mobility pattern.

The traditional cached contents are mainly files and videos. Files are usually time tolerant data while videos are time sensitive data. On the other hand, the IoT data have much less lifetimes because they are usually sensed data obtained by the IoT sensors. Stale IoT data cannot reflect the actual environment status. Hence, caching strategies should consider the freshness of the IoT data.

Request patterns are usually extracted and analyzed to predict future request patterns, which provide information for proactive caching. Most works assume that the request arrivals follow the Poisson process or Markov process. However, these mathematical models are not necessarily applicable to real traffics. Machine learning or big data analysis techniques may be utilized by observing the past requests.

The content popularity is an indispensable factor for the content placement problem. It reflects the probabilities of users requesting certain contents. Most works utilize the Zipf distribution to characterize the content popularity. However, Zipf is only demonstrated to be accurate in video downloading and may not be suitable to characterize other data types, e.g., IoT data. Machine learning could be an efficient tool to model the content popularity, e.g., regression and classification models. Practical systems, e.g., Hadoop-based big data processing platform, have been built to predict the content popularity by collecting and processing the past requests.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2019.2908280, IEEE Communications Surveys & Tutorials

User mobility pattern affects the content placement strategies. Most works characterize the user mobility patterns as a Markov chain because the future position is highly related to the current position. The user mobility pattern involves the spatial property and temporal property. The spatial property depends on the physical location. The common model is the random waypoint model. The temporal property reflects the time-related features of the user mobility, which are often measured by the contact time and inter-contact time. As users who have strong social relationships tend to have more similar mobility patterns, estimating the user mobility pattern by analyzing social relationships requires further studies.

VII. Content Exploration

When a user requests a content from a wireless network, the content should be searched through the network. The content query problem is to determine the means of searching for the requested content. To determine where the content is, the content placement problem should be taken into consideration. We will next discuss these problems.

A. Content Query Problem

Upon receiving a content request, a sequence of steps will be executed to process this request. In general, the UE first searches for the requested content in its local cache. If the content is not cached in its local storage, it will search the neighboring UEs for the content. If a user cannot find the desired content in other UEs, the request will be sent to the BS in a small region (e.g. FBS, PBS, SBS) and then to MBSs. If the requested content cannot be found at any of the BSs, it will be forwarded to the mobile core network. At the last resort, this request will go through the Internet to content providers. The forwarding and delivery of the content query request are shown in Fig. 8.

In wireless networks with overlapped coverages of different BSs, caching may change the user association. Users may not always be associated with the nearest BS when considering delays and traffic loads. Instead, associating users with the BSs that cache the requested contents may be more beneficial because they can obtain the content directly from the BSs. Hence, the user association problem plays an important role in mobile edge caching.

Poularakis *et al.* [141] proposed an approximation algorithm to jointly optimize the user association and content placement problem in SBS caching networks with the objective to maximize the requests served by the SBSs. The problem is constrained by cache storages and SBS backhaul capacities. They further studied the joint problem of user association and content placement in the context of video delivery in cache-enabled SBS networks to minimize the average user experienced delay in [142]. A Lagrangian relaxation algorithm was then designed to solve this problem. However, their works both assumed that different SBSs operate at orthogonal frequency bands and hence neglected the interference among SBSs. With consideration of the interferences, Wang *et al.* [68] jointly optimized the content placement and user association problem in SBS caching networks to minimize the average download delay. They demonstrated that this joint problem is NP-hard. They then separated coupled variables (i.e., caching variable and user association variable) to reduce the problem complexity. A Lagrangian relaxation algorithm was then designed to address the transformed problem.

From the perspective of game theory, Pantisano *et al.* [143] modeled the user association problem between SBSs and UEs as a one-to-many matching game with the aim to minimize the average file download delay. In this matching game, each SBS creates a preference list of UEs based on the current caching status and backhaul congestion state, while each UE chooses its preference list of SBSs according to the downlink channel capacity. A stable matching solution can be obtained after several iterations.

Han *et al.* [144] additionally considered the energy consumption of all SBSs in the cache-enabled SBS networks with hybrid power supplies. They tried to solve the user association problem between UEs and SBSs to minimize the average traffic delivery latency with the consideration of SBS backhaul capacities, green power utilization and cache hit ratio. They assumed that the cache hit ratio of each SBS can be estimated from historical statistics. The total delay consists of both the wireless channels and backhaul links which can be reduced with a high cache hit ratio.

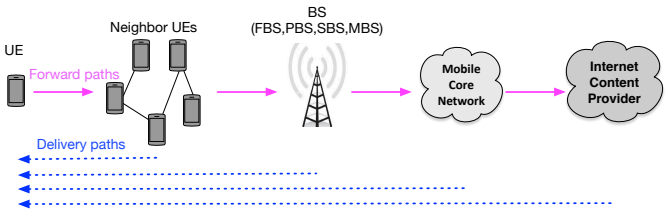


Fig. 8. Content query request forwarding and delivery.

B. Content Placement Problem

The content placement problem is always the most important issue in any caching schemes. It determines which content should be placed at each cache. It also addresses the size of each cache and where the cache should be located at different network nodes. Note that distributing the contents to cache nodes introduces traffic into the network; the additional traffic overhead should be considered in designing the content placement strategies.

Golrezaei *et al.* [54] investigated the caching policy in femtocaching networks to minimize the average file download delay under the constraint of helper cache storages. Users can fetch contents from either a helper which has low-rate backhaul but high storage capacity or the BS which possesses the whole file library. They transformed the average download delay minimization problem into the cache hit ratio maximization problem. However, their work did not consider the wireless channel conditions and assumed the delays between users and helpers are constant. Song *et al.* [145], instead, considered the wireless channel conditions in femocaching networks. They formulated the content placement problem as an ILP model and demonstrated that the closed-form solution

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2019.2908280, IEEE Communications Surveys & Tutorials

of this model was difficult to obtain. They then designed a greedy algorithm to obtain the suboptimal solution. In their simulations, they demonstrated that caching the most popular files was not always the best choice and wireless channel conditions also affected the caching performance. On the other hand, Peng *et al.* [146], by considering the backhaul constraints, studied the content placement problem in a cache-enabled wireless network where BSs are equipped with caches and a central controller can transmit files to users if files are not cached. They aimed to minimize the file transmission delay consisting of the delay in the backhaul links and wireless transmissions. They first formulated this problem as a mixed-integer nonlinear programming problem and then designed a relaxation-based heuristic algorithm to obtain the suboptimal solutions.

User social relationships can impact the content placement strategies in mobile edge caching. Wu *et al.* [22] explained that the challenges of social-aware caching root in the fact that the contact among UEs is usually opportunistic in practical systems owing to the short-range D2D communications. Because of the scarce spectrum resources and contact durations, the transmitted contents can be very limited. They also discussed four content placement strategies: 1) global content placement based on the user contact information; 2) social aware content placement based on the respective community; 3) individual content placement based on the user's own interest; 4) random content placement.

User mobility is also a critical factor of the content placement problem. Users can move from one location to another and stay in one cell for a certain time (i.e., cell sojourn time). During the cell sojourn time, the user is associated with one BS. Therefore, the user can only receive the data from certain BS during that time. Hence, the cell sojourn time can impact the content placement strategy in BSs. Wang *et al.* [137] investigated the content placement problem by utilizing the user mobility information in mobile edge caching. They discussed caching at BSs and UEs separately with the objective of maximizing the traffic offloading probability. Caching at BSs is constrained by the wireless transmission rate and the sojourn time. The content placement problem for caching at BSs is formulated as a convex optimization problem for uncoded caching and a mixed integer programming (MIP) problem for coded caching. Caching at UEs is constrained by the UE transmission distance and user contact times. The content placement problem for caching at UEs falls into the problem of maximizing the monotone submodular function.

C. Summary and Discussion

In this section, we discuss the content exploration phase of the caching process, including the content query problem and content placement problem.

The content query problem determines how the requests are forwarded in order to find the contents and which network nodes should serve the users (i.e., user association problem). In general, the requested content is first searched in neighbor UEs, then in BSs (e.g., FBSs, PBSs and MBSs), and finally from the mobile core network or remote Internet content servers. Unlike the traditional user association problem in wireless network, users may not be associated with the nearest BSs in mobile edge caching. Instead, it is more beneficial to associate users to the caches that have the requested content. Since the cache locations affects the user association strategies, the user association problem is usually jointly optimized with the content placement problem in most works. In cooperative caching where caches share contents, the joint optimization of content placement and user association has not been addressed yet and needs further study.

