IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 19, 2017, accepted March 9, 2017, date of publication March 21, 2017, date of current version June 7, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2685434

A Survey on Mobile Edge Networks: Convergence of Computing, Caching and Communications

SHUO WANG[1], XING ZHANG[1], (Senior Member, IEEE), YAN ZHANG[2], (Senior Member, IEEE), LIN WANG[1], JUWO YANG[1], AND WENBO WANG[1], (Senior Member, IEEE)

[1]School of Information and Communications Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China
[2]University of Oslo, Oslo 1364, Norway

Corresponding authors: Xing Zhang (zhangx@ieee.org) and Yan Zhang (yanzhang@ieee.org)

This work was supported in part by the National Science Foundation of China under Grant 61372114, Grant 61571054, and Grant 61631005 in part by the New Star in Science and Technology of Beijing Municipal Science and Technology Commission through Beijing Nova Program under Grant Z151100000315077.

ABSTRACT As the explosive growth of smart devices and the advent of many new applications, traffic volume has been growing exponentially. The traditional centralized network architecture cannot accommodate such user demands due to heavy burden on the backhaul links and long latency. Therefore, new architectures, which bring network functions and contents to the network edge, are proposed, i.e., mobile edge computing and caching. Mobile edge networks provide cloud computing and caching capabilities at the edge of cellular networks. In this survey, we make an exhaustive review on the state-of-the-art research efforts on mobile edge networks. We first give an overview of mobile edge networks, including definition, architecture, and advantages. Next, a comprehensive survey of issues on computing, caching, and communication techniques at the network edge is presented. The applications and use cases of mobile edge networks are discussed. Subsequently, the key enablers of mobile edge networks, such as cloud technology, SDN/NFV, and smart devices are discussed. Finally, open research challenges and future directions are presented as well.

INDEX TERMS Mobile edge computing, mobile edge caching, D2D, SDN, NFV, content delivery, computational offloading.

I. INTRODUCTION

During the past several decades, mobile cellular networks have been evolving steadily and significantly from the 1st generation (1G) voice only systems to current 4th generation (4G) all-IP based LTE-Advanced networks. The system capacity and average data rate have improved greatly with the technology advancements in physical layer such as WCDMA, OFDMA, MIMO, CoMP and in network layer such as heterogeneous network (HetNet) and cloud radio access network (C-RAN). According to a recent report from Cisco [1], the mobile data traffic has grown 4000-fold during the past 10 years and will continue grow at a rate of 53 percent annually from 2015 to 2020. In particular, mobile video traffic accounts for more than half of total mobile data traffic and this percentage keeps increasing. Besides, mobile devices are getting smarter in their computing capabilities, and new machine type devices appear such as wearable devices and sensors in addition to human type devices. This leads to massive M2M connections in next generation mobile networks.

Machine type communications (MTC) bring a wide range of new applications and services in wireless networks. Shariatmadari *et al.* [2] presented the current status and challenges of MTC for cellular systems. The most important challenges include massive number of MTC devices, small data bursts, low-latency, and low power consumption. Various solutions have been proposed to accommodate these challenges [3], [4]. Since the processing capabilities of MTC devices are constrained, one promising solution is to offload their tasks to places that have powerful processing capabilities. The ubiquitous connectivity of MTC leads to the strong heterogeneous networking paradigm. Research efforts have been made to accommodate such MTC applications from 4G to the emerging 5G systems [5].

The preliminary mobile computing scheme adopted a 2-level hierarchy which originally called ''servers'' and ''clients'' [6]. Later on, The terminology ''cloud'' was used to represent a collection of servers with computational and information resources, which leads to the research on mobile

cloud computing (MCC). Mobile cloud computing considers various mobile-related factors compared to the traditional computation offloading techniques, such as device energy, bandwidth utilization cost, network connectivity, mobility, context awareness and location awareness [7], [8]. Various survey articles have been published focusing on different aspects of MCC. Guan *et al.* [9] and Dinh *et al.* [10] presented generic issues on mobile cloud computing including architecture, technical challenges and applications. In [11], existing works on mobile cloud platforms and access schemes were discussed. The authors compared two mobile cloud platforms, the Hyrax platform [12] and virtual machine (VM) based cloudlets [13], and then reviewed intelligent access schemes utilizing the user's location and context [14]. The authors in [7] elaborated the entities affecting computation offloading decision and presented detailed application models classification and the latest mobile cloud application models. Fernando *et al.* in [15] presented a detailed taxonomy of mobile cloud computing based on the key issues and the approaches to tackle them, such as operational issues, end user issues, service level issues, security, context-awareness and data management. User authentication is significant in securing cloud-based computing and communications. Alizadeh *et al.* [16] surveyed the state-of-the-art authentication mechanism in MCC and compare it with that in cloud computing. The merits of MCC can be summarized as follows. Firstly, it can provide sufficient resources for mobile devices and has great flexibility. Secondly, the cost of MCC can be reduced due to centralized management of resources. Finally, since all the tasks are processed in the cloud, MCC support multiple platforms. Fig. 1 illustrates the general architecture of MCC, which contains 2 tiers: the cloud and the mobile devices.

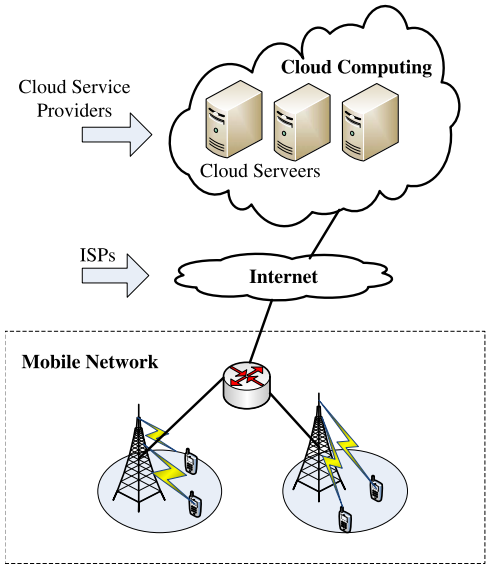

FIGURE 1. Architecture of mobile cloud computing.

Despite the merits of mobile cloud computing, it faces the inevitable problems such as long latency and backhaul bandwidth limitation due to the long distance from mobile devices to the Internet cloud. Mobile Edge computing (MEC), which deploy cloud servers in base stations, is a promising solution for the problem since the computation capability is closer to the mobile devices [17]. The benefits of MEC consist of low latency, proximity, high bandwidth, real time radio network information and location awareness. MEC is recognized as one of the key technologies for the next generation 5G networks by the European 5G PPP (5G Infrastructure Public Private Partnership) [18]. Ahmed and Ahmed in [19] described the taxonomy of MEC based on different aspects including its characteristics, access technologies, applications, objectives and so on. They also identified some of the open issues in MEC. Beck *et al.* [20] classified several applications deployed at the mobile edge according to the technical metrics: power consumption, delay, bandwidth usage and scalability. The benefits of MEC each stakeholder can get were also analyzed. For mobile users and network operators, the main profit is reduced delay which leads to fast services, For the application service providers (ASP), they can benefit from the utilization of user-related information.

Fog computing is another edge paradigm which supports the future internet of things (IoT) applications [21]. It uses near-user edge devices such as edge routers to carry out substantial amount of computing tasks. Although it is a concept similar to MEC in some aspects, it distinguishes itself as more suitable for the context of IoT [22]. The OpenFog Consortium was founded by Cisco Systems, ARM Holdings, Dell, Intel, Microsoft and Princeton University in 2015 to promote development and interests in fog computing. The representative applications and various aspects of research issues of fog computing were highlighted in [22]. Three driving scenarios which will benefit from fog computing are augmented reality and real-time video analytics, content delivery and mobile big data analytics. The quality of service (QoS) metrics for fog services include four aspects: connectivity, reliability, capacity and delay. The interfacing and programming model, resource management, security and privacy are key challenges fog computing encounters. Yi *et al.* [23] surveyed the new security and privacy challenges fog computing faces in addition to those same as cloud computing, such as secure data storage and secure computation. The security threats and challenges in the edge paradigms were analyzed in [24]. The authors presented specific challenges and promising solutions in the following aspects: identity and authentication, access control system, protocol and network security, trust management, intrusion detection system, privacy and virtualization.

A similar edge computing concept proposed by the academia is called Cloudlet [13], [25]. The cloudlet is an extension of distant "cloud". It is a "data center in a box" which is self-managed, energy efficient and simple to deploy on a business premises such as a coffee shop or an office room. One approach for deploying cloudlet infrastructure is to integrate cloudlet and WiFi access point hardware into a single entity. However, the management of widespread

deployed cloudlet is challenging. The solution is transient customization of cloudlet infrastructure using hardware VM (virtual machine) technology [25].

The increasing demand for massive multimedia services over mobile cellular network poses great challenges on network capacity and backhaul links. The emergence of mobile edge caching and delivery techniques are promising solutions to cope with those challenges [26]. In traditional centralized mobile network architecture, content requests of end users are served by remote internet content providers. In that situation, duplicate traffic has to be transmitted through the whole mobile network, which leads to network congestion and a waste of network resources. Caching popular contents at the network edge (e.g., gateways, base stations and end user devices) can avoid duplicate transmissions of same content and improve users' quality of experiences due to reduced latency. The concept of caching is not new in cellular networks. It has been used in web caching and information centric networks. Ali *et al.* [27] investigated the web caching and prefetching techniques in improving the web performance. A classification of caching policies was presented such as recency-based policies (e.g., LRU), frequency-based policies (e.g. LFU), size-based policies, function-based policies and randomized policies. Furthermore, Podlipnig and Böszörmenyi [28] described the advantages and disadvantages of cache replacement strategies and outlined potential research topics in modern proxy caches. Zhang *et al.* [29] and Fang *et al.* [30] surveyed caching mechanisms in information centric networks and its energy efficiency of caching respectively. However, due to the characteristics of wireless cellular networks, the above mentioned caching techniques cannot be directly applied. The caching schemes in cellular networks should be thoroughly investigated.

Despite the merits of the above edge paradigms, they make the wireless network heterogeneous and difficult to manage in the traditional way. Emerging technologies such as network function virtualization (NFV) and software defined networks (SDN) are promising solutions which enable the network to be flexible and easy to maintain. Network function virtualization is a recently proposed network architecture that utilizes IT virtualization technologies to virtualize network node functions on top of standard general purpose hardware, which changes the communication network infrastructure to be more cost-efficient [31]. Software defined network is a computing network architecture that split the function of control plane and data plane. In legacy mobile networks, both the control plane and user plane (or data plane) are integrated on the macro base stations. This network architecture cannot meet the explosive traffic and connection growth in future 5G networks. Based on the concept of SDN, new architectures are proposed. Zhang *et al.* [32] proposed a macro-assisted data-only carrier scheme which split the control function and data function to macro base stations (MBSs) and small base stations (SBSs), respectively. Pentikousis *et al.* [33] introduced and validated a software-defined mobile network architecture which improved the capability of the operator and reduced time to market for new services. A summary of existing survey articles on mobile computing and caching is shown in Table 1.

TABLE 1. Summary of existing survey articles on mobile computing and caching.

Aspects	Survey Papers	Contributions
Mobile Cloud Computing	[7]	A classification of application models and investigation of latest mobile cloud application models
	[9]	A summary of challenges for MCC, application partition and offloading technologies, classification of contexts and context management methods.
	[10]	An overview of MCC definition, architecture, and applications, as well as the generic issues and exi sting solutions. Discussions of the future research directions of MCC.
	[11]	An investigation of existing works on representative platforms and intelligent access schemes of MCC.
	[15]	A detailed taxonomy of mobile cloud computing based on the key issues and the approaches to tackle them.
	[16]	A comprehensive survey of the state-of-the-art authentication mechanism in MCC and comparison with that in cloud computing.
Mobile Edge Computing	[19]	A taxonomy of MEC based on different aspects including its characteristics, access technologies, applications, objectives and so on. Identification of some of the open issues in MEC.
	[20]	A classification of applications deployed at the mobile edge according to the technical metrics and the benefits of MEC for stakeholders in the network.
	[24]	A discussion of the security threats and challenges in the edge paradigms, as well as the promising solution for each specific challenge.
Fog Computing	[22]	Highlighting the representative applications and various aspects of research issues of fog computing.
	[23]	Survey of the new security and privacy challenges fog computing faces in addition to those same as cloud computing.
Caching	[27]	An investigation of the web caching and prefetching techniques in improving the web performance as well as a classification of caching policies.
	[28]	A description of the advantages and disadvantages of cache replacement strategies, outlining potential research topics in modern proxy caches.
	[29]	A survey of caching mechanisms in information centric networks.
	[30]	A survey of the energy efficiency of caching in information centric networks.

The above mentioned advantages and progress of mobile edge networks motivate us to perform a comprehensive

literature survey. The main contributions of this article are summarized as follows:

1) A comprehensive survey of mobile edge network architectures is presented, including MEC, Fog Computing, Cloudlets and edge caching. A comparison of different edge computing proposals is summarized. The advantages of mobile edge networks are pointed out.
2) A comprehensive survey of the key technologies in mobile edge computing and caching is presented. In particular, as for computing at the network edge, the literatures related to computation offloading, cooperation between the edge and core network, combination with 5G and proposed platforms are elaborated. As for edge caching, research progress on content popularity, caching policies, scheduling, and mobility management are surveyed.
3) The applications and use cases of mobile edge networks are comprehensively summarized. The key enablers of mobile edge networks are pointed out, including cloud technology, software defined network, network function virtualization and smarter mobile devices.
4) The open issues and challenges related to mobile edge networks are identified, such as network heterogeneity, realtime analytics, pricing, scalability, utilization of wireless big data, context awareness and so on.

The rest of the paper is organized as follows. In Section II, an overview of mobile edge networks including the definition, architecture and advantages is presented. In Section III, an elaborated survey of literatures on computing related issues at mobile edge networks is given. In Section IV, the research efforts on edge caching are fully surveyed. The advances in communication techniques with synergy of computing and caching is discussed in Section V. In Section VI, The applications and use cases of mobile edge networks are explained. In Section VII, the key enabler technologies of mobile edge networks are summarized. The open challenges and future directions are shown in Section VIII. Finally, conclusions are drawn in Section IX. For convenience, a summary of all abbreviations is shown in Table 2.

