Mobile Edge Computing for Ultra-Reliable and Low-Latency Communications

Kai Jiang, Huan Zhou, Xin Chen, and Haijun Zhang

Abstract

Driven by the recent technological advances and the flourishing innovative applications, mobile edge computing (MEC) has emerged as a promising computing paradigm, aiming to extend cloud computing services from the centralized cloud to the edges of networks. Indeed, it is an essential component in the fifth generation architecture, which can empower users with rapid and powerful computation, cache capacity, energy efficiency, mobility, location, and context awareness support. However, as network access methods become diverse, how to ensure ultra-reliable and low-latency communications in MEC is a very challenging task. This article tries to present a comprehensive survey of relevant research and technological developments in MEC. We first highlight the advantages of MEC in comparison with mobile cloud computing, based on which we then make a holistic overview of MEC, including its related architectures and key enablers. Subsequently, we introduce several specific themes and the existing techniques in MEC, where we are particularly concerned with efficient communication techniques and reliable mechanisms in MEC. Furthermore, we elaborate on the evolution of MEC standardization and discuss several typical application scenarios. Finally, we strive to shed light on some challenges and potential research directions for MEC, which may facilitate the transformation of MEC from theory to practice.

Introduction

With the popularity of mobile devices and the advances of wireless access technologies in the fifth generation (5G) networks, mobile applications, especially increasingly computation-intensive applications such as online interactive gaming, facial recognition, assisted driving, and augmented/virtual reality have led to explosive growth of data traffic. Meanwhile, these flourishing applications generally have intense requirements on quality of service (QoS), which brings higher energy consumption than traditional applications [1]. Considering physical size and cost constraints, current end devices have suffered from the limitation of computation resources and energy power, which may become an inevitable bottleneck to address the accompanying challenges in processing extensive applications [2].

Mobile cloud computing (MCC) can enable convenient access to a shared resource pool in the centralized cloud. It was once considered as a promising solution to relieve the increasingly prominent conflict between application requests and resource-constrained end devices. However, the cloud servers are spatially far from end devices, which makes MCC infeasible for latency-sensitive applications as the long transmission distance may contribute to extra cost and delay. Furthermore, despite the continuous efforts that have been spent on enhancing the channel bandwidth, the utilization efficiency of the radio spectrum is notably reaching its theoretical boundary [3]. Thus, such efforts will not be sufficient, and a fundamental innovation that breaks the bottleneck of massive task processing in 5G networks is urgently required.

All these challenges have accelerated the development of mobile edge computing (MEC). Specifically, MEC has emerged as an important component to cope with the innovative applications in the 5G architecture [4], for which it can move computation, caching, and network functions toward the network edges. Compared to MCC, MEC extends cloud computing services (e.g., caching and computation capacity) from the centralized cloud to the edges of networks, and enables users to offload workloads to nearby MEC servers directly by leveraging base stations (BSs) and access points (APs). Such a pattern not only accommodates the expansion requirements of the computation capabilities of end devices, but also improves the QoS of mobile applications with considerably reduced latency and energy consumption [5].

As seen in recent studies [3, 6, 7], the successful realization of MEC is still in its infancy. At the same time, as the network access methods become diverse, how to ensure ultra-reliable and low-latency communications in MEC is a very challenging task. Thus, providing an overview of the state-of-the-art developments will elicit fruitful discussions, provide useful insights into this area's current status, and further inspire more potential research directions in MEC. In this article, we aim to provide a comprehensive survey in this young field with a focus on the ultra-reliable and low-latency communications perspective. We first highlight the advantages of MEC by comparison with MCC, based on which we then provide a holistic overview of MEC, including its related architectures and key enablers. Subsequently, we introduce several specific themes and the existing techniques in MEC, where we are particularly concerned with the efficient communication techniques and reliable mechanisms in MEC. Furthermore, we elaborate on the evolution of MEC standardization and discuss several typical application scenarios. Finally, we strive to shed light on

Digital Object Identifier:
10.1109/MCOMSTD.001.2000045

Kai Jiang, Huan Zhou, and Xin Chen are with China Three Gorges University; Haijun Zhang is with the University of Science and Technology Beijing.

some challenges and potential research directions for MEC, which may facilitate the transformation of MEC from theory to practice.