The content placement problem, which determines how to distribute contents to different caches, is the most important issue in caching related studies. The objective of this problem is usually to minimize the average file download delay and maximize the cache hit probability. As the placement decision involves the 0-1 variables, most works formulate this problem as an ILP model which is usually solved by designing heuristic algorithms to obtain suboptimal solutions. The content placement problem in D2D caching usually considers the user social relationships which affects the user contact times and location distances. However, most works neglect the additional traffic overhead caused by caching the contents. The content placement strategies considering the overhead are still required.

VIII. Content Delivery

The content delivery problem addresses the issues of how to transmit contents to the users. Specifically, these issues involves the locations the content should be transmitted from, the transmission power and the frequency bands the content should be transmitted. Hence, the specific physical conditions (e.g., available spectrums, wireless channel conditions and interference) should be taken into consideration when designing content delivery strategies.

A. Multicast Transmission

Multicast enables simultaneous content transmission to multiple destinations by broadcasting. Via multicast, a BS can concurrently serve multiple users who request identical contents. Hence, multicast can help reduce duplicated transmissions and energy consumption. Users' requests may be initiated at different times. For the delay-tolerant file downloading, users can endure some initial delays so they can wait for each other before initiating the multicast transmission. On the other hand, users may suffer from the initial delays and the QoE degradation for delay sensitive video streaming. Therefore, multicast schemes should consider different data types. Maddah-Ali and Niesen [106] proposed a novel coded multicast strategy, which coded parts of popular contents and pre-cached them. In the content delivery phase, the content server can serve multiple requests with a single multicast transmission. However, this method scales badly in practice because the coding complexity grows exponentially. Poularakis *et al.* [147] proposed a caching policy for small cell networks to minimize the service cost which is related to the incurred traffic. In their system model, the SBSs can use multicast to transmit cached contents to users under their coverages, and the MBS can also transmit

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2019.2908280, IEEE Communications Surveys & Tutorials

contents to users by multicast transmission but incur more service cost than that of SBSs. They designed several heuristic algorithms to address this problem and then demonstrated that the multicast-aware caching scheme performs 52% better than the multicast-agnostic one. Liao *et al.* [148] utilized the multicast technique to deliver content from MBS to SBSs in a cache-enabled small cell network. They designed caching policies to minimize the long-term average backhaul load for two different networks (i.e., small scale networks and large scale networks). For the small scale networks, they proposed a greedy algorithm to provision content delivery. For the large scale networks, a multicast-aware in-cluster cooperative caching algorithm was developed in which different SBSs can share content with each other.

Instead of utilizing multicast to optimize the caching policy, Zhou *et al.* [149] tried to optimize the multicast content delivery policy under a given content distribution in HetNets consisting of a MBS and SBSs to minimize the average network delay and power. The MBS provides full coverage of the network and generates higher power consumption. The SBS cells are not overlapped and there is no interference among all SBSs. Considering the special structure of monotonicity of the value function, they developed an approximate dynamic programming algorithm to solve this problem. They extended their work in [150], where they additionally considered the impact of content sizes and the wireless channel conditions on the multicast scheduling policy. They formulated the optimization problem as a Markov decision process and designed a suboptimal algorithm to solve this problem.

B. Relay Multihop

The D2D delivery enables one UE to retrieve contents via D2D links from nearby UEs. A higher UE density can lead to a higher probability that the required contents can be provided. The multi-hop delivery through D2D relays allows neighboring UEs to serve as relays for content delivery. This relay-based mechanism allows a broader range of content delivery. Moreover, when the requested contents are cached in multiple UEs, they can cooperatively deliver contents to provide a higher transmission rate. Xia *et al.* [151] investigated the scenario where UEs cooperate to download the content via multihop relaying. Specifically, their scheme grouped UEs into multicast groups which can download files from the BS by wireless multicast transmissions. Then, each multicast group can act as the relay to transmit files to other multicast groups. Hence, the content delivery is carried out group by group. They discussed the problem of how to form efficient groups to minimize the power consumption of all UEs. They demonstrated by simulations that the total power consumption can be saved significantly by grouping UEs in multihop D2D networks.

C. Joint Optimization of Content Placement and Content Delivery

In practice, the content placement and content delivery can impact each other. On one hand, the content placement determines the distribution of contents and impacts the content delivery paths. On the other hand, the statistics of the content delivery over a long time can be utilized to explore the popularity distribution. The caches can be periodically updated based on these statistics. Therefore, studying the coupling between the content distribution and delivery is very important.

Maddah-Ali and Niesen [106] jointly optimized the caching and coded multicast delivery, and demonstrated that the joint optimization of caching and delivery can improve the caching performance. Cui and Jiang [61] jointly considered the caching placement and multicast delivery in a cache-enabled two-tier HetNet consisting of the MBS tier and the PBS tier. They considered identical caching in the MBS tier and random caching in the PBS tier. In the MBS tier, all MBSs cache the same set of files. In the PBS tier, all PBSs randomly cache different files except the files that have already been cached at MBSs. Both MBSs and SBSs can transmit files to user by multicasting. They formulated the joint problem as a mixed discrete-continuous optimization problem with the objective to maximize the successful transmission probability and designed a near optimal algorithm to solve it. Gregori *et al.* [56] jointly optimized caching and transmission policies to minimize the MBS energy consumption. In their system model, SBSs and UEs are equipped with cache storages. A SBS can serve multiple users simultaneously by multicasting and UEs can share data through D2D communications. They formulated this joint problem as a finite dimensional convex problem and then designed a projected subgradient algorithm to solve it.

D. Summary and Discussion

In this section, we discuss the content delivery problem on how to transmit contents to users.

The delivery strategies can be generally classified into unicast and multicast transmission. In unicast transmission, the BS uses different time and frequency resources to serve different users and content delivery strategies are usually designed to minimize the network latency and backhaul traffic. In contrast, multicast transmission enables the BS to simultaneously serve multiple users, who request the same content in the same cell, at the same time and using the same frequency resources. However, requests do not necessarily arrive in time. Hence, the serving BS has to delay some requests, collect multiple requests in a certain time window, and then serve multiple users at the same time. Therefore, how to determine this time window is critical in order not to compromise the user QoS. A longer time window implies that more users can be served at the same time and thus increases the spectrum efficiency. On the contrary, a shorter time window leads to a shorter delay and hence higher user QoS. Therefore, there is a tradeoff between spectrum efficiency and user QoS in determining the time window; this tradeoff requires further research.

The content placement and delivery strategies couple with each other. Content placement determines the locations of the cached content, and further affects the content delivery, which decides the transmission paths from the caching locations to the users. On the other hand, the statistics of content delivery paths can be leveraged to explore the content popularity,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2019.2908280, IEEE Communications Surveys & Tutorials

which impacts the content placement strategies. Moreover, if multicast is adopted for the content delivery strategy, different BSs prefer to cache different contents to increase the cache hit ratio (i.e., an important metric to evaluate the content placement strategy).

IX. Content Update

Caching strategy using the outdated information (e.g., content popularities, user locations, network traffic loads, etc) may degrade performance because it may not reflect the current network status. Hence, it is critical to update caches at intervals. The cache update process generally takes place after the content delivery is completed. The content replacement problem is about what contents should be removed from caches, when to remove them and how to cache new contents.

A. Content Replacement Problem

Several content replacement strategies have been proposed such as least frequently used (LFU), least recently used (LRU), and their variants [152]. LRU updates each cache to keep the most recently requested contents while LFU keeps the most frequently requested contents. Taking LRU as an example, upon the arrival of a request, it is first forwarded to the local cache and served by the cache if the request file is cached. Otherwise, it will be served by the content provider, and then the cache evicts the least recently used file and cache the newly requested file.

Pedarsani *et al.* [153] assumed that the server has two delivery modes. They used the coded delivery for the partially cached files in one mode and uncoded delivery for entirely cached files in another mode. In their caching system, the caches are updated after the delivery phase is executed. If the requested file is not partially cached, it is delivered uncoded from the content provider. Then, all caches remove the least recently used contents and cache chosen bits of the requested file from the content provider, so that the requested file is then partially stored at each cache after this procedure. By this approach, the changes of the requested files can be traced and utilized to learn the distribution of file popularities over time.