II. OVERVIEW OF MOBILE EDGE NETWORKS

The evolvement of mobile cellular networks has experienced 4 generations in the last two decades with the advancements in information and telecommunications technology. At the same time, users' demands for mobile networks also become more and more strict such as ultra high data rate and extremely low latency. Moreover, various new requirements are appearing due to the advent of new kinds of smart devices and new applications, such as virtual reality and the internet of things. The traditional base station centric network architecture cannot fulfill these requirements any more. The mobile cellular network architecture is evolving from BS-centric to device centric [34] and content centric network in the future 5G system, where the center of gravity moves

TABLE 2. Summary of abbreviations.

5G	5th generation
5GPPP	5G Infrastructure Public Private Partnership
API	application interface
AR	augmented reality
ASE	area spectral efficiency
ASP	application service provider
BS	base station
C-RAN	cloud radio access network
CAPEX	capital expenditures
CDN	content delivery network
CoMP	coordinated multiple point
CSI	channel state information
D2D	device-to-device
DC	data center
EPC	evolved packet core
ETSI	European Telecommunications Standards Institute
HetNet	heterogeneous network
IA	interference alignment
ICN	information centric network
IoT	internet of things
IRM	independent reference model
ISP	internet service provider
IP	internet protocol
LFU	least frequently used
LRU	least recently used
LTE	long term evolution
M2M	machine to machine
MBS	macro base station
MCC	mobile cloud computing
MEC	mobile edge computing
MEN	mobile edge networks
MIMO	multiple in multiple out
MPV	most popular video
MTC	machine type communication
Multi-RAT	multiple radio access technology
NFV	network function vitualization
OFDMA	orthogonal frequency division multiple access
OPEX	operating expenses
QoE	quality of experience
RACS	radio application cloud servers
RAN	radio access network
RNC	radio network controller
RTT	round trip time
SBS	small base station
SDN	software defined network
SINR	signal to interference plus noise ratio
SNM	shot noise model
SVC	scalable video coding
TDMA	time division multiple access
UGC	user generated content
UPP	user preference profile
VM	virtual machine
VR	virtual reality
WCDMA	wideband code division multiple access

from the network core to the edge [35]. In this section, we will firstly explain what is mobile edge networks. Then the architecture of mobile edge networks is presented. At last, we discuss the advantages of mobile edge networks.

A. WHAT IS MOBILE EDGE NETWORKS

The core idea of mobile edge networks is to move network functions, contents and resources closer to end users, i.e., the network edge, by utilizing SDN and NFV technologies. The network resources mainly include computing, storage or caching, and communication resources. While in some literature caching is included in computing resources [20], in this paper we discuss them separately since the service types and problems they are related to are different.

Edge computing in mobile networks is evolved from mobile cloud computing, which is an architecture moving the computing power and data storage away from mobile devices and into the cloud to leverage the powerful computing and storage capability of cloud platform [10]. However, mobile cloud computing faces several challenges such as long latency and high backhaul bandwidth consumption, therefore it is not suitable for real-time applications. Dinh *et al.* [10] list the technical challenges that MCC faces in detail. In the mobile communication side, the challenges include low bandwidth, service availability and heterogeneity due to the characteristics of wireless networks such as scarce radio resources, traffic congestion, and multiple radio access technologies (multi-RAT). In the computing side, efficient and dynamic computing offloading under environment changes is challenging, as well as the security issues for users and data, efficiency of data access and context awareness. In comparison, edge computing enables the network edge to have cloud computing capabilities.

Three different edge computing schemes have been proposed by industrial and academic parties: mobile edge computing, fog computing and cloudlet. Mobile edge computing was proposed by the standards organization European Telecommunications Standards Institute (ETSI) [17]. It is based on a virtualized platform which enables applications running at the network edge. Meanwhile, the infrastructure of NFV can be reused by applications which is beneficial for network operators. MEC servers can be deployed at various location at the network edge such as the LTE macro base station (eNodeB), at the 3G Radio Network Controller (RNC), and at an aggregation point. The deployment location may be affected by scalability, physical constraints, performance criteria and so on. MEC applications can be seamless deployed on different MEC platform intelligently and flexibly based on the technical parameters such as latency, required resources, availability, scalability and cost. The objective of ETSI MEC is to deliver a standard architecture and industry-standardized APIs for 3rd party aplications [36].

Fog computing is another edge computing architecture with the aim to accommodate IoT applications originally proposed by Cisco [36]. Fog computing is an extension of the cloud computing paradigm to the wireless network edge [21]. The name *fog computing* comes from the analogy that the fog is closer to people than the clouds. Similarly, the distance of an IoT device is closer to a fog computing platform than to lage-scale data centers. The necessity of fog computing maintained by Cisco is that a 2-tiered deployment of IoT applications is not sufficient for the requirements of low latency, mobility, and location awareness [36]. The solution is a multi-tiered architecture which deploys an intermediate fog platform between the device and the main cloud. The main characteristic of fog computing is that is is a completely distributed, multi-layer cloud computing architecture where the fog nodes are deployed in different network tiers [37].

The concept of Cloudlet is developed by an academic team at Carnegie Mellon University [25]. It can be deployed in both Wi-Fi networks and cellular networks. The key features of cloudlets are near-real-time provisioning of applications to edge nodes and handoff of virtual machine images among edge nodes when a device moves [36].

Mobile edge caching is proposed for tackling the challenges of massive content delivery in future mobile networks. The advances in storage enable the network to exploit the large amount of low cost storage resources at different places in the network. The traffic load of cellular network is dynamically varying in the spatial and temporal domain [38]. Proactive caching is an approach that exploits such traffic dynamicity by proactively cache popular content during off-peak periods, which reduces peak traffic demands [39]. Since the cache units are deployed at the network edge, a lot of information can be exploited to improve the caching efficiency. For example, social structures of users can be leveraged to cache and disseminate content via D2D communications.

Based on the above discussion, we define mobile edge networks as: *"A mobile network architecture that deploys and utilizes flexible computing and storage resources at the mobile network edge, including the radio access network, edge routers, gateways and mobile devices etc., with the help of SDN and NFV technologies"*.

B. ARCHITECTURE OF MOBILE EDGE NETWORKS

The mobile edge networks introduce new way of manipulating computing and storage resources. Both industries and the academia bring up proposals on the architecture of MEN. We will present the specific architectures including ETSI MEC, Fog Computing, Cloudlet and Edge Caching. Then we will summarize these proposals and give the general architecture of MEN.

1) MOBILE EDGE COMPUTING

Mobile edge computing has drawn much attention of industries and the academia. In industries, the ETSI has launched an Industry Specification Group (ISG) on MEC in December 2014. The ISG produces specifications what enable the hosting of 3rd-party innovative applications in a standard MEC environment [17]. The group has delivered several specifications on service scenarios, requirements,

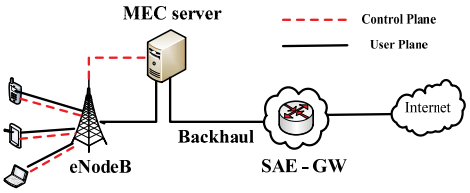

FIGURE 2. Architecture of MEC.

architecture and APIs. Fig. 2 shows the architecture of MEC. MEC servers are located in proximity of base stations. They can either handle a user request and respond directly to the UE or forward the request to remote data centers and content distribution networks (CDNs) [40].

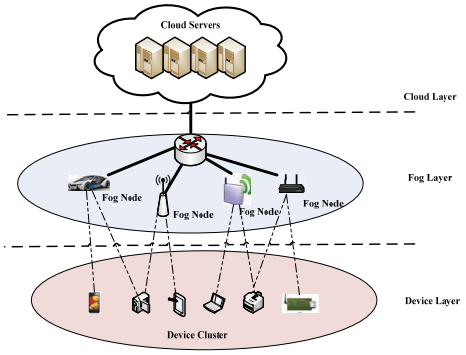

FIGURE 3. Architecture of Fog computing.

2) FOG COMPUTING

Fog computing is a platform designed mainly for Internet of Things use cases. Its component fog nodes are massively distributed in wide area. The main feature of fog is that it utilizes collaborations among multiple end user clients or near-user edge devices to help processing and storage of mobiled devices [41]. Compared to Cloud, Fog has advantages in three dimensions: exploiting storage, computing and control functions, communication and networking at or near the end user [42]. In the view of fog computing, the edge is part of the core network and a data center. Fog and Cloud complement each other to make computing, storage and communication possible anywhere along the continuum between the cloud and endpoints. Fog computing is also integrated to the C-RAN architecture to formulate the Fog RAN architecture [43]. The architecture of fog computing is shown in Fig. 3. It contains three layers: cloud layer, fog layer and device layer. The Fog layer may contain multiple tiers according to the requirement. The Fog node could be small BSs, vehicles, WiFi Access Point and even user terminals. The devices choose the most appropriate fog node to associate with.

3) CLOUDLET

The cloudlet proposal is a 3-tier architecture: ''device - cloudlet - cloud'' [36]. Cloudlets could be deployed at WiFi

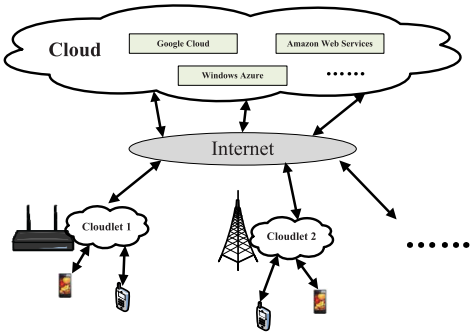

FIGURE 4. Architecture of cloudlets.

access points or LTE base stations [13]. Many new applications require end to end latency of 1 ms. Theoretically, 1 ms of propagation delay requires a cloudlet within 300 km even at the speed of light, In reality, cloudlets should be deployed much closer to ensure the delay requirement. The combination of 5G cellular networks and cloudlets will make this possible [44]. Fig. 4 illustrate the general architecture of Cloudlet systems. To overcome the limited capabilities of single cloudlet, cooperation among different cloudlets is necessary in order to meet the user demands [45]. A comparison of different mobile computing architectures is summarized in Table 3.

4) EDGE CACHING

Caching in the mobile edge network has been proved beneficial. The future mobile networks will be heterogeneous due to dense deployment of different types of base stations. Thus, cache can be deployed at various places in the mobile networks. In legacy cellular system, the content requested by users has to be fetched from the Internet CDN node far away from the mobile networks. Then, caching content at the mobile core network is implemented. However, the backhaul links are still constrained. In addition, with the evolvement of base station and low cost storage unit, deploying cache at macro base stations and small base stations become feasible. In the future 5G networks, D2D communication enables the storage unit at user devices to be exploited for content sharing according to the social relations among users. A general architecture of edge caching is shown in Fig. 5.

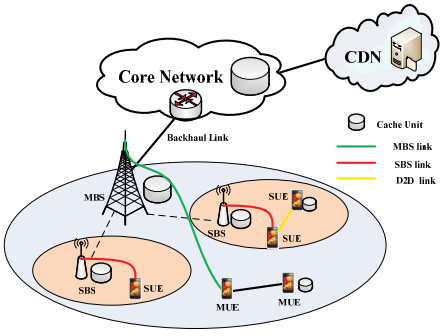

FIGURE 5. Architecture of edge caching.

Based on the above discussion, the general architecture of mobile edge networks is shown in Fig. 6.

TABLE 3. Comparison of different mobile computing architectures.

Item	MCC	MEC	Fog Computing	Cloudlet
Originally proposed by	Not specific	ETSI	Cisco	Prof. Satyanarayanan
Hierarchy	2 tiers	3 tiers	3 or more tiers	3 tiers
Latency	High	Low	Low	Low
Ownership	Centralized by cloud providers: Amazon, Microsoft, etc.	Mobile operators	Decentralized Fog node Owners	Local business
Sharing Population	large	Medium	Small	Small
Location	Large data center	RAN	between devices and DC	between devices and DC, or directly in a device
Context awareness	No	Yes	Yes	Could be
Cooperation between nodes	No	No	Yes	No

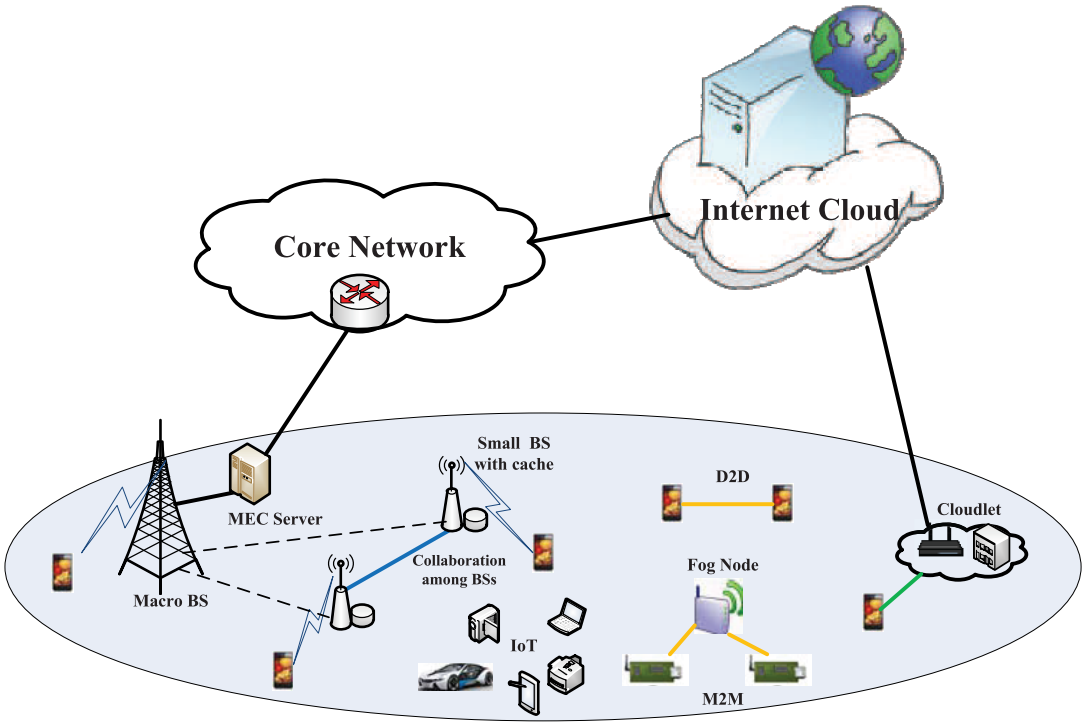

FIGURE 6. Architecture of mobile edge networks.