Overview of Mobile Edge Computing

In this section, we first highlight the advantages of MEC by comparison with MCC, and then we provide a holistic overview of MEC, including its related architectures and key enablers.

Why MEC vs. MCC

There are often significant disparities between MEC and MCC in terms of latency, energy efficiency, and so on. Now we describe the advantages of MEC in detail.

Reducing Delay/Overhead: With the soaring demands for data services, MCC faces the inherent limitation as cloud servers are usually spatially far from end users. In contrast, the computation and storage capabilities in MEC are in proximity to end users, which causes a significant reduction in transmission delay and overhead. Besides, MCC must pass through multiple networks, including radio access networks, backhaul networks, and the Internet. In these networks, flow control, routing, and other network management operations may cause excessive delays. Compared to MCC, MEC is more affordable and accessible, and the delay only occurs in the radio access network. Although cloud servers' computing capacities are orders of magnitude higher than edge servers, their computing capacities have to be shared by more users.

Privacy/Security Enhancement: Compared to MCC, improving the mobile applications' privacy and security is an obvious advantage of MEC. In general, the cloud computing platform can be considered as a large remote public data center, which is more vulnerable due to the high concentricity of users' data resources. Meanwhile, the ownership and management of users' data are separated in MCC, which may explicitly lead to the leakage and loss of private data. Fortunately, edge servers, which are in closer proximity to users, can provide effective and reliable services to avoid these problems. On one hand, MEC servers are typically characterized by distributed deployment, small scale, and more scattered valuable information, which all significantly reduce the likelihood of becoming security attack targets. On the other hand, many MEC servers are private or third-party, which will also alleviate the information leakage by forbidding uploading restricted data to the public data center.

Utilization of Context Awareness: Another significant advantage of MEC is that it can empower users with proximity services, which is to provide end devices with related services based on context awareness. Distributed edge infrastructures deployed at different places in the radio access network are able to obtain the fine-grained information of end users (e.g., real-time behavior, location, and state) by implementing interactions. Based on this information, the network resources can be allocated more efficiently, and the quality of the user's experience can be vastly improved. A typical example is a cooperative cruise control system [14], which uses BS fingerprints to track and analyze the trajectory of a large community of users for traffic-related information management, navigation, and personalized itinerary scheduling.