The schemes (e.g., LRU and LFU), which are performed immediately upon request arrivals, are called online schemes. However, online schemes only consider recent and current status and may not obtain the global optimal solutions. Hence, the future content popularity should also be taken into consideration. Li *et al.* [154] proposed a novel cache replacement method (PopCaching) to learn and estimate the content popularity. They then determined which content should be evicted from the cache. They demonstrated that PopCaching converges quickly and approximates the optimal cache hit rate.

B. Summary and Discussion

In this section, we discuss the content update phase of the caching process, which involves the content replacement problem.

Caches should be updated, i.e., the unpopular content should be replaced with the popular ones, because the content popularity varies with time. Caches prefer to maintain the popular contents so that more users can be served. The content replacement problem determines which contents should be removed and when to remove them. Conventional content replacement strategies include LFU and LRU which maintain the least frequently used and least recently used contents, respectively. As the content popularity is an important factor in updating caches, machine learning can be utilized to learn and estimate the content popularity. The frequency of content update is also a pressing issue to investigate. Updating contents too frequently introduces heavy traffic to the network. In contrast, the caches may not satisfy the user demands without updating caches for a long time. Hence, the tradeoff between the network traffic and content update frequency is yet to be resolved.

X. Challenges and Future Work

To exploit the full potential of mobile edge caching, the unique challenges in wireless networks should be considered, e.g., uncertain channel, interference, user mobility, limited UE battery life and user privacy. In this section, we discuss these challenges faced by mobile edge caching and identify the future research directions.

A. Caching in IoT networks

IoT networks, connecting billions of devices and sensors towards a smarter physical world, have gained attention from both the industries and academics [155]. The fundamental application of IoT is the sensing service that allows users to monitor ambient environment (e.g., temperature and air pollution level) via sensors [156]. Frequently activating IoT devices to transmit data not only injects tremendous traffic into the network but also exhausts the limited battery of IoT sensors. Caching in IoT networks, as shown in Fig. 9, equips the IoT gateway with caches to store IoT data so that they can be directly transmitted to the mobile users without activating the IoT sensors [80]. Quevedo *et al.* [157] introduced caching technology in the IoT networks and demonstrated that caching reduces both energy consumption of IoT sensors and bandwidth usage of wireless networks. Sun and Ansari [158] designed an energy efficient caching strategy in IoT networks to minimize the system delay. However, the above works did not take the lifetime of IoT data into consideration. As the ambient environment is dynamic, the cached IoT data may not accurately reflect the actual physical status. Hence, the cache should be updated according to the lifetime of the IoT data. How to measure the lifetime of IoT data and how to design the caching strategies by incorporating the IoT data lifetime require further investigation.

B. Fading and Interference

In most current literature, caching policies fail to consider the channel fading and interference, which are critical to wireless networks. Hence, the optimized policy may not perform well in practice. When the path loss and inter-cell interference (ICI) are considered, the caching policies may not improve

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2019.2908280, IEEE Communications Surveys & Tutorials

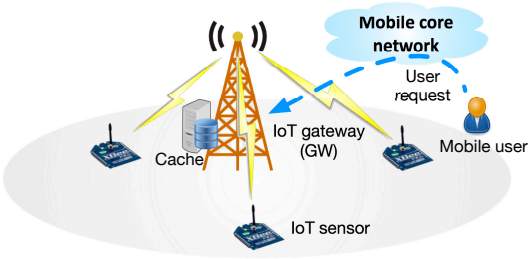


Fig. 9. Caching in IoT networks.

SE and EE despite of the increase of the cache hit probability because the network capacity is limited by the path loss and interference [50]. When a farther BS caches the requested content instead of the nearest one, the signal from the farther BS may be weaker than the nearest one due to the path loss. The nearest BS may generate interference towards the user. Hence, the content placement problem should take the interference into consideration.

The caching process itself may also generate interference to other ongoing transmissions. The conventional method is to pre-cache the content during the off-peak hours. However, pre-caching highly depends on the accurate estimation of the content popularity. Hence, content popularity estimation is still an ongoing research. On the other hand, caching with the consideration of the caching overhead (e.g., interference to other transmissions, bandwidth usage, etc.) remains a challenging issue.

C. Security and Privacy

The purpose of security (e.g., network security and storage security) is to prevent the information and resources from attacks. In order to defend attacks, security protocols should be designed. Mukherjee *et al.* [159] surveyed security and privacy issues in fog computing. Kim *et al.* [160] addressed the content poisoning attack in name data networks. The name data network is a new network paradigm which enables users to request content by specifying the content name rather than the IP address. Leguay *at el.* [161] proposed a security protocol to enable caching of encrypted content in content delivery networks. Although these works have exploited many security solutions, they may not be suitable for mobile edge caching because of the more dynamic wireless channels and mobile traffic faced by mobile edges. Therefore, security solutions tailored for mobile edge caching still require further investigation.

Owing to the broadcast nature of wireless networks, transmissions are vulnerable to various malicious attacks, including the passive eavesdropping for data interception and the active jamming for disrupting legitimate transmissions. Eavesdroppers can easily overhear the wireless communication sessions. The conventional method is to utilize cryptographic techniques to prevent attackers from intercepting data. The jamming attack disrupts the legitimate transmission by maliciously generating interferences. The conventional method to mitigate jamming is to frequently change the central frequency of the transmitted waveform so that the attacker cannot interrupt the legitimate transmissions. The criteria for designing a viable security safeguard include the authenticity, confidentiality, integrity and availability [162]. Authenticity identifies the authorized users from the unauthorized ones. Confidentiality limits the data access to the authorized users and prevents data disclosure to unauthorized ones. Integrity ensures the data accuracy without any falsification and modification by unauthorized users. Availability enables authorized users to access the requested data anytime and anywhere. Mobile edge caching strategies, which aim to meet both the user QoS and security requirements, still remain very challenging.

A privacy violation refers to the disclosure of the private and sensitive information to unauthorized individuals [163]. The conventional way to preserve privacy is anonymization, which encrypts or removes personally identifiable information from the transmission data. However, advances of big data technology have empowered attackers to combine anonymized data with non-anonymized public data collected from users (e.g., shopping and reading preferences, locations and photos) to identify individuals [164]. Furthermore, sensitive data can be sniffed and extracted directly by observing and analyzing data with data mining and machine learning techniques. Caching policies that can prevent attackers from extracting sensitive information are urgently sought after.

The user privacy conflicts with the information sharing (e.g., cached contents, user preferences, social ties and user mobility patterns). When the content popularity is obtained to improve the caching performance, the content information should be extracted and analyzed with respect to the user privacy. Hence, user privacy regulations should be well designed to balance the tradeoff between the user privacy and caching performance. Therefore, how to extract useful information without compromising the user privacy should be taken into consideration. The tradeoff between privacy and caching efficiency is an interesting research undertaking.

D. Software Defined Networking (SDN) based Mobile Edge Caching

The SDN technology is assumed to be one of the main candidates for the 5G core network in order to meet the restrict latency requirement of the 5G networks [7]. SDN is proposed as a new network paradigm to enable the network flexibility by separating the data and control plane [165]. A centralized controller in SDN architecture is aware of all states of network entities [166]. With all the information, the controller can manage network resources more efficiently and flexibly. Coordinated decisions can be made by the controller and guide the network to optimal operating conditions. Hence, integrating SDN into mobile networks has been studied in many researches [167]. Furthermore, the user mobility information can also be obtained and used for designing caching schemes to achieve better performance. The architecture of SDN based mobile edge caching is illustrated in Fig. 10. However, the centralized SDN controller faces high risks of a single point of failure and can be a target for attacks. If the SDN controller is broken down, how to guarantee service continuity still remains a challenging issue.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/COMST.2019.2908280, IEEE Communications Surveys & Tutorials

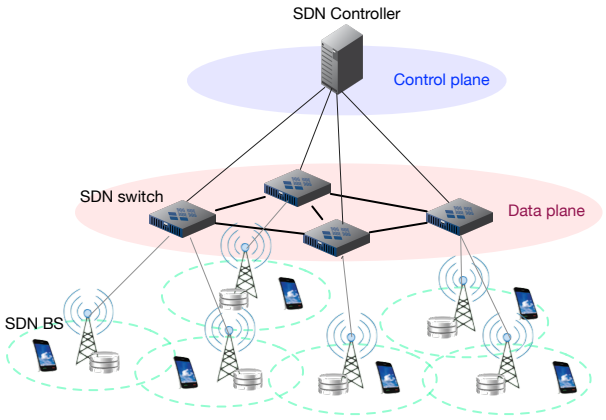


Fig. 10. SDN based mobile edge caching.