C. ADVANTAGES OF MOBILE EDGE NETWORKS

Compared to traditional centralized network architecture, researchers found the mobile edge networks have advantages in various aspects. Next we will discuss them in detail.

1) REDUCED LATENCY

Since the processing and storage capabilities are in proximity to the end users, the communication delay can be reduced significantly. The main applications which benefit from this are computational offloading and video content delivery. The work of [46] shows the offload packet delay can be reduce without affecting the network performance using a solution that combine MEC and cloud. Hu *et al.* [47] confirm that offloading through edge computing platforms can achieve great improvement of latency for highly interactive and compute intensive application such as augmented reality and cognitive assistance, in both WiFi and LTE networks. Gao *et al.* [13] present experimental results from Wi-Fi and 4G LTE networks showing that offloading to cloudlets can improve response times by 51% compared to cloud offloading.

2) BANDWIDTH REDUCTION

The deployment of edge servers in mobile edge network infrastructure can save operation cost by up to 67% for the bandwidth-hungry applications and compute intensive

applications [48]. Research results show that backhaul savings can be up to 22% exploiting proactive caching scheme [39]. Higher gains are possible if the storage capability is increased.

3) HIGH ENERGY EFFICIENCY

The experiment results show that edge computing can save energy consumption significantly for different applications in both WiFi networks and LTE networks [47]. The energy consumption of applications using central data centers (DCs) in cloud computing and those using nano data centers in fog computing is compared in [49]. Results show that using nano DCs can save energy, which is affected by the following factors: type of access network, ratio of active time to idle time of nano DCs and type of applications. It is found that the energy consumption in a mobile device can be reduced by up to 42% through offloading to cloudlets compared to cloud offloading [13].

4) PROXIMITY SERVICES

The architecture of mobile edge networks has great advantages in providing proximity services since the edge servers are closer to end users and D2D communication technology can be exploited [50]. Therefore, the traffic load on radio access network can be reduced.

5) UTILIZATION OF CONTEXT INFORMATION

The MEC server deployed at different places in the radio access network can obtain detailed context information including network level information, device level information [50]. With this information, the resources of network can be allocated more efficiently and the user experience can be improved. For example, location based applications can be deployed directly on MEC server rather than on the Internet cloud far from the users.

III. COMPUTING AT MOBILE EDGE NETWORKS

Computing is a major resource in mobile networks. Many compute intensive applications are appearing recently such as augmented reality, high definition video streaming and interactive gaming. However, the computation capability is very constrained in mobile devices. In addition, the power consumption of these computing tasks is very high for the battery capacity of current mobile devices. Edge computing paradigm enables the possibility of offloading the computation tasks to more powerful edge servers. How beneficial is computing at the edge networks? Many works have been done to find the answer to this question. In this section, we will survey the state-of-the-art research efforts on this issue.

A. OBJECTIVES

The benefits of edge computing is various. Different applications or system may have different performance requirements. We will present some of the common performance objectives that edge computing can obtain.

1) MINIMIZE ENERGY CONSUMPTION

Many works have been done to evaluate the energy efficiency of edge computing. Various optimization schemes have been proposed to minimize energy consumption in both the network side and the device side. For computation offloading in 5G heterogeneous networks, the energy cost both the task computing and file transmission should be considered. Zhang *et al.* [51] design an energy efficient computation offloading scheme, which jointly optimizes radio resource allocation and offloading to minimize the energy consumption of the offloading system under the latency constraint. In their scheme, the devices are classified into three types according to their ability and requirements firstly. Then they allocate the wireless channels of MBSs and SBSs to mobile devices according to their priority iteratively until the radio resources are used up or all the devices have been allocated required channels. At each iteration step, the scheme ensures the system obtains minimum energy cost. The results show that the proposed scheme has lower energy consumption that computing without offloading especially with large number of mobile devices. Borylo *et al.* [52], investigate the energy aware interaction between fog and cloud. It is demonstrated that the overall energy consumption of data centers can be reduced without a significant deterioration of the network performance.

2) MAXIMIZE CAPACITY

The next generation 5G networks require support of 1000 times higher mobile data volume per area than current 4G LTE networks [53]. This requires more capacity in the RAN, backhaul and fronthaul. Offloading is one of the combination of technologies that address these challenges in the RAN in addition to more spectrum, network densification and higher spectrum efficiency [54]. The strategy that combine the fog and cloud operations can achieve high system capacity while providing low latency for requested services [55].

3) MINIMIZE LATENCY

Latency is an important performance metric that affects user experience. The latency requirements of next generation 5G networks is 1 ms round trip time (RTT), which is almost 10 times reduction from the 10 ms RTT in 4G [34]. For real-time applications, the delay incurred by offloading tasks to the cloud is unacceptable. Enhancing the high density SBS with computing capabilities is a much more feasible way. Mach and Becvar [56] propose a distributed cloud-aware power control algorithm that is suitable for delay sensitive applications. In [57], a delay minimization problem is formulated under the constraint of power consumption. The authors design an optimal computation task scheduling policy for MEC systems. The delay of general traffic flows in the LTE downlink can be minimized by service level scheduling via MEC server deployed at the eNodeB [58]. Fog also provides low delay but with low capacity limitation, while the combined operation of fog and cloud can minimize the service

TABLE 4. Summary of literatures on computation offloading.

Work Area	Related Work	Key Points
Single User System	[8], [59]. [63]	• Computation offloading choice • optimal scheduling policies • partial offloading problem
Multi User System	[64], [65], [66], [67], [68]	• Multi-user computation offloading problem is NP-hard • Distributed game theoretic approach • Joint optimization of radio and computational resources • Heuristic optimization algorithm • Minimize energy consumption and reduce delay
Offloaded to MEC Server	[70], [71], [72], [60]	• Offload objects produced by multiple IoT devices to MEC server • Reduce latency, energy consumption and execution cost • Load balancing • AP clustering algorithm
Offloaded to Devices	[62]	• Using D2D technology • A collection of co-located mobile devices to provide cloud services • task scheduling problem
Mobility Awareness	[74], [75]	• dynamic contact time between users and MEC servers • Mobility model • mobility aware offloading strategy

latency and guarantee the capacity requirements at the same time [55].

B. COMPUTATION OFFLOADING

One of the main purposes of edge computing is computation offloading to break the limitations of mobile devices such as computational capabilities, battery resources and storage availability. When and how to offload the computation tasks is a hard problem. Various approaches have been proposed to tackle this problem under many kinds of scenarios such as single user case, multi user case and in vehicular networks [59]–[61]. Moreover, in next generation heterogeneous networks, the computation tasks can not only be offloaded to servers but also to devices by utilizing D2D communication [51], [62]. A summary of literatures on computation offloading is shown in Table 4.

1) SINGLE USER CASE

For the single user case, the optimal selection of executing mobile applications in the mobile device (mobile execution) or offloading to the cloud need to be analyzed [8]. One of the common design objective is to save energy for the mobile device. Consider the stochastic channel condition in wireless networks, the optimal scheduling policies should be obtained. In [61], a threshold based scheduling policy is derived which depends on the energy consumption model and wireless channel model. In [59], a low complexity online algorithm is proposed for a MEC system with energy harvesting devices. The algorithm is called the Lyapunov optimization-based dynamic computation offloading algorithm which jointly determines the offloading decision, the CPU-cycle frequencies and the transmit power for offloading. Wang *et al.* [63] investigate the partial computation offloading problem with the aims of minimizing device energy consumption and latency of application execution. They consider both the single server scenario and multiple servers scenario. The results show the conditions under which local execution is optimal and conclude that total offloading is not the optimal choice when the device has the capability of dynamic voltage scaling.

2) MULTI-USER CASE

The Offloading problem in multi-user case is more complex than that in single user case. Many research efforts have been made on this topic. Chen *et al.* [64] show that the multi-user computation offloading problem is NP-hard and propose a distributed game theoretic approach for efficient computation offloading decision. Sardellitti *et al.* [65] jointly optimize the radio resources and computational resources to minimize the overall users' energy consumption with latency constraints in a MIMO multicell system where multiple users offload computing tasks to a common cloud server. An iterative algorithm based on successive convex approximation technique is proposed to solve the nonconvex problem. You *et al.* [66] study the optimal resource allocation for a multi user mobile edge computation offloading system with both time division

multiple access (TDMA) and orthogonal frequency division multiple access (OFDMA). Ketykó *et al.* [67] form the multiuser computation offloading as a multiple knapsack problem and propose a heuristic algorithm to solve it. In [68], the inter-cell interference environment of dense small cells and limitation of computation resources of MEC are considered. An adaptive sequential offloading game approach is proposed which adaptively adjust the number of offloaded users to reduce the queuing delay. Rehman *et al.* [69] propose an opportunistic computation offloading scheme for online data stream tasks. The scheme is evaluated using the event data stream of 5 million activities collected from 12 users for 15 days and results show significant data reduction by 98%.

3) OFFLOADED TO MEC SERVER

The common place to offload computation tasks is MEC servers. The REPLISOM architecture proposed in [70] offload the memory objects produced by multiple IoT devices to the edge cloud located at the LTE eNodeB. It reduces the latency and cost during offloading. In [71], the video encoding process during video call is offloaded to the MEC edge server based on a communication protocol for negotiating the offloading strategy, which reduces energy consumption of mobile devices during video call. In [72], the issue of load balancing in multiuser fog computing scenario is addressed and the authors propose a low complexity algorithm for fog clustering. A two layer radio access points clustering algorithm is proposed which achieve better quality of experience (QoE) than the centralized and decentralized only strategies [60]. Tianze *et al.* [73] take into consideration the energy consumption, time delay and execution unit cost to the design the optimal scheme for task offloading. They define a concept of opportunity consumption based on the limited energy and computing units of a mobile device.

4) OFFLOADED TO DEVICES

With the technology advancement of smart devices, more computation resources can be leveraged using D2D technology. A collection of co-located mobile devices can be utilized to provide cloud services at the edge [62]. In this scenario, computation tasks can be offloaded to other mobile devices nearby other than to the MEC server. The task scheduling problem should be reinvestigated which is different from that of offloading to servers.

5) MOBILITY AWARENESS

The user mobility cannot be ignored in mobile edge networks. Due to user mobility, the contact time between users and MEC servers is dynamic which will impact the strategy of offloading, specifically, where and what to offload [74]. Li and Wang [75] propose a mobility model which suggests that the inter-contact time between any two users complies with an exponential distribution. With this model, the mobility aware computation offloading strategy is developed in [74].

C. COOPERATION BETWEEN THE EDGE AND THE CORE

Despite its advantages of mobile edge networks such as low latency and high energy efficiency, the computing resources of edge networks are still limited. To exploit the merits of both the edge networks and the cloud platforms at the core network, the cooperation between them is valuable. The workload allocation problem is studied in a combined fog-cloud computing system in [76]. The authors design an approximation solution to obtain the tradeoff between power consumption and delay. Simulation results show that cooperation between fog and cloud can significantly improve the performance of cloud computing. A hybrid approach combining fog computing and public/private IoT clouds is proposed in [77] to enable integrated IoT applications. The proposed solution can improve the number of task requests being successfully executed. Souza *et al.* [55] combine the operation of fog and cloud to simultaneously improve the latency performance and system capacity.

D. PLATFORMS

In addition to the theoretical discussions of mobile edge computing, some preliminary implementation of MEC servers are developed. Nokia Neworks introduced a real-world MEC platform in 2014: the Radio Application Cloud Servers (RACS) [78]. RACS use a VM hypervisor to deploy virtual machine images which execute MEC applications. In [79], an adaptive operation platform (AOP) is proposed to manage fog computing infrastructure based on the operational requirements of industrial IoT process. The AOP includes several functional elements: the Model Building element, the Rule Mapper element and the Rule Deployer element. Vallati *et al.* [80] exemplify the M2M Fog platform and evaluate three different deployment to connect Fog nodes to the LTE network: the legacy deployment based on macro cells, D2D based and small cell based deployment.

IV. CACHING AT MOBILE EDGE NETWORKS

To exploit caching technology in the mobile networks, many issues should be studied. Where to cache? What to cache? How to cache? The traditional caching schemes in content centric networks have not consider the wireless networks characteristics such as dynamic traffic load and interference [81]. In this section, we will survey the research efforts that have been made in the mobile edge networks. The related issues include caching places, content description, caching policies, content delivery and so on. A summary of literatures on edge caching is shown in Table 5.

A. CACHING PLACES

There are many places where caching units can be deployed in the mobile networks. In the all-IP based cellular networks, three main places that can deploy cache are the core network, the radio access network (RAN) and user devices. Currently the widely deployed places of caching is the evolved packet core (EPC) [26]. By caching content at the mobile

TABLE 5. Summary of literatures on edge caching.

Work Area	Related Work	Key Points
Content Popularity	[91], [92]. [93], [94]	• Static IRM model, power law distribution • Dynamic model, SNM model
Caching Policies and Algorithms	[95], [82], [97], [98], [100]	• Conventional caching policies • User preference based polices • Learning based policies • Non-cooperative policies • Cooperative policies
Caching File Types	[82], [94], [101], [104]	• Multimedia data • IoT data
Mobility Awareness	[105], [106], [107], [108], [109], [110]	• Spatial and temporal properties of user mobility • Discrete time Markov chain model • Short connection duration problem in mmWave 5G networks

core network, the mobile traffic can be reduced by one to two thirds. Moreover, deploying caching at the EPC is technically easier than at the RAN. The caching places at the edge networks are discussed as follows.

1) MBS CACHING

In heterogeneous networks, MBSs have more coverage areas and can serve more users. Caching at MBS can obtain better cache hit probability. In [82], the performance of reactive and proactive caching at MBS is studied. A video aware backhaul and wireless channel scheduling technique is proposed combining with edge caching. The results demonstrate that the video capacity can be significantly increased and the stalling probability of videos is reduced. Gu *et al.* [83] investigate the storage allocation problem in MBS caching. They propose a heuristic method to solve the NP-hard problem.

2) SBS CACHING

SBSs are densely deployed in next generation heterogeneous networks. Therefore, caching at SBSs is another good choice since the SBSs are more close to the end users and usually provide higher data rate. Many literatures have studied the performance of caching at SBSs [84]–[88].