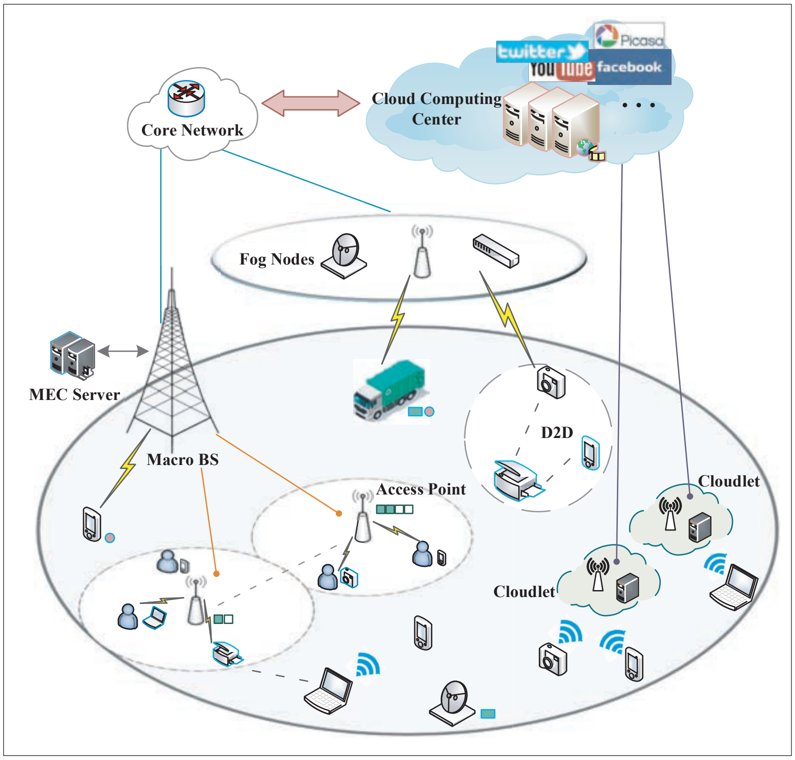

Figure 1. Architecture of MEC.

Architecture of MEC

MEC is conceived to remedy the transmission distance between the high-powered servers and various end devices. Apart from MEC, there have been some other similar edge paradigms, such as cloudlets and fog computing. As shown in Fig. 1, different edge paradigms usually tend to coexist with MEC in many application scenarios.

Cloudlet: As a similar concept that has gained widespread academic acceptance, cloudlet was first proposed by a team at Carnegie Mellon University in 2009 [8]. Cloudlet is a dynamic virtual-machine-based computation offloading system, and its basic principle is to deploy a trusted and resource-rich cluster of computers that are well connected to the Internet in a strategic location at the network edge. This system can realize the important functions of MEC and harness the synergies between different cloudlets. However, the penetration of cloudlet's applications is slower than MEC and fog computing because of the requirement of dedicated devices. Furthermore, it was inadequate that cloudlets must be placed closer to the users due to the limited WiFi coverage. There may be a fatal drawback of the WiFi connection between users and cloudlets. Users have to switch between the mobile network and WiFi when they use cloudlet services, making it hard to satisfy the QoS requirements of applications.

Fog Computing: Fog computing, alternatively known as a generalized form of MEC architecture, was initially introduced by Cisco in 2012. The concept of fog computing is derived from the analogy that the fog is closer to the end users than the clouds. Accordingly, fog computing is an extension of cloud computing to the network

Computing paradigms	Cloudlet	Fog computing	Mobile edge computing
Origin	Carnegie Mellon University, 2009	Cisco, 2012	ETSI, 2015
The relationship with the cloud	An extension of cloud Cooperate with the cloud	An extension of cloud Cooperate with the cloud	Relatively independent of the cloud
Node location	Local or outdoor installation	Varying between end devices and cloud	Base station or access point
Node devices	Data center in a box	Gateways, data centers, access points, vehicles, end devices.	MEC servers running in base stations or access points.
Software architecture	Cloudlet agent is required	Based on fog abstraction layer	Mobile orchestrator is used
Context awareness	Low	Medium	High
Proximity	One hop	One or multiple hops	One hop
Virtualization	Virtual machine	Virtual machine and container	Virtual machine and container
Access mechanism	Wi-Fi	Mobile networks, Wi-Fi, Bluetooth.	Mobile networks
Inter-node connection	Partial	Full	Partial

TABLE 1. Comparison of different edge computing paradigms.

edge, which can provide cloud services with much lower latency. It provides storage, computing, and management functions at the edge of the network [9]. Meanwhile, the notion of fog nodes is pervasive, such as resource-rich data centers, APs, vehicles, and end users. Finally, it is worth noting that although the term fog computing is frequently used in place of MEC, and their areas are overlapping, there are still some limitations in fog computing, as shown in Table 1.