E. Integration of Wireless Caching and Wired Caching

Caching mechanisms have been well studied in both the wired and wireless networks. However, fixed mobile convergence networks has not been studied yet, i.e., caching in wired and wireless networks have only been investigated separately. As content requests are first forwarded to the wireless segment and then to the wired segment of the network, the latency is attributed from these two parts. The throughput can also be limited by the capacity of both segments. For example, the user QoS is generally considered as the total latency including the wired and wireless latencies (i.e., $T = T_{wired} + T_{wireless}$). Latency incurred via the wired segment can be lessened by better caching strategies (i.e., T_{wired} decreases). Hence, the latency requirement for wireless network can be reduced (i.e., $T_{wireless}$ increases) when a certain user QoS is maintained (i.e., T keeps the same). Therefore, the BS transmission power can be reduced accordingly as $T_{wireless}$ increases, thus reducing the energy consumption. As a result, a better caching strategy in the wired network can help reduce the power consumption in the wireless network. In order to further provide better network performances in terms of the energy consumption, system throughput and network delay, caching incorporated with joint optimization of both wireless and wired networks is worth further investigation.

F. Net Neutrality

Net neutrality originally aims to ensure equal and fair passage of all packets in IP networks so that Internet end-users have fair access to other network endpoints to distribute and access contents [168]. However, it is hardly applicable to wireless networks because of the scarcity of spectrum and channel variations [169]. For example, users in the same cell may experience different data rates because of the diverse channel conditions. The mobile operators may limit the bandwidth for file sharing while providing more bandwidth for video streaming. Net neutrality in wireless networks should ensure QoS for all users. In order to measure net neutrality in mobile edge caching, the performance information should be extracted and analyzed while contents are being delivered. However, no existing work has been found on how to measure net neutrality in mobile edge caching, and how the caching strategies should be adjusted when traffic discrimination has been identified.

XI. Conclusion

Mobile edge caching is a practical means to reducing duplicated traffic through wireless backhaul links and improving user QoE for 5G networks, which require stricter latency and higher throughput. In our work, we have conducted a comprehensive survey regarding different aspects of mobile edge caching. We began with a brief introduction of mobile edge computing and mobile edge caching. We have discussed the caching schemes based on different caching locations and different performance criteria. In addition, we have delineated the caching process, summarized as four phases including the content request, exploration, delivery, and update, respectively. At the end, we have enlisted the challenges of mobile edge caching that require further investigation.

References

[1] Cisco Systems, "Cisco visual networking index: Global mobile data traffic forecast update, 2016–2021 white paper," San Jose, CA, USA, Mar. 2017.

[2] M. Agiwal, A. Roy, and N. Saxena, "Next generation 5G wireless networks: A comprehensive survey," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 3, pp. 1617–1655, Third Quarter 2016.

[3] M. R. Rahimi, J. Ren, C. H. Liu, A. V. Vasilakos, and N. Venkatasubramanian, "Mobile cloud computing: A survey, state of art and future directions," *Mobile Networks and Applications*, vol. 19, no. 2, pp. 133–143, Apr. 2014.

[4] A. Kiani and N. Ansari, "Edge computing aware NOMA for 5G networks," *IEEE Internet of Things Journal*, vol. 5, no. 2, pp. 1299–1306, Apr. 2018.

[5] J. G. Andrews, S. Buzzi *et al.*, "What will 5G be?" *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1065–1082, 2014.

[6] 5GPPP Architecture Working Group, "View on 5G architecture," 5GPPP Initiative, Tech. Rep., Dec. 2017. [Online]. Available: https://5g-ppp.eu/wp-content/uploads/2018/01/5G-PPP-5G-Architecture-White-Paper-Jan-2018-v2.0.pdf

[7] I. Parvez, A. Rahmati, I. Guvenc, A. I. Sarwat, and H. Dai, "A survey on low latency towards 5G: RAN, core network and caching solutions," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 4, pp. 3098–3130, Fourth quarter 2018.

[8] X. Sun and N. Ansari, "Latency aware workload offloading in the cloudlet network," *IEEE Communications Letters*, vol. 21, no. 7, pp. 1481–1484, Jul. 2017.

[9] N. Chalaemwongwan and W. Kurutach, "Mobile cloud computing: A survey and propose solution framework," in *2016 13th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, Jun. 2016, pp. 1–4.

[10] 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, pp. 6757–6779, 2017.

[11] L. Li, G. Zhao, and R. S. Blum, "A survey of caching techniques in cellular networks: Research issues and challenges in content placement and delivery strategies," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 3, pp. 1710–1732, third quarter 2018.

[12] Z. Hu, Z. Zheng, T. Wang, L. Song, and X. Li, "Caching as a service: Small-cell caching mechanism design for service providers," *IEEE Transactions on Wireless Communications*, vol. 15, no. 10, pp. 6992–7004, 2016.

[13] 5G Infrastructure PPP Association *et al.*, "5G vision-the 5G infrastructure public private partnership: the next generation of communication networks and services," *White Paper, Feb.*, 2015.

[14] ETSI GS MEC 001 V1.1.1, *Mobile Edge Computing (MEC) Terminology*, Mobile Edge Computing (MEC) ETSI Industry Specification Group (ISG), Mar. 2016.

[15] ETSI GS MEC 002 V1.1.1, *Mobile Edge Computing (MEC);Framework and Reference Architecture*, Mobile Edge Computing (MEC) ETSI Industry Specification Group (ISG), Mar. 2016.

[16] ETSI GS MEC 003 V1.1.1, *Mobile Edge Computing (MEC); Technical Requirements*, Mobile Edge Computing (MEC) ETSI Industry Specification Group (ISG), Mar. 2016.

[17] W. Ali, S. M. Shamsuddin, and A. S. Ismail, "A survey of web caching and prefetching," *Int. J. Advance. Soft Comput. Appl*, vol. 3, no. 1, pp. 18–44, 2011.

[18] G. Zhang, Y. Li, and T. Lin, "Caching in information centric networking: A survey," *Computer Networks*, vol. 57, no. 16, pp. 3128–3141, 2013.

[19] M. Zhang, H. Luo, and H. Zhang, "A survey of caching mechanisms in information-centric networking," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 3, pp. 1473–1499, 2015.

[20] T. Han, N. Ansari, M. Wu, and H. Yu, "On accelerating content delivery in mobile networks," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 3, pp. 1314–1333, Third Quarter, 2013.

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

[22] Y. Wu, S. Yao, Y. Yang, T. Zhou, H. Qian, H. Hu, and M. Hamalainen, "Challenges of mobile social device caching," *IEEE Access*, vol. 4, pp. 8938–8947, 2016.

[23] E. Nygren, R. K. Sitaraman, and J. Sun, "The Akamai network: a platform for high-performance internet applications," *ACM SIGOPS Operating Systems Review*, vol. 44, no. 3, pp. 2–19, 2010.

[24] S. Scellato, C. Mascolo, M. Musolesi, and J. Crowcroft, "Track globally, deliver locally: improving content delivery networks by tracking geographic social cascades," in *Proceedings of the 20th international conference on World wide web*. ACM, 2011, Conference Proceedings, pp. 457–466.

[25] S. Borst, V. Gupta, and A. Walid, "Distributed caching algorithms for content distribution networks," in *2010 Proceedings IEEE INFOCOM*, 2010, Conference Proceedings, pp. 1–9.

[26] ETSI White Paper No. 24, "MEC deployments in 4G and evolution towards 5G," First Edition, Feb. 2018.

[27] 3GPP TS 23.501 V15.0.0, "3rd generation partnership project; technical specification group services and system aspects; system architecture for the 5G system; stage 2 (Release 15)," Dec. 2017.

[28] H. T. Dinh, C. Lee, D. Niyato, and P. Wang, "A survey of mobile cloud computing: architecture, applications, and approaches," *Wireless communications and mobile computing*, vol. 13, no. 18, pp. 1587–1611, 2013.

[29] Y. C. Hu, M. Patel, D. Sabella, N. Sprecher, and V. Young, "Mobile edge computing—a key technology towards 5G," *ETSI White Paper*, vol. 11, no. 11, pp. 1–16, 2015.

[30] A. Kiani and N. Ansari, "Toward hierarchical mobile edge computing: An auction-based profit maximization approach," *IEEE Internet of Things Journal*, vol. 4, no. 6, pp. 2082–2091, Dec 2017.

[31] X. Sun and N. Ansari, "EdgeIoT: Mobile edge computing for the internet of things," *IEEE Communications Magazine*, vol. 54, no. 12, pp. 22–29, Dec. 2016.