3) DEVICE CACHING

D2D communication is one of the key technologies in next generation 5G networks. The storage resources in mobile devices can be exploited. The QoE of users can be greatly improved by caching contents in mobile devices if the caching strategy is carefully designed. A caching based D2D communication scheme is proposed in [89] considering the social relations among users and their common interests. In [90], an opportunistic cooperative D2D transmission scheme exploiting caching capability at the devices is proposed. In this scheme, the D2D users are divided into clusters and different popular files are cached at the users within a cluster. Results show that the proposed strategy can provide 5 to 6 times throughput gain over existing D2D caching scheme when the popularity distribution is skewed.

B. CONTENT POPULARITY

To decide what to cache in the edge networks, the popularity of content should be considered to maximize the hit probability of cache, i.e., the probability that the content request by users is cached in the edge networks.

1) STATIC MODEL

Most of the current works on mobile caching assume that the content popularity is static and adopt an independent reference model (IRM): the request of contents is generated according to an independent Poisson process whose rate is related to content popularity modeled by a power law [91]. The commonly used popularity model is the zipf model observed in web caching [92].

2) DYNAMIC MODEL

The static IRM model cannot reflect the real content popularity which is time varying [91]. A dynamic popularity model called the shot noise model (SNM) is proposed in [93]. The model uses a pulse with two parameters to model each content: the duration reflects the content life span and the height reflects its instantaneous popularity. Cha *et al.* [94] analyze the statistical properties of user generated content (UGC) popularity distributions and discuss opportunities to leverage the ''long tail'' video demands.

C. CACHING POLICIES AND ALGORITHMS

Various Caching policies and algorithms have been proposed in mobile caching. Some of the conventional caching policies in wired networks are revised for wireless networks.

TABLE 6. Summary of literatures on caching policies and algorithms.

Work Area	Related Work	Key Points
Conventional Policies	[95], [96]	• Least frequently used (LFU) • Least recently used (LRU) • most popular video (MPV)
User Preference Based Policies	[82]	• Considering local content popularity • User's preference toward specific video categories
Learning Based Policies	[97], [98]	• Estimation of timely content popularity with reinforcement learning • Q-learning based cache replacement strategy • Coded caching scheme
Non-cooperative Polices	[82], [98]	• Distributed caching neglecting caches in other cells • Cache replacement problem modeled by MDP
Cooperative Policies	[99], [100], [101], [102], [90], [88], [103]	• Cooperative cache management among base stations, UEs • Minimize bandwidth cost • Maximize traffic served from cache • Collaborative video caching and scheduling among cells • Tradeoff between caching redundancy and diversity • Joint caching and routing design

In addition, new schemes such as learning based policies and cooperative caching policies are also proposed. The literature [95] reviews in detail the conventional caching policies and forwarding mechanism in information centric networks. we will present a taxonomy of the caching policies in wireless mobile networks in the following, which is summarized in Table 6.

1) CONVENTIONAL CACHING POLICIES

Content replacement policies such as the least frequently used (LFU) and least recently used (LRU) have been adopted in a large number of caching policies [95], [96]. These strategies are simple and efficient with uniform size objects. However, these policies ignore the download latency and size of objects. Another proactive caching policy used in content deliver networks is the MPV policy, which caches the most popular videos based on the global video popularity distribution [82]. However, the cache size of the RAN is very limited compared to that of CDN. The hit probability achieved by MPV policy could be too low for RAN caches.

2) USER PREFERENCE BASED POLICIES

Ahlehagh and Dey [82] propose a user preference profile (UPP) based caching policy. It is observed that local video popularity is significantly different from national video popularity and users may show strong preferences toward specific video categories. The UPP of each user is defined as the probability that a user requests videos of a specific video category.

3) LEARNING BASED POLICIES

In fact, the content popularity is time-varying and is not known in advance. Therefore, the track and estimation of timely content popularity is an important issue. Based on machine learning technology, learning based caching polices are proposed in [97]. Sengupta *et al.* [97] solve the problem of distributed caching in SBSs from a reinforcement learning view. By adopting coded caching, the caching problem is reduced to a linear program that considering the network connectivity and the coded aching scheme performs better than the uncoded scheme. Gu *et al.* [98] solve the cache replacement problem with a Q-learning based strategy.

4) NON-COOPERATIVE CACHING

Some of the existing caching policies decide the content to cache at each base station without consider the cooperation among BSs. In [82], the proposed scheme makes caching decision based on the UPP of active uses in a specific cell without considering the impact of caches in other cells. In [98], the cache replacement problem modeled as Markov Decision Process (MDP) is solved in a distributed way using Q-learning method, without exchanging extra information about cached data between the BS. This strategy outperforms the conventional ones such as the LFU, LRU and randomized strategy.

5) COOPERATIVE CACHING

A lot of existing works have studied the cooperation among cache places when designing the caching policies. In [99],

a light-weight cooperative cache management algorithm is developed to maximize the traffic volume served from cache and minimize the bandwidth cost. In [100], the cooperation between femto base stations and user equipment (UE) for content caching and delivery is investigated. The cooperative caching problem is formulated as an integer-linear programming problem and solved by using the subgradient method. The content delivery policy is formulated as an unbalanced assignment problem and solved by using Hungrian algorithm. Yu *et al.* [101] explore the exploiting of scalable video coding (SVC) technique in collaborative video caching and scheduling among cells to further improve the the caching capacity and the QoE of users. Wang *et al.* [102] investigate the cooperation among caches in the RAN and obtain the optimal redundancy ratio of content cached in each base station. The cooperation among D2D users is exploited for cache enabled D2D communication in [90]. A network coding based content placement method is proposed in [103]. The strategy increases the amount of available data to users and results in a fair distribution of data at the same time. Poularakis *et al.* in [88] jointly design the caching and routing scheme to maximize the content requests served by small cell base stations under the constraint of BS bandwidth. The problem is reduced to a facility location problem and is solved using bounded approximation algorithms.

D. CACHING OF DIFFERENT FILE TYPES

The most common file types for caching is multimedia files such as popular videos and audio files. The internet of things is one of the main use cases of the next generation 5G networks. Thus the caching of IoT data is also important as the IoT data volume is increasing and IoT data have different characteristics compared to multimedia data.

1) LARGE DATA FILES OR MULTIMEDIA DATA

The caching design of large multimedia data especially video files is studied in most of the existing literatures [82], [94], [101]. The characteristic of multimedia data is that many users have the same interest for popular videos. Therefore, caching the popular files in the RAN can benefit from high hit ratio.

2) IoT DATA

The low-rate monitoring, measurement, and automation data generated by IoT applications ruuning on billions of devices need to be cached to reduce total traffic load. However, IoT data are different from multimedia data in that the IoT data have short lifetimes. Thus, different caching policies are required. Vural *et al.* [104] propose a model that takes into consideration both the communication costs and freshness of a transient IoT data item. The network load can be reduced significantly especially for highly requested data.

E. MOBILITY AWARENESS

The user mobility is a unique feature of wireless network, thus it should be considered in caching at the network edge. Many works have been done on this issue. Wang *et al.* [105] propose a general framework for mobility aware caching in content centric wireless networks. Both the spatial and temporal properties of user mobility are modeled. In [106], the mobility pattern of users is considered and the mobility aware caching problem is formulated as an optimization problem aiming at maximize the caching utility and the authors propose a polynomial-time heuristic solution to solve the problem. The impact of user mobility on the hit performance of edge caching is analyzed in [107]. The user mobility is modeled as a discrete time Markov chain in [108]. The authors consider a scenario where segments of encoded content files are stored in a set of base stations in a cell with a main base station in it. The caching algorithm is designed to minimize the probability of using main base station for file delivery. Different from the prior works that assume user mobility is known, the continuity of content service for the handover users is investigated based on mobility prediction algorithm when user mobility is unknown in [109]. Gomes *et al.* [110] exploit the information centric networking and the mobile Follow-Me Cloud approach to enhance the migration of content-caches located at the mobile network edge. The proposed content relocation algorithm improves content-availability by up to 500% compared with existing solutions.

F. IMPACT ON SYSTEM PERFORMANCE

1) CAPACITY

Existing works on edge caching have proved that caching at the network edge can significantly improve system capacity. For example, the solution proposed in [82] can improve capacity by 3 times compared to having no cache in the RAN.

2) DELAY

Caching at the network edge can significantly reduce content delivery delay due to the proximity of caches to end users. In [82], the initial delay and stalls of a video session is reduced by jointly scheduling RAN backhaul and wireless channels, therefore the video QoE is improved. In [111], the tradeoff between delivery latency and fronthaul and aching resources is derived.

3) SPECTRAL EFFICIENCY

Liu and Yang [112] compare a cache-enabled 2-tier HetNet with a conventional HetNet without cache using stochastic geometry theory. The numerical results show that the cache-enabled helper density can be reduced by 3/4 compared to the pico BS density without cache to achieve the same area spectral efficiency (ASE). Given the total capacity within an area, an optimal cache enabled SBS density exists that maximizes the ASE.

4) ENERGY EFFICIENCY

Energy Efficiency is another important performance metric for the next generation 5G networks. In [113], the impact of caching at BSs on the energy efficiency of downlink

networks is analyzed. The results show that the energy efficiency will be improved when the file catalog size is small and caching at multiple pico BSs is more energy efficient than caching at a macro BS. In [114], the fundamental tradeoffs between energy efficiency and small cell density in software defined heterogeneous networks is demonstrated. The energy efficiency of cache enabled heterogeneous networks is much higher than current LTE networks.

V. ADVANCES IN COMMUNICATIONS TECHNIQUES WITH SYNERGY OF COMPUTING AND CACHING

The combination of computing and caching resources with communication systems will change the design philosophy of communication networks to a large extent. Combining 5G with MEC would make certain inter and intra domain use cases feasible such as automotive services and e-Health services [50]. By coordinating the cloud enabled small cells in a cluster, a distributed light data center can be created inside the communication networks [51], [115]. In this section, we will discuss the communication techniques and the synergy of communication, computing and caching.

A. mmWave COMMUNICATION

The use of mm-wave spectrum in future networks enables high data rate transmission. Various schemes of integrating 5G BSs with legacy cellular networks have been proposed such as standalone mm-wave systems and hybrid systems of mm-wave BS and 4G BS [34]. The mm-wave spectrum is mainly used for data communications. The narrow beams used by mm-wave BSs can improve the link quality between BSs and large number of users. However, this kind of deployment also raise new challenges of short connection durations and frequent handoffs for high mobility users, making video streaming suffer from long latency. Qiao *et al.* [116] propose a caching-based mm-wave framework that precaches contents at the BS for handoff users. The proposed solution can provide consistent high quality video streaming for high mobility users in 5G mm-wave small cells.

B. D2D COMMUNICATION

D2D communication enables the direct transmission between end devices in proximity. It has been recognized as one of the main technologies for 5G networks. The advantages of using D2D communication include one-hop communication, higher spectral efficiency, low transmission power, coverage extension and spectrum reuse [117], [118]. In addition, with the help of D2D communication, massive MTC devices can be offloaded from MBSs to SBSs, which improves overall network capacity and avoids traffic congestion at MBSs [119]. Many previous research efforts on D2D communication have been focus on the issues of spectrum reuse, peer discovery, power control, connection establishment and interference management [120], [121]. With the concept of edge caching and computing, the utilization of computing and storage resources on large number of smart devices draws the attention of researchers. Computation offloading and content sharing problems via D2D communications are under investigation by many researchers [62], [89], [90].

C. TRANSMISSION SCHEMES

The ubiquitously deployed low-cost caches promote the development of content centric transmission. These caches at the mobile networks move the content near end users. However, the improvement of the effectiveness of small caches in small cells is a challenge. Since the content cached is popular for the local users, some requests for the same file may happen at nearby time. Therefore, these requests can be served via a single multicast transmission [87]. Compared to caching schemes using unicast transmission, the multicast-aware scheme can reduce cost by up to 88%. When the requested contents are all different in some time slots, a interweaved transmission scheme can be adopted to balance the traffic of sequential time slots of each SBS [122].

D. INTERFERENCE MANAGEMENT

In future ultra dense heterogeneous networks, interference is a key challenge. One of the potential technique for interference management of small cell networks is interference alignment (IA). Zhao *et al.* [122] investigate the IA problem of content centric communication with caching and computing exploited. Utilizing the content centric principle via caching, the topology of interference network is simplified and interference management becomes easier. Thanks to the high computational capacity of cloud computing platform, the solution of IA can be easily calculated. The proposed framework reduces backhaul load and the overhead of CSI feedback, and improve the throughput at the same time.

E. COMMUNICATION RESOURCES ALLOCATION AND SCHEDULING

With the integration of caching and computing resources in communication systems, the allocation and scheduling of communication resources is different from the legacy networks. In addition to cache placement, the scheduling of communication resources also affect the efficiency of caching. In [82], joint design of video-aware backhaul scheduling and caching can improve capacity by more than 50% than conventional policies. The careful design of wireless channel scheduling can result in significantly higher video capacity when the wireless channel bandwidth is constrained. The joint caching and video scheduling strategy that combines collaborative caching with SVC is studied in [101].

F. SYNERGY OF COMMUNICATION, COMPUTING AND CACHING

By enabling the mobile network with more computing capability, the scarce communication resources could be saved. In [123], a content-slimming system is proposed, which detect redundant video content and clip the from the original frames by computing and only transmit the necessary video content. This scheme reduce the transmission bandwidth consumption at least by 50% compared to H.264 without sacrifice video quality and visual experience.

TABLE 7. Applications and use cases of mobile edge networks.