MEC: MEC was first introduced by the European Telecommunications Standards Institute (ETSI) in 2015, and attracted widespread attention in both industry and academia in the 5G era. According to ETSI: "Mobile Edge Computing provides an IT service environment and cloud computing capabilities at the edge of the mobile network, within the Radio Access Network and in close proximity to mobile subscribers." Specifically, MEC can support cloud servers by deploying cloud capacity and network functions to the edges of networks, which provides end users with powerful computation, caching capacity, mobility, location, and context awareness support, and significantly reduces the transmission delay. Moreover, MEC can provide an ultra-low-latency environment with high bandwidth and convenient access to radio and network, which makes it more suitable to meet the responsiveness and privacy requirements of real-time services. As shown in Fig. 1, MEC servers are deployed at the BS. On one hand, they can process arriving task requests and deliver responses to users. On the other hand, they can also forward the task requests to remote data centers, which can provide processing capabilities that the edge servers cannot afford. In brief, MEC promises dramatic potential for developing various emerging applications where low latency and energy consumption are required, bringing innovation and promoting businesses in materializing the 5G vision.

Key Enablers

Several technologies are identified as enabling technologies for MEC paradigm realization, including cloud technology, software-defined networking (SDN), network function virtualization (NFV), network slicing, and so on.

Cloud Technology: MEC extends cloud computing functions to the edge of the mobile network. Advances in cloud technology have made it easier to deploy virtual machines on a large number of general-purpose servers in locations such as BSs and gateways. The cloud can provide powerful processing capabilities and abundant resources. It turns out that the integration of the cloud and the Internet of Things (IoT)is beneficial in delivering new services. MEC is integrated with cloud computing functions to provide an effective solution for management and configuration services.

Software Defined Networking: SDN is an innovation of computer networks, which separates the control layer's function from the data layer. The data layer contains user-generated messages and is responsible for forwarding them using the forwarding table prepared by the control layer. The integration of MEC and SDN can make centralized control more effective and reliable.

Network Function Virtualization: NFV is a supplementary technology of SDN proposed for future 5G network architectures. The purpose of NFV is to use software functions to virtualize network node functions and to transfer network functions from standard general-purpose hardware to computing platforms that can provide the same services as traditional mobile networks. Using NFV can reduce the capital investment and operating costs of network operators. The application of NFV will change the landscape of the telecommunications industry and bring many benefits such as accelerating time to market, optimizing network configuration and topology in near real time, and supporting multi-tenancy.

Network Slicing: Network slicing is a form of agile and virtual network architecture that can separate the network into different network segments, enabling self-contained, multiple logical network instances to be created on top of a common shared physical infrastructure. Each of the network instances is optimized for a specific request, supporting customized network operations and resource isolation. More importantly, the innate heterogeneity of MEC makes network slicing indispensable for ultra-reliable and low-latency communication, for which network slicing can support different services running across a sin-