[32] E. Bastug, M. Bennis, and M. Debbah, "Living on the edge: The role of proactive caching in 5G wireless networks," *IEEE Communications Magazine*, vol. 52, no. 8, pp. 82–89, 2014.

[33] N. Ansari and X. Sun, "Mobile edge computing empowers internet of things," *IEICE Transactions on Communications*, vol. 101, no. 3, pp. 604–619, 2018.

[34] M. T. Beck, M. Werner, S. Feld, and S. Schimper, "Mobile edge computing: A taxonomy," in *Proc. of the Sixth International Conference on Advances in Future Internet*. Citeseer, 2014, pp. 48–55.

[35] M. Peng, Y. Sun, X. Li, Z. Mao, and C. Wang, "Recent advances in cloud radio access networks: System architectures, key techniques, and open issues," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 3, pp. 2282–2308, Third Quarter, 2016.

[36] Mobile, China, "C-RAN: the road towards green RAN," *White Paper,*, vol. 2, 2011.

[37] T. Han and N. Ansari, "A traffic load balancing framework for software-defined radio access networks powered by hybrid energy sources," *IEEE/ACM Transactions on Networking*, vol. 24, no. 2, pp. 1038–1051, Apr. 2016.

[38] R. Wang, J. Zhang, S. H. Song, and K. B. Letaief, "Mobility-aware caching in D2D networks," *IEEE Transactions on Wireless Communications*, vol. 16, no. 8, pp. 5001–5015, Aug. 2017.

[39] A. Pyattaev, O. Galinina, S. Andreev, M. Katz, and Y. Koucheryavy, "Understanding practical limitations of network coding for assisted proximate communication," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 2, pp. 156–170, 2015.

[40] M. Ji, G. Caire, and A. F. Molisch, "Fundamental limits of distributed caching in D2D wireless networks," in *2013 IEEE Information Theory Workshop (ITW)*, Sep. 2013, pp. 1–5.

[41] M. Ji, G. Caire, and A. F. Molisch, "Wireless device-to-device caching networks: Basic principles and system performance," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 1, pp. 176–189, 2016.

[42] M. Sheng, C. Xu, J. Liu, J. Song, X. Ma, and J. Li, "Enhancement for content delivery with proximity communications in caching enabled wireless networks: architecture and challenges," *IEEE Communications Magazine*, vol. 54, no. 8, pp. 70–76, 2016.

[43] M. Ji, G. Caire, and A. F. Molisch, "Fundamental limits of caching in wireless D2D networks," *IEEE Transactions on Information Theory*, vol. 62, no. 2, pp. 849–869, 2016.

[44] Y. Zhang, Y. Xu, Q. Wu, X. Liu, K. Yao, and A. Anpalagan, "A game-theoretic approach for optimal distributed cooperative hybrid caching in D2D networks," *IEEE Wireless Communications Letters*, vol. 7, no. 3, pp. 324–327, June 2018.

[45] N. Golrezaei, A. G. Dimakis, and A. F. Molisch, "Scaling behavior for device-to-device communications with distributed caching," *IEEE Transactions on Information Theory*, vol. 60, no. 7, pp. 4286–4298, 2014.

[46] A. Altieri, P. Piantanida, L. R. Vega, and C. G. Galarza, "On fundamental trade-offs of device-to-device communications in large wireless networks," *IEEE Transactions on Wireless Communications*, vol. 14, no. 9, pp. 4958–4971, 2015.

[47] X. Wang, M. Chen, Z. Han, D. O. Wu, and T. T. Kwon, "TOSS: Traffic offloading by social network service-based opportunistic sharing in mobile social networks," in *IEEE INFOCOM 2014 - IEEE Conference on Computer Communications*, Apr. 2014, pp. 2346–2354.

[48] Y. Wu, S. Yao, Y. Yang, Z. Hu, and C. Wang, "Semigradient-based cooperative caching algorithm for mobile social networks," in *2016 IEEE Global Communications Conference (GLOBECOM)*, Dec. 2016, pp. 1–6.

[49] 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 Communications Magazine*, vol. 51, no. 4, pp. 142–149, 2013.

[50] D. Liu and C. Yang, "Energy efficiency of downlink networks with caching at base stations," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 4, pp. 907–922, 2016.

[51] M. Sheng, W. Han, C. Huang, J. Li, and S. Cui, "Video delivery in heterogenous crans: architectures and strategies," *IEEE Wireless Communications*, vol. 22, no. 3, pp. 14–21, 2015.

[52] S. A. R. Zaidi, M. Ghogho, and D. C. McLernon, "Information centric modeling for two-tier cache enabled cellular networks," in *2015 IEEE International Conference on Communication Workshop (ICCW)*, Jun. 2015, pp. 80–86.

[53] Z. Chang, Y. Gu, Z. Han, X. Chen, and T. Ristaniemi, "Context-aware data caching for 5G heterogeneous small cells networks," in *2016 IEEE International Conference on Communications (ICC)*, May 2016, pp. 1–6.

[54] N. Golrezaei, K. Shanmugam, A. G. Dimakis, A. F. Molisch, and G. Caire, "Femtocaching: Wireless video content delivery through distributed caching helpers," in *2012 Proceedings IEEE INFOCOM*, Mar. 2012, pp. 1107–1115.

[55] P. Blasco and D. Gündüz, "Learning-based optimization of cache content in a small cell base station," in *2014 IEEE International Conference on Communications (ICC)*, Jun. 2014, pp. 1897–1903.

[56] M. Gregori, J. Gómez-Vilardebó, J. Matamoros, and D. Gündüz, "Wireless content caching for small cell and D2D networks," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 5, pp. 1222–1234, 2016.

[57] A. Khreishah, J. Chakareski, and A. Gharaibeh, "Joint caching, routing, and channel assignment for collaborative small-cell cellular networks," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 8, pp. 2275–2284, 2016.

[58] S. T. ul Hassan, M. Bennis, P. H. J. Nardelli, and M. Latva-Aho, "Modeling and analysis of content caching in wireless small cell networks," in *2015 International Symposium on Wireless Communication Systems (ISWCS)*, Aug. 2015, pp. 765–769.

[59] S. Tamoor-ul Hassan, M. Bennis, P. H. Nardelli, and M. Latva-Aho, "Caching in wireless small cell networks: A storage-bandwidth

tradeoff," *IEEE Communications Letters*, vol. 20, no. 6, pp. 1175–1178, 2016.

[60] X. Li, X. Wang, and V. C. M. Leung, "Weighted network traffic offloading in cache-enabled heterogeneous networks," in *2016 IEEE International Conference on Communications (ICC)*, May 2016, pp. 1–6.

[61] Y. Cui and D. Jiang, "Analysis and optimization of caching and multicasting in large-scale cache-enabled heterogeneous wireless networks," *IEEE Transactions on Wireless Communications*, vol. 16, no. 1, pp. 250–264, 2017.

[62] K. Shanmugam, N. Golrezaei, A. G. Dimakis, A. F. Molisch, and G. Caire, "FemtoCaching: Wireless content delivery through distributed caching helpers," *IEEE Transactions on Information Theory*, vol. 59, no. 12, pp. 8402–8413, Dec. 2013.

[63] T. Wang, L. Song, and Z. Han, "Dynamic femtocaching for mobile users," in *2015 IEEE Wireless Communications and Networking Conference (WCNC)*, Mar. 2015, pp. 861–865.

[64] R. Pabst, B. H. Walke, D. C. Schultz *et al.*, "Relay-based deployment concepts for wireless and mobile broadband radio," *IEEE Communications Magazine*, vol. 42, no. 9, pp. 80–89, 2004.

[65] X. Wang, Y. Bao, X. Liu, and Z. Niu, "On the design of relay caching in cellular networks for energy efficiency," in *2011 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, 2011, pp. 259–264.

[66] M. Erol-Kantarci, "Cache-at-relay: energy-efficient content placement for next-generation wireless relays," *International Journal of Network Management*, vol. 25, no. 6, pp. 454–470, 2015.

[67] A. Liu and V. K. N. Lau, "Cache-enabled opportunistic cooperative mimo for video streaming in wireless systems," *IEEE Transactions on Signal Processing*, vol. 62, no. 2, pp. 390–402, Jan. 2014.

[68] Y. Wang, X. Tao, X. Zhang, and G. Mao, "Joint caching placement and user association for minimizing user download delay," *IEEE Access*, vol. 4, pp. 8625–8633, 2016.