Applications and Use Cases	Related Work	Key Points
Dynamic Content Delivery	[19], [125]	• Placing content close to users • Exploiting user's context information
AR/VR	[126]	• Real-time fast processing • Context aware
Intensive Computation Assistance	[126]	• Low latency, low cost devices • Collecting info. from multiple sources
Video Streaming and Analysis	[126], [127]	• Avoiding redundant video streams transmission • More capable of analysis
Internet of Thing	[21], [36], [79], [128], [129], [130], [131], [132], [133], [134], [135]	• Healthcare • Wireless sensor systems • Smart grid • Smart Home • Smart City
Connected Vehicles	[19], [36], [130]	• V2X communication • Automotive safety services • Traffic control and smart parking
Cognitive Assistance	[136]	• Augmenting human perception and cognition ability • Processing latency sensitive tasks
Wireless Big Data Analysis	[19], [37]	• reduce bandwidth consumption and network latency

The application types, user mobility and communication resources will influence the optimal position of deploying computing and caching resources [124]. For the low bandwidth, high persistence use cases that require computing tasks, computing resources should not be deployed at femtocells because frequent handovers and low computational power of the femtocells, especially when the backhaul transmission delays are minimal. For the high-bandwidth location-bound services require storage resources, such as AR, the computing and caching resources should be deployed as close to the end user as possible. In summary, computing and caching resources deployed at different layers of the network should be utilized according to the service types with the consideration of backhaul capacity.

VI. APPLICATIONS AND USE CASES

The new applications are the main driven force for the evolvement of network architecture. The requirements of emerging applications become more and more strict in data rate, latency, etc. In this section, we will summarize the applications and use cases in mobile edge networks as shown in Table 7.

A. DYNAMIC CONTENT DELIVERY

With the increasing demand of multimedia content, the backhaul links face congestion problems in conventional centralized network architecture. Caching at the edge networks can provide dynamic content delivery based on the information of network status and user's context-aware information [19], [125]. Since the content is placed close to mobile users, the QoE of mobile users is improved significantly.

B. AUGMENTED REALITY/VIRTUAL REALITY

The augmented reality (AR) and virtual reality (VR) technology is seen as the most promising application that will change our life style. This application needs real-time information of user's status such as the position and direction they are facing. The MEC server is capable of exploiting local context information and has powerful processing ability, which is very suitable for the AP/VR applications [126].

C. INTENSIVE COMPUTATION ASSISTANCE

The computational capability is often sacrificed in order to lower the cost of devices. Therefore, computation offloading requiring very low processing time with low latency is necessary for such applications with intensive

computation tasks. MEC servers are equipped with high computational capabilities and can process the offloaded computation tasks in very short timescales. Moreover, the MEC server can collect information from multiple sources, which helps those devices perform tasks that require information from multiple sources [126].

D. VIDEO STREAMING AND ANALYSIS

It is observed that video traffic accounts for more than half of the total mobile data traffic in current networks and the percentage is still increasing. The adoption of edge caching avoids numerous redundant video streams transport through the core network to Internet CDNs. The use of MEC server allow video analysis at the more capable cloud platforms at network edge other than at the source producing videos [126], [127].

E. INTERNET OF THINGS

The idea of Internet of Things is becoming a reality with the technology advancements in smart sensors, communications, and Internet protocols [128]. Fog computing is an advantageous architecture in support of IoT applications. Gazis *et al.* [79] propose an adaptive operations platform which enable the application of Fog Component in industrial IoT context. The VM-based cloudlets at mobile edge networks enable edge analytics for crowd-sourced video content in IoT [129]. MEC will provide new IoT services that would not be feasible before [130], [131]. The specific applications and use cases of MEN in the Internet of Things is presented below.

1) HEALTHCARE

Real-time processing and event response are important for healthcare applications. Experiments proved that the healthcare system utilizing fog computing responded faster and was more energy efficient than cloud-only approaches [132]. For example, fog computing can be used to detect falls of stroke patients.

2) WIRELESS SENSOR SYSTEMS

Various scenarios that make use of wireless sensor systems can be use cases of mobile edge cloud computing platforms, such as in oil & gas industry, building industry and environmental monitoring [36].

3) SMART GRID

The analysis of the data generated in the smart grid environment is a very challenging task due to the complex parameters. The using of mobile edge computing can improve performance in throughput, response time and transmission delay [133]. Smart Grid is a typical use case that require the interplay between the fog and the cloud [21]. The fog collect and process the local data generated by grid sensors and devices. The cloud provides global coverage and restore data that have a life cycle of months and years.

4) SMART HOME

Smart home systems have become a new trend for future housewide ecosystems. Smart home is a kind of small-scale IoT system with limited spatial occupancy and localized communications. Deploying MEC servers as IoT gateways close to the smart objects will enable direct M2M interactions in future networks [134]. MEC node, which could be deployed on femtocells, home routers, set-top boxes and smartphones, is beneficial for low latency, localized and plug-and play services for smart home.

5) SMART CITY

The vision of smart city is to improve quality of life by utilizing technological advancements. The concept of MEC is helpful for time-critical events because the contents generated by large number of connected devices can be exploited to discover the occurrence of anomalous events [135]. Another useful component of smart city system is smart traffic lights. For instance, the smart traffic lights can send warning signals to approaching vehicles, or detect the pedestrians and cyclists who is crossing the street, or warn the vehicles of risks of running a red light [36].

F. CONNECTED VEHICLES

Edge computing approaches can play an important role in connected vehicles, V2X communication and automotive safety services such as real-time warning of ice on motorway and coordinated lane change maneuvers [36]. The applications run on MEC servers are in close proximity to the vehicles and can provide roadside functionality with low latency [130]. Traffic control and smart parking can be achieved since the edge network is able to collect and analyze real-time data from sensor devices installed ubiquitously [19].

G. COGNITIVE ASSISTANCE

Cognitive assistance applications are used to augment human perception and cognition ability. Satyanarayanan *et al.* [136] demonstrate how cloudlet can help in that cloudlet is only one wireless hop away cloud proxy. Therefore, it is the ideal offload site selection for cognitive assistance. All the latency-sensitive processing tasks are offloaded to cloudlet which the user device associates to. When the user moves away from the proximity of one cloudlet, the use will be handed off to another cloudlet in its proximity.

H. WIRELESS BIG DATA ANALYSIS

Big data is commonly characterized along three dimensions: volume, velocity and variety. Bonomi*et al.* [37] suggests geo-distribution as a fourth dimension to the characterization of big data. For example, the large number of sensors and actuators are naturally distributed. Fog computing, as a distributed intelligent platform that manages distributed networking, compute and storage resources, is a promising choice to handle these big data. Compared to big data

analytics performed at the core network, doing big data analytics at the network edge will reduce bandwidth consumption and network latency [19].

VII. KEY ENABLERS

In this section, the key technologies that enable the concept of mobile edge networks become reality are presented. These technologies provide flexibility, scalability and operating efficiency to the mobile edge networks.

A. CLOUD TECHNOLOGY

The concept of mobile edge networks is an extension of cloud computing capabilities to the edge of mobile networks. The advancement in cloud technology makes it easier to deploy virtual machines on high volume general purpose Servers in places such as base stations and gateways [19]. The cloud can offer powerful processing capabilities and huge amount of resources. The integration of cloud and IoT has been proved beneficial for delivering new services [137]. The mobile edge networks are integrated with cloud computing capabilities and offer effective solutions for service management and provision.

B. SOFTWARE DEFINED NETWORK

SDN technology enable the network to be intelligent, programmable, and more open [138]. The main idea of SDN is to separate the control and data planes. The benefits of SDN include creating network control planes on common hardware, exposing network capabilities through APIs, remotely controlling network equipment, and logically decoupling network intelligence into different software-based controllers [138]. SDN technology overcomes the shortcoming of management complexity due to large scale deployments of servers and applications [139]. Applying the SDN paradigm will enable management at different levels needed in the MEC platforms [140].

C. NETWORK FUNCTION VIRTUALIZATION

Network function virtualization (NFV) is a complementary technology of SDN proposed for future 5G networks. The purpose of NFV is to virtualize a set of network functions by moving them from dedicated hardware to general purpose computing platforms using software technologies, which can provide the same services as legacy mobile networks. As a result, the capability of managing vast heterogeneous devices is improved as well as the scalability and flexibility of the network [5]. With NFV, both the capital expenditures (CAPEX) and operating expenses (OPEX) of network operators are likely to be reduced. Network virtualization leads to the problem of virtual network embedding (VNE), VNE algorithms are currently being studied [141]. The application of NFV changes the landscape of telecommunications industry and brings many benefits such as reducing the time to market, optimizing network configuration and topology in near real time, and supporting multi-tenancy [142].

D. SMARTER MOBILE DEVICES

In the legacy generations of cellular networks (2G, 3G, 4G), the system design is oriented by having complete control at the network infrastructure side. However, with the mobile devices become more powerful and smarter, this design philosophy should be changed to utilize intelligence at the device side [35]. In the future networks, the devices will play a more active role with more smartness. One important technology is the D2D communication. In current network, data traffic has become the main traffic types instead of voices. There are many situations where devices in close proximity would like to share content or interact with each other, for example, gaming and social networking. The direct D2D communication can improve network efficiency in several aspects. Firstly, it saves a lot of signaling resources and reduces transmission latency. Secondly, it can save large amount of energy compared to transmission through the help of base stations. Furthermore, the spectral efficiency can be improved since the path losses are much lower than BS to device communication. The D2D communication is expected to be natively supported in 5G network. Another useful technology is local caching, which is intended to strike a balance between data storage and data transfer. For wireless devices, the marginal cost of transferring information is non-negligible. With more and more memory units installed in today's mobile devices, caching popular contents such as video or audio files at devices is clearly cheaper and more efficient than to transmit such content repeatedly via unicast since the demand is asynchronous.

VIII. OPEN CHALLENGES AND FUTURE DIRECTIONS

Mobile edge network is an revolution in the architecture of wireless mobile networks. It has many new features comparing to the existing 3G/4G cellular systems with better QoE performance and flexibility. Therefore, a wide variety of research challenges and opportunities exists for future research works. In this section, we point out the major open challenges in mobile edge networks and shed light on the possible future research directions.

A. OPEN RESEARCH CHALLENGES

The stringent requirements of the next generation mobile networks, such as ultra high throughput, extremely low latency and high energy efficiency, pose great challenges on current research of both the academia and industry. We summarize the key research issues related to mobile edge networks below:

1) HETEROGENEITY

In the future networks, with the development of IoT and noval applications, the heterogeneity in networking, communication, and devices becomes a critical issue. This heterogeneity causes other related problems such as the asynchronism and non orthogonality [34]. The challenge of handling this heterogeneity under a unified network architecture need to be fully investigated.

2) COMPUTATION MODELING

To validate the accuracy of analysis and simulation works, the model of computation resources should be accurate. In current literatures, the computation resources are modeled as CPU cycles per second [66]. Although this model is simple for analysis, whether it fully reflects the characteristics of computing still remains a question. Hence, more accurate computation models should be developed.

3) ENABLING REALTIME ANALYTICS

A lot of novel applications require realtime analytics such as VR/AR and e-Health. The dynamic resource management should be determined which schedules the analytic tasks to the most appropriate edge server guaranteeing the latency and throughput [132].

4) USER MOBILITY

User mobility is a key challenge in mobile edge networks. It has non-negligible impact on caching and computation offloading decisions. The frequent mobility of users will cause frequent handovers among edge servers, Mobility management technique considering both horizontal and vertical mobility should be implemented allowing users to access edge servers seamlessly [19].

5) PRICING POLICY

In mobile edge networks, the storage, computing and communication resources are allocated dynamically according to users' demand. Thus, the optimal pricing policy is different from the legacy systems. From a commercial perspective, the profits of all the stakeholders in the system should be balanced. The profit of the edge cloud is significantly influenced by the pricing policy when users care about the price a lot [143]. Literature [144] shed lights on the pricing and resource allocation problem in video caching systems from a game theoretic perspective.

6) SCALABILITY

Scalability is an important feature that mobile edge networks provide compared to the legacy systems. The increasing number of mobile devices such as IoT devices will require scalability of services by applying load balancing mechanism [19]. A cloud orchestrator which flexibly manage the edge node may provide network scalability [145].

7) SECURITY

The deployment of edge cloud servers is creating novel security challenges due to the exploiting of the mobile device information [146]. The growing rate of the evolution of security solutions cannot keep up with the pace of the new security challenges. Many existing security protocols assume full connectivity [147], which is not realistic in mobile edge networks since many links are intermittent by default. A defense technique is proposed in [146] for malicious node in mobile edge computing platforms called HoneyBot. The HoneyBot nodes can detect, track, and isolate D2D insider attacks. The speed and accuracy of this technique are impacted by the placement and number of HoneyBot nodes.

The security solutions for cloud computing may not suitable for edge computing because the working surroundings of edge devices will face many new threats different from well managed cloud [148]. The authentication at different levels of gateways and smart meters is another important security issue. Some solutions have been proposed for the authentication problem, such as public key infrastructure (PKI) based solutions [149], Diffie-Hellman key exchange based solutions [150].

Intrusion detection techniques are also required in fog computing. Some detection methods have been proposed for different applications [148]. For example, signature-based method observes the behavior patterns and checks against a database of possible misbehaviors. The anomaly-based method detects intrusion by comparing an observed behavior with expected behavior to check the deviation.

8) PRIVACY

Privacy issues deal with hiding details. The content sharing and computation cooperation among users via D2D communication draw the concern on user's privacy. How to exploit the resources of mobile devices without invasion of privacy remains a challenge. In the typical use case of fog computing, the smart grid, the encryption of data from the smart meters and the aggregated result in Fog devices ensures data privacy. The original data can only be decrypted at the operation center [151]. In addition, many data privacy mechanisms have been developed for MCC to enforce privacy policies among collaborating mobile devices and to conceal the location of a set of clients [24]. These mechanisms shed light on the design of privacy mechanisms for collaborative edge data centers in mobile edge networks. Most of the existing privacy solutions only require a trusted platform module (TPM), which can be deployed in edge data centers.

In the private networks cases such as personal cloudlets and corporate networks, the users and the edge data centers have a trustworthy relationship. Thus, privacy helper entities can be deployed in the edge data centers [24]. Various data privacy mechanisms and services can be implemented on these helpers. In addition, the edge paradigm is helpful in strengthen the privacy feature of certain services. For instance, the anonymity of the users of location-based services can be protected by deploying a crowd-souring platform in a trusted edge server [152].

9) USER PARTICIPATION

A major idea in mobile edge networks is give the opportunity to utilize user terminals' available resources. Uses could be empowered with toolkits that enable users play an active role in technology design [137]. However, this cooperation among users depends on users' willingness to participate. User incentive mechanisms need to be considered when designing computation offloading and caching strategies [64].