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

[70] J. Yao and N. Ansari, "QoS-aware joint BBU-RRH mapping and user association in Cloud-RANs," *IEEE Transactions on Green Communications and Networking*, vol. 2, no. 4, pp. 881–889, Dec. 2018.

[71] S.-H. Park, O. Simeone, and S. S. Shitz, "Joint optimization of cloud and edge processing for fog radio access networks," *IEEE Transactions on Wireless Communications*, vol. 15, no. 11, pp. 7621–7632, 2016.

[72] R. Tandon and O. Simeone, "Cloud-aided wireless networks with edge caching: Fundamental latency trade-offs in fog radio access networks," in *2016 IEEE International Symposium on Information Theory (ISIT)*, Jul. 2016, pp. 2029–2033.

[73] M. Tao, E. Chen, H. Zhou, and W. Yu, "Content-centric sparse multicast beamforming for cache-enabled cloud RAN," *IEEE Transactions on Wireless Communications*, vol. 15, no. 9, pp. 6118–6131, 2016.

[74] S. Mosleh, L. Liu, H. Hou, and Y. Yi, "Coordinated data assignment: A novel scheme for big data over cached Cloud-RAN," in *2016 IEEE Global Communications Conference (GLOBECOM)*, Dec. 2016, pp. 1–6.

[75] R. G. Stephen and R. Zhang, "Green OFDMA resource allocation in cache-enabled CRAN," in *2016 IEEE Online Conference on Green Communications (OnlineGreenComm)*, Nov. 2016, pp. 70–75.

[76] T. X. Tran and D. Pompili, "Octopus: A cooperative hierarchical caching strategy for cloud radio access networks," in *2016 IEEE 13th International Conference on Mobile Ad Hoc and Sensor Systems (MASS)*, Oct. 2016, pp. 154–162.

[77] J. Yao and N. Ansari, "Joint caching in fronthaul and backhaul constrained C-RAN," in *2017 IEEE Global Communications Conference (GLOBECOM)*, Dec. 2017.

[78] ——, "Joint content placement and storage allocation in C-RANs for IoT sensing service," *IEEE Internet of Things Journal*, vol. 6, no. 1, pp. 1060–1067, Feb. 2019.

[79] M. Chen, W. Saad, C. Yin, and M. Debbah, "Echo state networks for proactive caching in cloud-based radio access networks with mobile users," *IEEE Transactions on Wireless Communications*, vol. 16, no. 6, pp. 3520–3535, 2017.

[80] X. Wang, S. Leng, and K. Yang, "Social-aware edge caching in fog radio access networks," *IEEE Access*, vol. 5, pp. 8492–8501, 2017.

[81] K. Zheng, Q. Zheng, P. Chatzimisios, W. Xiang, and Y. Zhou, "Heterogeneous vehicular networking: A survey on architecture, challenges, and solutions," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 4, pp. 2377–2396, Fourth quarter 2015.

[82] F. Pantisano, M. Bennis, W. Saad, and M. Debbah, "In-network caching and content placement in cooperative small cell networks," in *1st International Conference on 5G for Ubiquitous Connectivity*, Nov. 2014, pp. 128–133.

[83] N. Golrezaei, P. Mansourifard, A. F. Molisch, and A. G. Dimakis, "Base-station assisted device-to-device communications for high-throughput wireless video networks," *IEEE Transactions on Wireless Communications*, vol. 13, no. 7, pp. 3665–3676, 2014.

[84] Z. Chen and M. Kountouris, "D2D caching vs. small cell caching: Where to cache content in a wireless network?" in *2016 IEEE 17th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, Jul. 2016, pp. 1–6.

[85] B. Blaszczyszyn and A. Giovanidis, "Optimal geographic caching in cellular networks," in *2015 IEEE International Conference on Communications (ICC)*, Jun. 2015, pp. 3358–3363.

[86] D. Liu and C. Yang, "Cache-enabled heterogeneous cellular networks: Comparison and tradeoffs," in *2016 IEEE International Conference on Communications (ICC)*, May 2016, pp. 1–6.

[87] B. Perabathini, E. Baştuğ, M. Kountouris, M. Debbah, and A. Conte, "Caching at the edge: A green perspective for 5G networks," in *2015 IEEE International Conference on Communication Workshop (ICCW)*, Jun. 2015, pp. 2830–2835.

[88] A. A. AlMomani, A. Argyriou, and M. Erol-Kantarci, "A heuristic approach for overlay content-caching network design in 5G wireless networks," in *2016 IEEE Symposium on Computers and Communication (ISCC)*, Jun. 2016, pp. 621–626.

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

[90] T. M. Nguyen, W. Ajib, and C. Assi, "Designing wireless backhaul heterogeneous networks with small cell buffering," *IEEE Transactions on Communications*, vol. 66, no. 10, pp. 4596–4610, Oct. 2018.

[91] H. Hsu and K. C. Chen, "A resource allocation perspective on caching to achieve low latency," *IEEE Communications Letters*, vol. 20, no. 1, pp. 145–148, 2016.

[92] J. Liu, B. Bai, J. Zhang, and K. B. Letaief, "Content caching at the wireless network edge: A distributed algorithm via belief propagation," in *2016 IEEE International Conference on Communications (ICC)*, May 2016, pp. 1–6.

[93] M. Dehghan, A. Seetharam, B. Jiang, T. He, T. Salonidis, J. Kurose, D. Towsley, and R. Sitaraman, "On the complexity of optimal routing and content caching in heterogeneous networks," in *2015 IEEE Conference on Computer Communications (INFOCOM)*, Apr. 2015, pp. 936–944.

[94] J. Rao, H. Feng, and Z. Chen, "Exploiting user mobility for D2D assisted wireless caching networks," in *2016 8th International Conference on Wireless Communications Signal Processing (WCSP)*, Oct. 2016, pp. 1–5.

[95] J. Yao and N. Ansari, "QoS-aware rechargeable UAV trajectory optimization for sensing service," in *IEEE International Conference on Communications (ICC)*, May 2019.

[96] H. Ahlehagh and S. Dey, "Video-aware scheduling and caching in the radio access network," *IEEE/ACM Transactions on Networking*, vol. 22, no. 5, pp. 1444–1462, 2014.

[97] J. Tadrous, A. Eryilmaz, and H. E. Gamal, "Proactive content download and user demand shaping for data networks," *IEEE/ACM Transactions on Networking (TON)*, vol. 23, no. 6, pp. 1917–1930, 2015.

[98] T. Hou, G. Feng, S. Qin, and W. Jiang, "Proactive content caching by exploiting transfer learning for mobile edge computing," in *2017 IEEE Global Communications Conference (GLOBECOM)*, Dec. 2017, pp. 1–6.

[99] A. Gharaibeh, A. Khreishah, B. Ji, and M. Ayyash, "A provably efficient online collaborative caching algorithm for multicell-coordinated systems," *IEEE Transactions on Mobile Computing*, vol. 15, no. 8, pp. 1863–1876, 2016.

[100] L. Marini, J. Li, and Y. Li, "Distributed caching based on decentralized learning automata," in *IEEE International Conference on Communications (ICC)*, 2015, pp. 3807–3812.

[101] F. R. Kschischang, B. J. Frey, and H. A. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Transactions on Information Theory*, vol. 47, no. 2, pp. 498–519, Feb. 2001.

[102] J. Li, Y. Chen, Z. Lin, W. Chen, B. Vucetic, and L. Hanzo, "Distributed caching for data dissemination in the downlink of heterogeneous networks," *IEEE Transactions on Communications*, vol. 63, no. 10, pp. 3553–3568, 2015.

[103] W. Jiang, G. Feng, and S. Qin, "Optimal cooperative content caching and delivery policy for heterogeneous cellular networks," *IEEE Transactions on Mobile Computing*, vol. 16, no. 5, pp. 1382–1393, May 2017.

[104] J. Song, M. Sheng, T. Q. S. Quek, C. Xu, and X. Wang, "Learning-based content caching and sharing for wireless networks," *IEEE Transactions on Communications*, vol. 65, no. 10, pp. 4309–4324, Oct. 2017.

[105] T. Ho and D. Lun, *Network coding: an introduction*. Cambridge University Press, 2008.

[106] M. A. Maddah-Ali and U. Niesen, "Fundamental limits of caching," *IEEE Transactions on Information Theory*, vol. 60, no. 5, pp. 2856–2867, 2014.

[107] ——, "Decentralized coded caching attains order-optimal memory-rate tradeoff," *IEEE/ACM Transactions on Networking*, vol. 23, no. 4, pp. 1029–1040, 2015.

[108] ——, "Coding for caching: fundamental limits and practical challenges," *IEEE Communications Magazine*, vol. 54, no. 8, pp. 23–29, 2016.

[109] N. Karamchandani, U. Niesen, M. A. Maddah-Ali, and S. N. Diggavi, "Hierarchical coded caching," *IEEE Transactions on Information Theory*, vol. 62, no. 6, pp. 3212–3229, 2016.

[110] Z. Hu, Z. Zheng, T. Wang, L. Song, and X. Li, "Game theoretic approaches for wireless proactive caching," *IEEE Communications Magazine*, vol. 54, no. 8, pp. 37–43, 2016.

[111] G. Demange, D. Gale, and M. Sotomayor, "Multi-item auctions," *Journal of Political Economy*, vol. 94, no. 4, pp. 863–872, 1986.

[112] K. Hamidouche, W. Saad, M. Debbah, J. B. Song, and C. S. Hong, "The 5G cellular backhaul management dilemma: To cache or to serve," *IEEE Transactions on Wireless Communications*, vol. 16, no. 8, pp. 4866–4879, Aug. 2017.

[113] K. Hamidouche, W. Saad, and M. Debbah, "Many-to-many matching games for proactive social-caching in wireless small cell networks," in *2014 12th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*, May 2014, pp. 569–574.

[114] J. Yao and N. Ansari, "Reliability-aware fog resource provisioning for deadline-driven IoT services," in *2018 IEEE Global Communications Conference (GLOBECOM)*, Dec. 2018.

[115] X. Sun and N. Ansari, "EdgeIoT: Mobile edge computing for the internet of things," *IEEE Communications Magazine*, vol. 54, no. 12, pp. 22–29, Dec. 2016.

[116] J. Yao and N. Ansari, "QoS-aware fog resource provisioning and mobile device power control in IoT networks," *IEEE Transactions on Network and Service Management*, vol. 16, no. 1, pp. 167–175, Mar. 2019.

[117] S. Vural, P. Navaratnam, N. Wang, C. Wang, L. Dong, and R. Tafazolli, "In-network caching of internet-of-things data," in *2014 IEEE International Conference on Communications (ICC)*, Jun. 2014, pp. 3185–3190.

[118] D. Niyato, D. I. Kim, P. Wang, and L. Song, "A novel caching mechanism for internet of things (IoT) sensing service with energy harvesting," in *2016 IEEE International Conference on Communications (ICC)*, May 2016, pp. 1–6.

[119] J. Yao and N. Ansari, "Caching in energy harvesting aided internet of things: A game-theoretic approach," *IEEE Internet of Things Journal*, DOI: 10.1109/JIOT.2018.2880483, early access.

[120] H. Che, Y. Tung, and Z. Wang, "Hierarchical web caching systems: modeling, design and experimental results," *IEEE Journal on Selected Areas in Communications*, vol. 20, no. 7, pp. 1305–1314, Sep. 2002.

[121] R. Urgaonkar, S. Wang, T. He, M. Zafer, K. Chan, and K. K. Leung, "Dynamic service migration and workload scheduling in edge-clouds," *Performance Evaluation*, vol. 91, pp. 205–228, 2015.

[122] G. Li, M. Wang, J. Feng, L. Xu, B. Ramamurthy, W. Li, and X. Guan, "Understanding user generated content characteristics: A hot-event perspective," in *2011 IEEE International Conference on Communications (ICC)*, Jun. 2011, pp. 1–5.

[123] A. Tatar, M. D. de Amorim, S. Fdida, and P. Antoniadis, "A survey on predicting the popularity of web content," *Journal of Internet Services and Applications*, vol. 5, no. 1, p. 8, 2014.

[124] D. Niu, Z. Liu, B. Li, and S. Zhao, "Demand forecast and performance prediction in peer-assisted on-demand streaming systems," in *2011 Proceedings IEEE INFOCOM*, Apr. 2011, pp. 421–425.

[125] Z. Wang, L. Sun, C. Wu, and S. Yang, "Enhancing internet-scale video service deployment using microblog-based prediction," *IEEE Transactions on Parallel and Distributed Systems*, vol. 26, no. 3, pp. 775–785, 2015.

[126] M. Rowe, "Forecasting audience increase on youtube," in *Workshop on User Profile Data on the Social Semantic Web, 8th Extended Semantic Web Conference 2011 (ESWC 2011)*, May 2011.

[127] M. Leconte, G. Paschos, L. Gkatzikis, M. Draief, S. Vassilaras, and S. Chouvardas, "Placing dynamic content in caches with small population," in *IEEE INFOCOM 2016 - The 35th Annual IEEE International Conference on Computer Communications*, Apr. 2016, pp. 1–9.

[128] J. Famaey, F. Iterbeke, T. Wauters, and F. De Turck, "Towards a predictive cache replacement strategy for multimedia content," *Journal of Network and Computer Applications*, vol. 36, no. 1, pp. 219–227, 2013.

[129] L. Breslau, C. Pei, F. Li, G. Phillips, and S. Shenker, "Web caching and Zipf-like distributions: evidence and implications," in *INFOCOM '99 - Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies*, vol. 1, 1999, pp. 126–134.

[130] J. Liu, F. Liu, and N. Ansari, "Monitoring and analyzing big traffic data of a large-scale cellular network with Hadoop," *IEEE Network*, vol. 28, no. 4, pp. 32–39, Jul. 2014.

[131] 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, pp. 36–42, 2016.

[132] H. Nakayama, Z. M. Fadlullah, N. Ansari, and N. Kato, "A novel scheme for WSAN sink mobility based on clustering and set packing techniques," *IEEE Transactions on Automatic Control*, vol. 56, no. 10, pp. 2381–2389, Oct. 2011.

[133] H. Nishiyama, T. Ngo, N. Ansari, and N. Kato, "On minimizing the impact of mobility on topology control in mobile ad hoc networks," *IEEE Transactions on Wireless Communications*, vol. 11, no. 3, pp. 1158–1166, Mar. 2012.

[134] K. Poularakis and L. Tassiulas, "Exploiting user mobility for wireless content delivery," in *2013 IEEE International Symposium on Information Theory*, 2013, pp. 1017–1021.

[135] D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A.-L. Barabasi, "Human mobility, social ties, and link prediction," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, 2011, pp. 1100–1108.

[136] M. Musolesi, S. Hailes, and C. Mascolo, "An ad hoc mobility model founded on social network theory," in *Proceedings of the 7th ACM International Symposium on Modeling, Analysis and Simulation of Wireless and Mobile Systems*. New York, NY, USA: ACM, 2004, pp. 20–24.

[137] R. Wang, X. Peng, J. Zhang, and K. B. Letaief, "Mobility-aware caching for content-centric wireless networks: modeling and methodology," *IEEE Communications Magazine*, vol. 54, no. 8, pp. 77–83, 2016.

[138] C. Bettstetter, H. Hartenstein, and X. Pérez-Costa, "Stochastic properties of the random waypoint mobility model," *Wireless Networks*, vol. 10, no. 5, pp. 555–567, 2004.

[139] J.-K. Lee and J. C. Hou, "Modeling steady-state and transient behaviors of user mobility: formulation, analysis, and application," in *Proceedings of the 7th ACM international symposium on Mobile ad hoc networking and computing*, 2006, pp. 85–96.

[140] Q. Lv, Y. Qiao, N. Ansari, J. Liu, and J. Yang, "Big data driven hidden markov model based individual mobility prediction at points of interest," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 6, pp. 5204–5216, Jun. 2017.

[141] K. Poularakis, G. Iosifidis, and L. Tassiulas, "Approximation algorithms for mobile data caching in small cell networks," *IEEE Transactions on Communications*, vol. 62, no. 10, pp. 3665–3677, 2014.

[142] K. Poularakis, G. Iosifidis, A. Argyriou, and L. Tassiulas, "Video delivery over heterogeneous cellular networks: Optimizing cost and performance," in *IEEE INFOCOM 2014 - IEEE Conference on Computer Communications*, Apr. 2014, pp. 1078–1086.

[143] F. Pantisano, M. Bennis, W. Saad, and M. Debbah, "Cache-aware user association in backhaul-constrained small cell networks," in *12th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*, 2014, pp. 37–42.