B. DISCUSSIONS ON FUTURE DIRECTIONS

1) UTILIZATION OF WIRELESS BIG DATA

The wireless big data generated in the mobile edge networks is a valuable resources for analysis and design of the networks. Context aware approach also need the analysis of large amount of context information data. For example, user information big data can be utilized for popularity estimation in edge caching systems [153]. Thus, the full utilization of wireless big data will provide new opportunities in the performance of mobile networks.

2) ONLINE CACHING

The caching problem include two phases: cache placement and content delivery. Many works have been done on the cache updating during placement phase. However, more efficient caching update rules during the delivery phase, i.e. online caching, is a future direction of caching researches [154].

3) CONTEXT AWARENESS

The mobile edge networks are advantageous in exploiting context information. The context provides information such as user location, other users in vicinity, and resources in the environment [155]. The real time context aware applications could be accomplished by collaborations among MEC platforms [50]. Different levels of context information (application, network and device level) could be used to proactively allocate resources [156].

4) SMART USER ASSOCIATION

In cache enabled ultra dense HetNet, caching may change the way of user association other than the conventional nearest distance or SINR based method [157]. The users may associate to the BS which caches the content it requested. In this way, the nearest BS may have strong interference on the users. Cache aware user association may overcome the backhaul capacity limitations and enhance users' QoE [158]. There could be more than one user access mode and users will select the best access mode for them [159].

5) INTEGRATION

The architecture of mobile edge networks involve in various resources: computing, storage and communications. The efficient integration of these resources to achieve the optimal performance for all users and application is not concluded. More comprehensive resource allocation schemes need to be developed. Research on this topic may continue with the evolving of the network.

IX. CONCLUSION

This paper surveys and summarizes the research efforts made on the mobile edge network, which is a paradigm integrating computing, caching and communication resources. The proposed architectures for edge computing and caching are presented including the ETSI MEC, Fog Computing, Cloudlet and edge caching. The related issues of computing, caching and communications are discussed respectively. For edge computing, the proposed schemes of computation offloading are extensively surveyed and classified. The issues of cooperations between the edge and the core as well as some existing edge computing platforms are presented. For edge caching, we make a detailed taxonomy on where, what and how to cache. Then the advances of communication techniques and the synergy with computing and caching are discussed. The novel applications and use cases are the driven force of the mobile edge network architecture. We summarize these applications and uses cases that mobile edge networks can fully enable. This new paradigm faces many challenges and opportunities. We point out the future research directions of this hot topic.

ACKNOWLEDGMENT

The authors would like to thank the editor and the anonymous reviewers for their constructive comments and suggestions, which improve the quality of this paper.

REFERENCES

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

[2] H. Shariatmadari *et al.*, "Machine-type communications: Current status and future perspectives toward 5G systems," *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 10–17, Sep. 2015.

[3] M. Condoluci, M. Dohler, G. Araniti, A. Molinaro, and K. Zheng, "Toward 5G densenets: Architectural advances for effective machine-type communications over femtocells," *IEEE Commun. Mag.*, vol. 53, no. 1, pp. 134–141, Jan. 2015.

[4] F. M. Awuor and C.-Y. Wang, "Massive machine type communication in cellular system: A distributed queue approach," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kuala Lumpur, Malaysia, May 2016, pp. 1–7.

[5] M. R. Palattella *et al.*, "Internet of Things in the 5G era: Enablers, architecture, and business models," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 3, pp. 510–527, Mar. 2016.

[6] M. Satyanarayanan, "Mobile computing: The next decade," *ACM SIGMOBILE Mobile Comput. Commun. Rev.*, vol. 15, no. 2, pp. 2–10, Apr. 2011.

[7] A. R. Khan, M. Othman, S. A. Madani, and S. U. Khan, "A survey of mobile cloud computing application models," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 1, pp. 393–413, Feb. 2014.

[8] W. Zhang, Y. Wen, K. Guan, D. Kilper, H. Luo, and D. O. Wu, "Energy-optimal mobile cloud computing under stochastic wireless channel," *IEEE Trans. Wireless Commun.*, vol. 12, no. 9, pp. 4569–4581, Sep. 2013.

[9] L. Guan, X. Ke, M. Song, and J. Song, "A survey of research on mobile cloud computing," in *Proc. IEEE/ACIS 10th Int. Conf. Comput. Inf. Sci. (ICIS)*, Sanya, China, May 2011, pp. 387–392.

[10] H. T. Dinh, C. Lee, D. Niyato, and P. Wang, "A survey of mobile cloud computing: Architecture, applications, and approaches," *Wireless Commun. Mobile Comput.*, vol. 13, no. 18, pp. 1587–1611, Dec. 2013.

[11] X. Fan, J. Cao, and H. Mao, "A survey of mobile cloud computing," *ZTE Commun.*, vol. 9, no. 1, pp. 4–8, 2011.

[12] E. E. Marinelli, "Hyrax: Cloud computing on mobile devices using mapreduce," Carnegie Mellon Univ., Pittsburgh, PA, USA, Tech. Rep. CMU-CS-09-164, 2009.

[13] Y. Gao, W. Hu, K. Ha, B. Amos, P. Pillai, and M. Satyanarayanan, "Are cloudlets necessary?" School Comput. Sci., Carnegie Mellon Univ., Pittsburgh, PA, USA, Tech. Rep. CMU-CS-15-139, Oct. 2015.

[14] A. Klein, C. Mannweiler, J. Schneider, and H. D. Schotten, "Access schemes for mobile cloud computing," in *Proc. 11th Int. Conf. Mobile Data Manage.*, Kansas City, MO, USA, 2010, pp. 387–392.

[15] N. Fernando, S. W. Loke, and W. Rahayu, "Mobile cloud computing: A survey," *Future Generat. Comput. Syst.*, vol. 29, no. 1, pp. 84–106, Jan. 2013.

[16] M. Alizadeh, S. Abolfazli, M. Zamani, S. Baharun, and K. Sakurai, "Authentication in mobile cloud computing: A survey," *J. Netw. Comput. Appl.*, vol. 61, pp. 59–80, Feb. 2016.
[17] Y. C. Hu, M. Patel, D. Sabella, N. Sprecher, and V. Young, "Mobile edge computing—A key technology towards 5G," Eur. Telecommun. Standards Inst., Sophia Antipolis, France, 2015, white paper 11.
[18] *5G Vision: The 5G Infrastructure Public Private Partnership: The Next Generation of Communication Networks and Services*, accessed on Oct. 1, 2016. [Online]. Available: https://5gppp.eu/wp-content/uploads/2015/02/5G-Vision-Brochure-v1.pdf
[19] A. Ahmed and E. Ahmed, "A survey on mobile edge computing," in *Proc. 10th Int. Conf. Intell. Syst. Control (ISCO)*, Coimbatore, India, 2016, pp. 1–8.
[20] M. T. Beck, M. Werner, S. Feld, and S. Schimper, "Mobile edge computing: A taxonomy," in *Proc. 6th Int. Conf. Adv. Future Internet*, 2014, pp. 48–54.
[21] F. Bonomi, R. Milito, J. Zhu, and S. Addepalli, "Fog computing and its role in the Internet of Things," in *Proc. 1st Ed. MCC Workshop Mobile Cloud Comput.*, 2012, pp. 13–16.
[22] S. Yi, C. Li, and Q. Li, "A survey of fog computing: Concepts, applications and issues," in *Proc. ACM Workshop Mobile Big Data*, Hangzhou, China, 2015, pp. 37–42.
[23] S. Yi, Z. Qin, and Q. Li, "Security and privacy issues of fog computing: A survey," in *Proc. Int. Conf. Wireless Algorithms, Syst., Appl.*, 2015, pp. 685–695.
[24] R. Roman, J. Lopez, and M. Mambo. (2016). "Mobile edge computing, fog *et al.*: A survey and analysis of security threats and challenges." [Online]. Available: https://arxiv.org/abs/1602.00484
[25] M. Satyanarayanan, P. Bahl, R. Caceres, and N. Davies, "The case for VM-based cloudlets in mobile computing," *IEEE Pervasive Comput.*, vol. 8, no. 4, pp. 14–23, Oct./Dec. 2009.
[26] X. Wang, M. Chen, T. Taleb, A. Ksentini, and V. C. M. Leung, "Cache in the air: Exploiting content caching and delivery techniques for 5G systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 131–139, Feb. 2014.
[27] W. Ali, S. M. Shamsuddin, and A. S. Ismail, "A survey of Web caching and prefetching," *Int. J. Adv. Soft Comput. Appl.*, vol. 3, no. 1, pp. 18–44, Mar. 2011.
[28] S. Podlipnig and L. Böszörmenyi, "A survey of Web cache replacement strategies," *ACM Comput. Surv.*, vol. 35, no. 4, pp. 374–398, Dec. 2003.
[29] M. Zhang, H. Luo, and H. Zhang, "A survey of caching mechanisms in information-centric networking," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 3, pp. 1473–1499, 3rd Quart., 2015.
[30] C. Fang, F. R. Yu, T. Huang, J. Liu, and Y. Liu, "A survey of energy-efficient caching in information-centric networking," *IEEE Commun. Mag.*, vol. 52, no. 11, pp. 122–129, Nov. 2014.
[31] H. Hawilo, A. Shami, M. Mirahmadi, and R. Asal, "NFV: State of the art, challenges, and implementation in next generation mobile networks (vEPC)," *IEEE Netw.*, vol. 28, no. 6, pp. 18–26, Nov./Dec. 2014.
[32] X. Zhang *et al.*, "Macro-assisted data-only carrier for 5G green cellular systems," *IEEE Commun. Mag.*, vol. 53, no. 5, pp. 223–231, May 2015.
[33] K. Pentikousis, Y. Wang, and W. Hu, "Mobileflow: Toward software-defined mobile networks," *IEEE Commun. Mag.*, vol. 51, no. 7, pp. 44–53, Jul. 2013.
[34] M. Agiwal, A. Roy, and N. Saxena, "Next generation 5G wireless networks: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 3, pp. 1617–1655, 3rd Quart., 2016.
[35] F. Boccardi, R. W. Heath, Jr., A. Lozano, T. L. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 74–80, Feb. 2014.
[36] G. I. Klas. (2015). *Fog Computing and Mobile Edge Cloud Gain Momentum Open Fog Consortium ETSI MEC and Cloudlets*. [Online]. Available: http://yucianga.info/?p=938
[37] F. Bonomi, R. Milito, P. Natarajan, and J. Zhu, "Fog computing: A platform for Internet of Things and analytics," in *Big Data and Internet of Things: A Roadmap for Smart Environments*. New York, NY, USA: Springer, 2014, pp. 169–186.
[38] S. Wang, X. Zhang, J. Zhang, J. Feng, W. Wang, and K. Xin, "An approach for spatial-temporal traffic modeling in mobile cellular networks," in *Proc. 27th Int. Teletraffic Congr.*, Ghent, Belgium, 2015, pp. 203–209.
[39] E. Bastug, M. Bennis, and M. Debbah, "Living on the edge: The role of proactive caching in 5G wireless networks," *IEEE Commun. Mag.*, vol. 52, no. 8, pp. 82–89, Aug. 2014.
[40] M. Patel *et al.*, "Mobile-edge computing," Eur. Telecommun. Standards Inst., Sophia Antipolis, France, White Paper, 2014.
[41] M. Chiang. (2016). "Fog networking: An overview on research opportunities." [Online]. Available: https://arxiv.org/abs/1601.00835
[42] M. Chiang and T. Zhang, "Fog and IoT: An overview of research opportunities," *IEEE Internet Things J.*, vol. 3, no. 6, pp. 854–864, Dec. 2016.
[43] R. Tandon and O. Simeone, "Harnessing cloud and edge synergies: Toward an information theory of fog radio access networks," *IEEE Commun. Mag.*, vol. 54, no. 8, pp. 44–50, Aug. 2016.
[44] M. Satyanarayanan *et al.*, "An open ecosystem for mobile-cloud convergence," *IEEE Commun. Mag.*, vol. 53, no. 3, pp. 63–70, Mar. 2015.
[45] Y. Jararweh, A. Doulat, O. AlQudah, E. Ahmed, M. Al-Ayyoub, and E. Benkhelifa, "The future of mobile cloud computing: Integrating cloudlets and mobile edge computing," in *Proc. 23rd Int. Conf. Telecommun. (ICT)*, Thessaloniki, Greece, 2016, pp. 1–5.
[46] B. P. Rimal, D. P. Van, and M. Maier, "Mobile-edge computing vs. centralized cloud computing in fiber-wireless access networks," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, San Francisco, CA, USA, Apr. 2016, pp. 991–996.
[47] W. Hu *et al.*, "Quantifying the impact of edge computing on mobile applications," in *Proc. 7th ACM SIGOPS Asia–Pacific Workshop Syst.*, Hong Kong, 2016, pp. 1–8.
[48] A. Mehta, W. Tärneberg, C. Klein, J. Tordsson, M. Kihl, and E. Elmroth, "How beneficial are intermediate layer data centers in mobile edge networks?" in *Proc. IEEE Int. Workshops Found. Appl. Self Syst.*, Sep. 2016, pp. 222–229.
[49] F. Jalali, K. Hinton, R. Ayre, T. Alpcan, and R. S. Tucker, "Fog computing May help to save energy in cloud computing," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 5, pp. 1728–1739, May 2016.
[50] S. Nunna *et al.*, "Enabling real-time context-aware collaboration through 5G and mobile edge computing," in *Proc. 12th Int. Conf. Inf. Technol. New Generat. (ITNG)*, Las Vegas, NV, USA, 2015, pp. 601–605.
[51] K. Zhang *et al.*, "Energy-efficient offloading for mobile edge computing in 5G heterogeneous networks," *IEEE Access*, vol. 4, pp. 5896–5907, 2016.
[52] P. Borylo, A. Lason, J. Rzasa, A. Szymanski, and A. Jajszczyk, "Energy-aware fog and cloud interplay supported by wide area software defined networking," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kuala Lumpur, Malaysia, May 2016, pp. 1–7.
[53] A. Osseiran *et al.*, "Scenarios for 5G mobile and wireless communications: The vision of the METIS project," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 26–35, May 2014.
[54] P. K. Agyapong, M. Iwamura, D. Staehle, W. Kiess, and A. Benjebbour, "Design considerations for a 5G network architecture," *IEEE Commun. Mag.*, vol. 52, no. 11, pp. 65–75, Nov. 2014.
[55] V. B. C. Souza, W. Ramírez, X. Masip-Bruin, E. Marín-Tordera, G. Ren, and G. Tashakor, "Handling service allocation in combined fog-cloud scenarios," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kuala Lumpur, Malaysia, May 2016, pp. 1–5.
[56] P. Mach and Z. Becvar, "Cloud-aware power control for real-time application offloading in mobile edge computing," *Trans. Emerg. Telecommun. Technol.*, vol. 27, no. 5, pp. 648–661, Dec. 2015.
[57] J. Liu, Y. Mao, J. Zhang, and K. B. Letaief, "Delay-optimal computation task scheduling for mobile-edge computing systems," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Barcelona, Spain, Jul. 2016, pp. 1451–1455.
[58] J. O. Fajardo, I. Taboada, and F. Liberal, "Radio-aware service-level scheduling to minimize downlink traffic delay through mobile edge computing," in *Proc. Int. Conf. Mobile Netw. Manage.*, 2015, pp. 121–134.
[59] Y. Mao, J. Zhang, and K. B. Letaief, "Dynamic computation offloading for mobile-edge computing with energy harvesting devices," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 12, pp. 3590–3605, Dec. 2016.
[60] J. Oueis, E. C. Strinati, and S. Barbarossa, "Distributed mobile cloud computing: A multi-user clustering solution," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kuala Lumpur, Malaysia, May 2016, pp. 1–6.
[61] K. Zhang, Y. Mao, S. Leng, A. Vinel, and Y. Zhang, "Delay constrained offloading for mobile edge computing in cloud-enabled vehicular networks," in *Proc. 8th Int. Workshop Resilient Netw. Design Modeling (RNDM)*, Halmstad, Sweden, Sep. 2016, pp. 288–294.
[62] K. Habak, M. Ammar, K. A. Harras, and E. Zegura, "Femto clouds: Leveraging mobile devices to provide cloud service at the edge," in *Proc. IEEE 8th Int. Conf. Cloud Comput.*, New York, NY, USA, Jun./Jul. 2015, pp. 9–16.