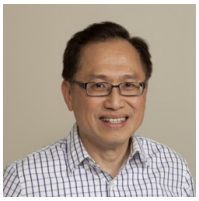
[144] T. Han and N. Ansari, "Network utility aware traffic load balancing in backhaul-constrained cache-enabled small cell networks with hybrid power supplies," *IEEE Transactions on Mobile Computing*, vol. 16, no. 10, pp. 2819–2832, Oct. 2017.

[145] J. Song, H. Song, and W. Choi, "Optimal caching placement of caching system with helpers," in *2015 IEEE International Conference on Communications (ICC)*, Jun. 2015, pp. 1825–1830.

[146] X. Peng, J.-C. Shen, J. Zhang, and K. B. Letaief, "Backhaul-aware caching placement for wireless networks," in *2015 IEEE Global Communications Conference (GLOBECOM)*, 2015, pp. 1–6.
[147] K. Poularakis, G. Iosifidis, V. Sourlas, and L. Tassiulas, "Multicast-aware caching for small cell networks," in *2014 IEEE Wireless Communications and Networking Conference (WCNC)*, 2014, pp. 2300–2305.
[148] J. Liao, K. Wong, Y. Zhang, Z. Zheng, and K. Yang, "Coding, multicast, and cooperation for cache- enabled heterogeneous small cell networks," *IEEE Transactions on Wireless Communications*, vol. 16, no. 10, pp. 6838–6853, Oct. 2017.
[149] B. Zhou, Y. Cui, and M. Tao, "Stochastic content-centric multicast scheduling for cache-enabled heterogeneous cellular networks," *IEEE Transactions on Wireless Communications*, vol. 15, no. 9, pp. 6284–6297, 2016.
[150] B. Zhou, Y. Cui, and M. Tao, "Optimal dynamic multicast scheduling for cache-enabled content-centric wireless networks," *IEEE Transactions on Communications*, vol. 65, no. 7, pp. 2956–2970, Jul. 2017.
[151] Z. Xia, J. Yan, and Y. Liu, "Cooperative content delivery in multicast multihop device-to-device networks," *IEEE Access*, vol. 5, pp. 6314–6324, 2017.
[152] M. Z. Shafiq, A. X. Liu, and A. R. Khakpour, "Revisiting caching in content delivery networks," in *ACM SIGMETRICS Performance Evaluation Review*, vol. 42, 2014, pp. 567–568.
[153] R. Pedarsani, M. A. Maddah-Ali, and U. Niesen, "Online coded caching," *IEEE/ACM Transactions on Networking*, vol. 24, no. 2, pp. 836–845, 2016.
[154] S. Li, J. Xu, M. v. d. Schaar, and W. Li, "Popularity-driven content caching," in *IEEE INFOCOM 2016 - The 35th Annual IEEE International Conference on Computer Communications*, 2016, pp. 1–9.
[155] J. Yao and N. Ansari, "Energy-aware task allocation for mobile IoT by online reinforcement learning," in *IEEE International Conference on Communications (ICC)*, May 2019.
[156] X. Sun and N. Ansari, "Traffic load balancing among brokers at the IoT application layer," *IEEE Transactions on Network and Service Management*, vol. 15, no. 1, pp. 489–502, Mar. 2018.
[157] J. Quevedo, D. Corujo, and R. Aguiar, "A case for ICN usage in IoT environments," in *2014 IEEE Global Communications Conference (GLOBECOM)*, Dec. 2014, pp. 2770–2775.
[158] X. Sun and N. Ansari, "Dynamic resource caching in the IoT application layer for smart cities," *IEEE Internet of Things Journal*, vol. 5, no. 2, pp. 606–613, Apr. 2018.
[159] M. Mukherjee, R. Matam, L. Shu, L. Maglaras, M. A. Ferrag, N. Choudhury, and V. Kumar, "Security and privacy in fog computing: Challenges," *IEEE Access*, vol. 5, pp. 19 293–19 304, 2017.
[160] D. Kim, J. Bi, A. V. Vasilakos, and I. Yeom, "Security of cached content in NDN," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 12, pp. 2933–2944, Dec. 2017.
[161] J. Leguay, G. S. Paschos, E. A. Quaglia, and B. Smyth, "CryptoCache: Network caching with confidentiality," in *2017 IEEE International Conference on Communications (ICC)*, May 2017, pp. 1–6.
[162] C. Kolias, G. Kambourakis, and S. Gritzalis, "Attacks and countermeasures on 802.16: Analysis and assessment," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 1, pp. 487–514, 2013.
[163] J. H. Abawajy, M. I. H. Ninggal, and T. Herawan, "Privacy preserving social network data publication," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 3, pp. 1974–1997, 2016.
[164] T. Wang, Z. Zheng, M. H. Rehmani, S. Yao, and Z. Huo, "Privacy preservation in big data from the communication perspective—a survey," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 1, pp. 753–778, First quarter 2019.
[165] W. Xia, Y. Wen, C. H. Foh, D. Niyato, and H. Xie, "A survey on software-defined networking," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 1, pp. 27–51, 2015.
[166] Q. Duan, N. Ansari, and M. Toy, "Software-defined network virtualization – an architectural framework for integrating SDN and NFV for service provisioning in future networks," *IEEE Network*, vol. 30, no. 5, pp. 10–16, Sep. 2016.
[167] I. T. Haque and N. Abu-Ghazaleh, "Wireless software defined networking: A survey and taxonomy," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 4, pp. 2713–2737, Fourth Quarter, 2016.
[168] T. Wu, "Network neutrality, broadband discrimination," *Journal of Telecommunications and High Technology Law*, vol. 2, p. 141, 2003.
[169] D. Miorandi, I. Carreras, E. Gregori, I. Graham, and J. Stewart, "Measuring net neutrality in mobile internet: Towards a crowdsensing-based citizen observatory," in *IEEE International Conference on Communications Workshops (ICC)*, Jun. 2013, pp. 199–203.

Jingjing Yao (S17) received the B.E. degree in information and communication engineering from Dalian University of Technology (DUT) and the M.E. degree in information and communication engineering from University of Science and Technology of China (USTC). She is currently working towards the Ph.D. degree in Computer Engineering at the New Jersey Institute of Technology (NJIT), Newark, New Jersey. Her research interests include cloud computing, cloud radio access networks, and Internet of Things.

Tao Han (S08-M15) received his Ph.D. in Electrical Engineering from New Jersey Institute of Technology (NJIT), Newark, NJ, USA. He is currently an Assistant Professor in the Department of Electrical and Computer Engineering at the University of North Carolina at Charlotte, Charlotte, NC, USA. He serves as an Associate Editor of IEEE Communications Letters. His research interest includes mobile edge networking, mobile X reality, 5G, Internet of Things, and smart grid.



Nirwan Ansari (S78-M83-SM94-F09) is Distinguished Professor of Electrical and Computer Engineering at the New Jersey Institute of Technology (NJIT). He has also been a visiting (chair) professor at several universities.

He authored Green Mobile Networks: A Networking Perspective (Wiley-IEEE, 2017) with T. Han, and co-authored two other books. He has also (co-)authored more than 600 technical publications, over 250 published in widely cited journals/magazines. He has guest-edited a number of special issues covering various emerging topics in communications and networking. He has served on the editorial/advisory board of over ten journals. His current research focuses on green communications and networking, cloud computing, drone-assisted networking, and various aspects of broadband networks.

He was elected to serve in the IEEE Communications Society (ComSoc) Board of Governors as a member-at-large, has chaired some ComSoc technical and steering committees, has been serving in many committees such as the IEEE Fellow Committee, and has been actively organizing numerous IEEE International Conferences/Symposia/Workshops. He has frequently been delivering keynote addresses, distinguished lectures, tutorials, and invited talks. Some of his recognitions include several Excellence in Teaching Awards, a few best paper awards, the NCE Excellence in Research Award, the ComSoc TC-CSR Distinguished Technical Achievement Award, the ComSoc AHSN TC Technical Recognition Award, the IEEE TCGCC Distinguished Technical Achievement Recognition Award, the NJ Inventors Hall of Fame Inventor of the Year Award, the Thomas Alva Edison Patent Award, Purdue University Outstanding Electrical and Computer Engineering

Award, NCE 100 Medal, and designation as a COMSOC Distinguished Lecturer. He has also been granted 38 U.S. patents.

He received a Ph.D. from Purdue University in 1988, an MSEE from the University of Michigan in 1983, and a BSEE (summa cum laude with a perfect GPA) from NJIT in 1982.