[63] Y. Wang, M. Sheng, X. Wang, L. Wang, and J. Li, "Mobile-edge computing: Partial computation offloading using dynamic voltage scaling," *IEEE Trans. Commun.*, vol. 64, no. 10, pp. 4268–4282, Oct. 2016.

[64] X. Chen, L. Jiao, W. Li, and X. Fu, "Efficient multi-user computation offloading for mobile-edge cloud computing," *IEEE/ACM Trans. Netw.*, vol. 24, no. 5, pp. 2795–2808, Oct. 2016.

[65] S. Sardellitti, G. Scutari, and S. Barbarossa, "Joint optimization of radio and computational resources for multicell mobile-edge computing," *IEEE Trans. Signal Inf. Process. Over Netw.*, vol. 1, no. 2, pp. 89–103, Jun. 2015.

[66] C. You, K. Huang, H. Chae, and B. H. Kim. (2016). "Energy-efficient resource allocation for mobile-edge computation offloading." [Online]. Available: https://arxiv.org/abs/1605.08518

[67] I. Ketykó, L. Kecskés, C. Nemes and L. Farkas, "Multi-user computation offloading as multiple knapsack problem for 5G mobile edge computing," in *Proc. Eur. Conf. Netw. Commun. (EuCNC)*, Athens, Greece, 2016, pp. 225–229.

[68] M. Deng, H. Tian, and X. Lyu, "Adaptive sequential offloading game for multi-cell mobile edge computing," in *Proc. 23rd Int. Conf. Telecommun. (ICT)*, Thessaloniki, Greece, 2016, pp. 1–5.

[69] M. H. U. Rehman, C. Sun, T. Y. Wah, A. Iqbal, and P. P. Jayaraman, "Opportunistic computation offloading in mobile edge cloud computing environments," in *Proc. 17th IEEE Int. Conf. Mobile Data Manage. (MDM)*, Porto, Portugal, Jun. 2016, pp. 208–213.

[70] S. Abdelwahab, B. Hamdaoui, M. Guizani, and T. Znati, "Replisom: Disciplined tiny memory replication for massive IoT devices in LTE edge cloud," *IEEE Internet Things J.*, vol. 3, no. 3, pp. 327–338, Jun. 2016.

[71] M. T. Beck, S. Feld, A. Fichtner, C. Linnhoff-Popien, and T. Schimper, "ME-VoLTE: Network functions for energy-efficient video transcoding at the mobile edge," in *Proc. 8th Int. Conf. Intell. Next Generat. Netw. (ICIN)*, Paris, France, 2015, pp. 38–44.

[72] J. Oueis, E. C. Strinati, and S. Barbarossa, "The fog balancing: Load distribution for small cell cloud computing," in *Proc. IEEE 81st Veh. Technol. Conf. (VTC Spring)*, Glasgow, Scotland, May 2015, pp. 1–6.

[73] L. Tianze, W. Muqing, and Z. Min, "Consumption considered optimal scheme for task offloading in mobile edge computing," in *Proc. 23rd Int. Conf. Telecommun. (ICT)*, Thessaloniki, Greece, 2016, pp. 1–6.

[74] M. Chen, Y. Hao, M. Qiu, J. Song, D. Wu, and I. Humar, "Mobility-aware caching and computation offloading in 5G ultra-dense cellular networks," *Sensors*, vol. 16, no. 7, pp. 974–986, Jun. 2016.

[75] Y. Li and W. Wang, "Can mobile cloudlets support mobile applications?" in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, Toronto, ON, Canada, Apr./May 2014, pp. 1060–1068.

[76] R. Deng, R. Lu, C. Lai, and T. H. Luan, "Towards power consumption-delay tradeoff by workload allocation in cloud-fog computing," in *Proc. IEEE Int. Conf. Commun. (ICC)*, London, U.K., Jun. 2015, pp. 3909–3914.

[77] I. Farris, L. Militano, M. Nitti, L. Atzori, and A. Iera, "Federated edge-assisted mobile clouds for service provisioning in heterogeneous IoT environments," in *Proc. IEEE 2nd World Forum Internet Things (WF-IoT)*, Milan, Italy, Dec. 2015, pp. 591–596.

[78] Intel and Nokia Siemens Networks. *Increasing Mobile Operators' Value Proposition With Edge Computing*, Accessed on Oct. 5, 2016. [Online]. Available: http://www.intel.co.id/content/dam/www/public/us/en/documents/technology-briefs/edge-computing-tech-brief.pdf

[79] V. Gazis, A. Leonardi, K. Mathioudakis, K. Sasloglou, P. Kikiras, and R. Sudhaakar, "Components of fog computing in an industrial Internet of things context," in *Proc. 12th Annu. IEEE Int. Conf. Sens., Commun., Netw.-Workshops (SECON Workshops)*, Seattle, WA, USA, Jun. 2015, pp. 1–6.

[80] C. Vallati, A. Virdis, E. Mingozzi, and G. Stea, "Exploiting LTE D2D communications in M2M Fog platforms: Deployment and practical issues," in *Proc. IEEE 2nd World Forum Internet Things (WF-IoT)*, Milan, Italy, Dec. 2015, pp. 585–590.

[81] B. Xia, C. Yang, and C. Yang, "Modeling and analysis for cache-enabled networks with dynamic traffic," *IEEE Commun. Lett.*, vol. 20, no. 12, pp. 2506–2509, Sep. 2016.

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

[83] J. Gu, W. Wang, A. Huang, and H. Shan, "Proactive storage at caching-enable base stations in cellular networks," in *Proc. IEEE 24th Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun. (PIMRC)*, London, U.K., Sep. 2013, pp. 1543–1547.

[84] E. Baştuğ, M. Bennis, and M. Debbah, "Cache-enabled small cell networks: Modeling and tradeoffs," in *Proc. 11th Int. Symp. Wireless Commun. Syst. (ISWCS)*, Barcelona, Spain, 2014, pp. 649–653.

[85] P. Blasco and D. Gündüz, "Learning-based optimization of cache content in a small cell base station," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Sydney, NSW, Australia, Jun. 2014, pp. 1897–1903.

[86] N. Golrezaei, K. Shanmugam, A. G. Dimakis, A. F. Molisch, and G. Caire, "FemtoCaching: Wireless video content delivery through distributed caching helpers," in *Proc. IEEE INFOCOM*, Orlando, FL, USA, Mar. 2012, pp. 1107–1115.

[87] K. Poularakis, G. Iosifidis, V. Sourlas, and L. Tassiulas, "Multicast-aware caching for small cell networks," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Istanbul, Turkey, Apr. 2014, pp. 2300–2305.

[88] K. Poularakis, G. Iosifidis, and L. Tassiulas, "Approximation caching and routing algorithms for massive mobile data delivery," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Atlanta, GA, USA, Dec. 2013, pp. 3534–3539.

[89] B. Bai, L. Wang, Z. Han, W. Chen, and T. Svensson, "Caching based socially-aware D2D communications in wireless content delivery networks: A hypergraph framework," *IEEE Wireless Commun.*, vol. 23, no. 4, pp. 74–81, Aug. 2016.

[90] B. Chen, C. Yang, and G. Wang, "Cooperative device-to-device communications with caching," in *Proc. IEEE 83rd Veh. Technol. Conf. (VTC Spring)*, Nanjing, China, May 2016, pp. 1–5.

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

[92] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and Zipf-like distributions: Evidence and implications," in *Proc. 18th Annu. Joint Conf. IEEE Comput. Commun. Soc. (INFOCOM)*, New York, NY, USA, Mar. 1999, pp. 126–134.

[93] S. Traverso *et al.*, "Temporal locality in today's content caching: Why it matters and how to model it," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 43, no. 5, pp. 5–12, Oct. 2013.

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

[95] A. Ioannou and S. Weber, "A survey of caching policies and forwarding mechanisms in information-centric networking," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 4, pp. 2847–2886, 4th Quart., 2016.

[96] N. Laoutaris. (2007). "A closed-form method for LRU replacement under generalized power-law demand." [Online]. Available: https://arxiv.org/abs/0705.1970

[97] A. Sengupta, S. Amuru, R. Tandon, R. M. Buehrer, and T. C. Clancy, "Learning distributed caching strategies in small cell networks," in *Proc. 11th Int. Symp. Wireless Commun. Syst. (ISWCS)*, Barcelona, Spain, 2014, pp. 917–921.

[98] J. Gu, W. Wang, A. Huang, H. Shan, and Z. Zhang, "Distributed cache replacement for caching-enable base stations in cellular networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Sydney, NSW, Australia, Jun. 2014, pp. 2648–2653.

[99] S. Borst, V. Gupta, and A. Walid, "Distributed caching algorithms for content distribution networks," in *Proc. IEEE INFOCOM*, San Diego, CA, USA, Mar. 2010, pp. 1–9.

[100] W. Jiang, G. Feng, and S. Qin, "Optimal cooperative content caching and delivery policy for heterogeneous cellular networks," *IEEE Trans. Mobile Comput.*, to be published.

[101] R. Yu *et al.*, "Enhancing software-defined RAN with collaborative caching and scalable video coding," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kuala Lumpur, Malaysia, May 2016, pp. 1–6.

[102] S. Wang, X. Zhang, K. Yang, L. Wang, and W. Wang, "Distributed edge caching scheme considering the tradeoff between the diversity and redundancy of cached content," in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC)*, Shenzhen, China, Nov. 2015, pp. 1–5.

[103] P. Ostovari, A. Khreishah, and J. Wu, "Cache content placement using triangular network coding," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Shanghai, China, Apr. 2013, pp. 1375–1380.

[104] S. Vural, P. Navaratnam, N. Wang, C. Wang, L. Dong, and R. Tafazolli, "In-network caching of Internet-of-Things data," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Sydney, NSW, Australia, Jun. 2014, pp. 3185–3190.

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

[106] Y. Guan, Y. Xiao, H. Feng, C.-C. Shen, and L. J. Cimini, "MobiCacher: Mobility-aware content caching in small-cell networks," in *Proc. IEEE Global Commun. Conf.*, Austin, TX, USA, Dec. 2014, pp. 4537–4542.

[107] C. Jarray and A. Giovanidis, "The effects of mobility on the hit performance of cached D2D networks," in *Proc. 14th Int. Symp. Modeling Optim. Mobile, Ad Hoc, Wireless Netw. (WiOpt)*, Tempe, AZ, USA, 2016, pp. 1–8.

[108] K. Poularakis and L. Tassiulas, "Exploiting user mobility for wireless content delivery," in *Proc. IEEE Int. Symp. Inf. Theory*, Istanbul, Turkey, Jul. 2013, pp. 1017–1021.

[109] H. Li and D. Hu, "Mobility prediction based seamless RAN-cache handover in HetNet," in *Proc. IEEE Wireless Commun. Netw. Conf.*, Doha, Qatar, Apr. 2016, pp. 1–7.

[110] A. S. Gomes *et al.*, "Edge caching with mobility prediction in virtualized LTE mobile networks," *Future Generat. Comput. Syst.*, vol. 70, pp. 148–162, May 2017.

[111] R. Tandon and O. Simeone, "Cloud-aided wireless networks with edge caching: Fundamental latency trade-offs in fog radio access networks," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Barcelona, Spain, Jun. 2016, pp. 2029–2033.

[112] D. Liu and C. Yang, "Cache-enabled heterogeneous cellular networks: Comparison and tradeoffs," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kuala Lumpur, Malaysia, Feb. 2016, pp. 1–6.

[113] D. Liu and C. Yang, "Will caching at base station improve energy efficiency of downlink transmission?" in *Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP)*, Atlanta, GA, USA, Dec. 2014, pp. 173–177.

[114] J. Zhang, X. Zhang, and W. Wang, "Cache-enabled software defined heterogeneous networks for green and flexible 5G networks," *IEEE Access*, vol. 4, pp. 3591–3604, 2016.

[115] J. Q. Fajardo *et al.*, "Introducing mobile edge computing capabilities through distributed 5G cloud enabled small cells," *Mobile Netw. Appl.*, vol. 21, no. 4, pp. 564–574, 2016.

[116] J. Qiao, Y. He, and X. S. Shen, "Proactive caching for mobile video streaming in millimeter wave 5G networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 10, pp. 7187–7198, Oct. 2016.

[117] K. Doppler, M. Rinne, C. Wijting, C. B. Ribeiro, and K. Hugl, "Device-to-device communication as an underlay to LTE-advanced networks," *IEEE Commun. Mag.*, vol. 47, no. 12, pp. 42–49, Dec. 2009.

[118] G. Fodor *et al.*, "Design aspects of network assisted device-to-device communications," *IEEE Commun. Mag.*, vol. 50, no. 3, pp. 170–177, Mar. 2012.

[119] W. Cao, G. Feng, S. Qin, and M. Yan, "Cellular offloading in heterogeneous mobile networks with D2D communication assistance," *IEEE Trans. Veh. Technol.*, to be published.

[120] *Feasibility Study for Proximity Services (ProSe) (Release 12)*, 3GPP, document TR 22.803, Jun. 2013.

[121] A. Laya, K. Wang, A. A. Widaa, J. Alonso-Zarate, J. Markendahl, and L. Alonso, "Device-to-device communications and small cells: Enabling spectrum reuse for dense networks," *IEEE Wireless Commun.*, vol. 21, no. 4, pp. 98–105, Aug. 2014.

[122] N. Zhao, X. Liu, F. R. Yu, M. Li, and V. C. M. Leung, "Communications, caching, and computing oriented small cell networks with interference alignment," *IEEE Commun. Mag.*, vol. 54, no. 9, pp. 29–35, Sep. 2016.

[123] C. Liu *et al.*, "Video content redundancy elimination based on the convergence of computing, communication and cache," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2016, pp. 1–6.

[124] S. Andreev *et al.*, "Exploring synergy between communications, caching, and computing in 5G-grade deployments," *IEEE Commun. Mag.*, vol. 54, no. 8, pp. 60–69, Aug. 2016.

[125] J. Zhu, D. S. Chan, M. S. Prabhu, P. Natarajan, H. Hu, and F. Bonomi, "Improving Web sites performance using edge servers in fog computing architecture," in *Proc. IEEE 7th Int. Symp. Service Oriented Syst. Eng. (SOSE)*, Redwood City, CA, USA, Mar. 2013, pp. 320–323.

[126] ETSI. (2015). *Mobile-Edge Computing (MEC); Service Scenarios* [Online]. Available: http://www.etsi.org/deliver/etsi_gs/MEC-IEG/001_099/004/01.01.01_60/gs_MEC-IEG004v010101p.pdf

[127] O. Mäkinen, "Streaming at the edge: Local service concepts utilizing mobile edge computing," in *Proc. 9th Int. Conf. Next Generat. Mobile Appl., Services Technol.*, Cambridge, MA, USA, Sep. 2015, pp. 1–6.

[128] A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash, "Internet of Things: A survey on enabling technologies, protocols, and applications," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 4, pp. 2347–2376, 4th Quart., 2015.

[129] M. Satyanarayanan *et al.*, "Edge analytics in the Internet of Things," *IEEE Pervasive Comput.*, vol. 14, no. 2, pp. 24–31, Apr./Jun. 2015.

[130] D. Sabella, A. Vaillant, P. Kuure, U. Rauschenbach, and F. Giust, "Mobile-edge computing architecture: The role of MEC in the Internet of Things," *IEEE Consum. Electron. Mag.*, vol. 5, no. 4, pp. 84–91, Oct. 2016.

[131] P. Corcoran and S. K. Datta, "Mobile-edge computing and the Internet of Things for consumers: Extending cloud computing and services to the edge of the network," *IEEE Consum. Electron. Mag.*, vol. 5, no. 4, pp. 73–74, Oct. 2016.

[132] A. V. Dastjerdi and R. Buyya, "Fog computing: Helping the Internet of Things realize its potential," *Computer*, vol. 49, no. 8, pp. 112–116, Aug. 2016.

[133] N. Kumar, S. Zeadally, and J. J. P. C. Rodrigues, "Vehicular delay-tolerant networks for smart grid data management using mobile edge computing," *IEEE Commun. Mag.*, vol. 54, no. 10, pp. 60–66, Oct. 2016.

[134] C. Vallati, A. Virdis, E. Mingozzi, and G. Stea, "Mobile-edge computing come home connecting things in future smart homes using LTE device-to-device communications," *IEEE Consum. Electron. Mag.*, vol. 5, no. 4, pp. 77–83, Oct. 2016.

[135] M. Sapienza, E. Guardo, M. Cavallo, G. La Torre, G. Leombruno, and O. Tomarchio, "Solving critical events through mobile edge computing: An approach for smart cities," in *Proc. IEEE Int. Conf. Smart Comput. (SMARTCOMP)*, St. Louis, MO, USA, May 2016, pp. 1–5.

[136] M. Satyanarayanan, Z. Chen, K. Ha, W. Hu, W. Richter, and P. Pillai, "Cloudlets: At the leading edge of mobile-cloud convergence," in *Proc. 6th Int. Conf. Mobile Comput. Appl. Services (MobiCASE)*, Austin, TX, USA, 2014, pp. 1–9.

[137] A. Botta, W. de Donato, V. Persico, and A. Pescapé, "On the integration of cloud computing and Internet of Things," in *Proc. Int. Conf. Future Internet Things Cloud (FiCloud)*, Barcelona, Spain, Aug. 2014, pp. 23–30.

[138] AT&T, "Domain 2.0 vision white paper," Dallas, TX, USA, AT&T. White Paper, Aug. 2013.

[139] Y. Jararweh, A. Doulat, A. Darabseh, M. Alsmirat, M. Al-Ayyoub, and E. Benkhelifa, "SDMEC: Software defined system for mobile edge computing," in *Proc. IEEE Int. Conf. Cloud Eng. Workshop (ICEW)*, Berlin, Germany, Apr. 2016, pp. 88–93.

[140] O. Salman, I. Elhajj, A. Kayssi, and A. Chehab, "Edge computing enabling the Internet of Things," in *Proc. IEEE 2nd World Forum Internet Things (WF-IoT)*, Milan, Italy, Dec. 2015, pp. 603–608.

[141] M. T. Beck and M. Marco, "Mobile edge computing: Challenges for future virtual network embedding algorithms," in *Proc. 8th Int. Conf. Adv. Eng. Comput. Appl. Sci. (ADVCOMP)*, 2014, pp. 65–70.

[142] C. Cui, H. Deng, D. Telekom, U. Mihel, and H. Damker, "Network function virtualisation: Network operator perspectives on industry progress," ISGNFV, ETSI, Sophia Antipolis, France, Updated White Paper, 2013.

[143] T. Zhao, S. Zhou, X. Guo, Y. Zhao, and Z. Niu, "Pricing policy and computational resource provisioning for delay-aware mobile edge computing," in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC)*, Chengdu, China, Jul. 2016, pp. 1–6.

[144] J. Li, H. Chen, Y. Chen, Z. Lin, B. Vucetic, and L. Hanzo, "Pricing and resource allocation via game theory for a small-cell video caching system," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 8, pp. 2115–2129, Aug. 2016.

[145] E. Cau *et al.*, "Efficient exploitation of mobile edge computing for virtualized 5G in EPC architectures," in *Proc. 4th IEEE Int. Conf. Mobile Cloud Comput., Services, Eng. (MobileCloud)*, Oxford, U.K., 2016, pp. 100–109.

[146] A. Mtibaa, K. Harras, and H. Alnuweiri, "Friend or Foe? Detecting and isolating malicious nodes in mobile edge computing platforms," in *Proc. IEEE 7th Int. Conf. Cloud Comput. Technol. Sci. (CloudCom)*, Vancouver, BC, USA, 2015, pp. 42–49.

[147] H. Deng, W. Li, and D. P. Agrawal, "Routing security in wireless ad hoc networks," *IEEE Commun. Mag.*, vol. 40, no. 10, pp. 70–75, Oct. 2002.

[148] I. Stojmenovic and S. Wen, "The fog computing paradigm: Scenarios and security issues," in *Proc. Fed. Conf. Comput. Sci. Inf. Syst. (FedCSIS)*, Warsaw, Poland, 2014, pp. 1–8.

[149] Y. W. Law, M. Palaniswami, G. Kounga, and A. Lo, "WAKE: Key management scheme for wide-area measurement systems in smart grid," *IEEE Commun. Mag.*, vol. 51, no. 1, pp. 34–41, Jan. 2013.
[150] Z. M. Fadlullah, M. M. Fouda, N. Kato, A. Takeuchi, N. Iwasaki, and Y. Nozaki, "Toward intelligent machine-to-machine communications in smart grid," *IEEE Commun. Mag.*, vol. 49, no. 4, pp. 60–65, Apr. 2011.
[151] R. Lu, X. Liang, X. Li, X. Lin, and X. Shen, "EPPA: An efficient and privacy-preserving aggregation scheme for secure smart grid communications," *IEEE Trans. Parallel Distrib. Syst.*, vol. 23, no. 9, pp. 1621–1631, Sep. 2012.
[152] J. Abdo, J. Demerjian, H. Chaouchi, T. Atechian, and C. Bassil, "Privacy using mobile cloud computing," in *Proc. 5th Int. Conf. Digit. Inf. Commun. Technol. Appl. (DICTAP)*, 2015, pp. 178–182.
[153] M. A. Kader *et al.*, "Leveraging big data analytics for cache-enabled wireless networks," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, San Diego, CA, USA, Dec. 2015, pp. 1–6.
[154] M. A. Maddah-Ali and U. Niesen, "Fundamental limits of caching," *IEEE Trans. Inf. Theory*, vol. 60, no. 5, pp. 2856–2867, May 2014.
[155] B. Schilit, N. Adams, and R. Want, "Context-aware computing applications," in *Proc. 1st Workshop Mobile Comput. Syst. Appl. (WMCSA)*, Santa Cruz, CA, USA, 1994, pp. 85–90.
[156] J. Guo, C. Yao, and C. Yang, "Proactive resource allocation planning with three-levels of context information," in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC)*, Chengdu, China, Jul. 2016, pp. 1–6.
[157] D. Liu, B. Chen, C. Yang, and A. F. Molisch, "Caching at the wireless edge: Design aspects, challenges, and future directions," *IEEE Commun. Mag.*, vol. 54, no. 9, pp. 22–28, Sep. 2016.
[158] F. Pantisano, M. Bennis, W. Saad, and M. Debbah, "Cache-aware user association in backhaul-constrained small cell networks," in *Proc. 12th Int. Symp. Modeling Optim. Mobile, Ad Hoc, Wireless Netw. (WiOpt)*, Hammamet, Tunisia, 2014, pp. 37–42.
[159] S. Yan, M. Peng, and W. Wang, "User access mode selection in fog computing based radio access networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kuala Lumpur, Malaysia, Jun. 2016, pp. 1–6.

SHUO WANG received the B.E. degree in information engineering from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2011, and the M.E. degree in communication and information systems from the China Academy of Space Technology, Beijing, China, in 2014. He is currently pursuing the Ph.D. degree with the Key Laboratory of Universal Wireless Communications, School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing, China. His research interests include 5G network technology, green communication, mobile edge caching, and cloud computing.

XING ZHANG (M'10–SM'14) is currently a Full Professor with the School of Information and Communications Engineering, Beijing University of Posts and Telecommunications, China. He has authored or co-authored two technical books and over 100 papers in top journals and international conferences and holds over 30 patents. His research interests are mainly in 5G wireless communications and networks, green communications, cognitive radio and cooperative communications, big data and Internet of Things. He received the Best Paper Awards in the Nineth International Conference on Communications and Networking in China (Chinacom 2014), the 17th International Symposium on Wireless Personal Multimedia Communications in 2014, and the Eight IEEE International Conference on Wireless Communications and Signal Processing in 2016. He is a Senior Member of the IEEE ComSoc and a member of CCF. He has served as a General Co-Chair of the third IEEE International Conference on Smart Data (SmartData-2017), as a TPC Co-Chair/TPC Member for a number of major international conferences.

YAN ZHANG (M'05–SM'10) received the Ph.D. degree from the School of Electrical and Electronics Engineering, Nanyang Technological University, Singapore. He is currently a Full Professor with the Department of Informatics, University of Oslo, Norway. His current research interests include: next-generation wireless networks leading to 5G, green and secure cyber-physical systems (e.g., smart grid, healthcare, and transport). He is the IEEE VTS (Vehicular Technology Society) Distinguished Lecturer. He is also a Senior Member of the IEEE ComSoc, the IEEE CS, the IEEE PES, and the IEEE VT Society. He is a fellow of the IET. He serves as the chair positions in a number of conferences, including the IEEE GLOBECOM 2017, the IEEE VTC-Spring 2017, the IEEE PIMRC 2016, the IEEE CloudCom 2016, the IEEE ICCC 2016, the IEEE CCNC 2016, the IEEE SmartGridComm 2015, and the IEEE CloudCom 2015. He serves as a TPC Member for numerous international conferences, including the IEEE INFOCOM, the IEEE ICC, the IEEE GLOBECOM, and the IEEE WCNC. He is an Associate Technical Editor of the *IEEE Communications Magazine*, an Editor of the IEEE TRANSACTIONS ON GREEN COMMUNICATIONS AND NETWORKING, an Editor of the IEEE COMMUNICATIONS SURVEYS & TUTORIALS, an Editor of the IEEE INTERNET OF THINGS JOURNAL, and an Associate Editor of the IEEE ACCESS.

LIN WANG received the M.E. degree in information and communication Engineering from the Chongqing University of Posts and Telecommunications, Chongqing, China, in 2009. She is currently pursuing the Ph.D. degree with the Key Laboratory of Universal Wireless Communications, School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing, China. Her research interests are mainly wireless communications and networks, energy harvesting, green communication, and 5G network architecture.

JUWO YANG received the B.S. degree from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2012. He is currently pursuing the Ph.D. degree with the Key Laboratory of Universal Wireless Communications, School of Information and Communication Engineering, BUPT. His current research interests include green communication, 5G network architecture, and the Internet of things technologies.

WENBO WANG received the B.S., M.S., and Ph.D. degrees from BUPT in 1986, 1989, and 1992, respectively. He is currently a Professor with the School of Information and Communications Engineering, and the Executive Vice Dean of the Graduate School, Beijing University of Posts and Telecommunications. He is currently the Assistant Director with the Key Laboratory of Universal Wireless Communication, Ministry of Education. He has authored over 200 journal and international conference papers, and six books. His current research interests include radio transmission technology, wireless network theory, and software radio technology.