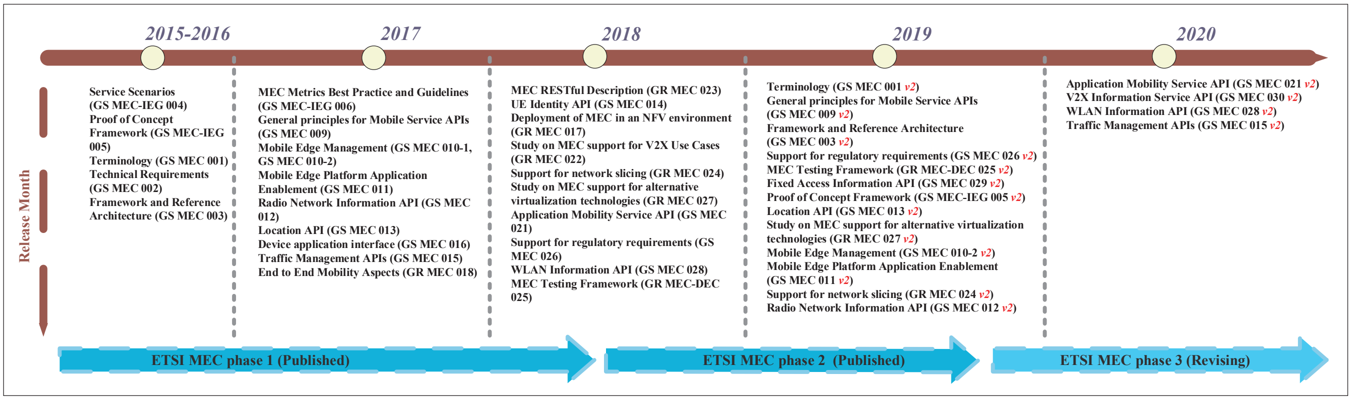
2015-2016
2017
2018
2019
2020
Release Month
Service Scenarios (GS MEC-IEG 004)
Proof of Concept Framework (GS MEC-IEG 005)
Terminology (GS MEC 001)
Technical Requirements (GS MEC 002)
Framework and Reference Architecture (GS MEC 003)
MEC Metrics Best Practice and Guidelines (GS MEC-IEG 006)
General principles for Mobile Service APIs (GS MEC 009)
Mobile Edge Management (GS MEC 010-1, GS MEC 010-2)
Mobile Edge Platform Application Enablement (GS MEC 011)
Radio Network Information API (GS MEC 012)
Location API (GS MEC 013)
Device application interface (GS MEC 016)
Traffic Management APIs (GS MEC 015)
End to End Mobility Aspects (GR MEC 018)
MEC RESTful Description (GR MEC 023)
UE Identity API (GS MEC 014)
Deployment of MEC in an NFV environment (GR MEC 017)
Study on MEC support for V2X Use Cases (GR MEC 022)
Support for network slicing (GR MEC 024)
Study on MEC support for alternative virtualization technologies (GR MEC 027)
Application Mobility Service API (GS MEC 021)
Support for regulatory requirements (GS MEC 026)
WLAN Information API (GS MEC 028)
MEC Testing Framework (GR MEC-DEC 025)
Terminology (GS MEC 001 v2)
General principles for Mobile Service APIs (GS MEC 009 v2)
Framework and Reference Architecture (GS MEC 003 v2)
Support for regulatory requirements (GS MEC 026 v2)
MEC Testing Framework (GR MEC-DEC 025 v2)
Fixed Access Information API (GS MEC 029 v2)
Proof of Concept Framework (GS MEC-IEG 005 v2)
Location API (GS MEC 013 v2)
Study on MEC support for alternative virtualization technologies (GR MEC 027 v2)
Mobile Edge Management (GS MEC 010-2 v2)
Mobile Edge Platform Application Enablement (GS MEC 011 v2)
Support for network slicing (GR MEC 024 v2)
Radio Network Information API (GS MEC 012 v2)
Application Mobility Service API (GS MEC 021 v2)
V2X Information Service API (GS MEC 030 v2)
WLAN Information API (GS MEC 028 v2)
Traffic Management APIs (GS MEC 015 v2)
ETSI MEC phase 1 (Published)
ETSI MEC phase 2 (Published)
ETSI MEC phase 3 (Revising)

Figure 2. The development of ETSI MEC standardization.

gle radio access network [10]. With the functions of network slicing performing in the MEC systems, the applications could be provisioned with flexible and dedicated network resources, as well as the dynamic allocation of network functions. In other words, network slicing not only can reduce the latency substantially, but also support traffic prioritization of MEC service subscribers. Moreover, network slicing can also provide excellent performance guarantees and security in various connections and services. This pattern further improves the flexibility and agility of MEC ecosystems.

Mobile Edge Computing for Ultra-Reliable and Low-Latency Communications

MEC introduces a new way of manipulating computing and storage resources. In this section, we first introduce several specific themes in MEC, including computation offloading, resource allocation, edge caching, and the integration of artificial intelligence (AI) approaches in these aspects. Then we summarize some efficient wireless communication techniques and reliable mechanisms in MEC.

Computation Offloading and Resource Management in MEC Systems

Computation Offloading: As mentioned earlier, the inevitable limitations of end devices make it impossible to processing numerous applications requiring high-speed Internet connectivity and high computation resources. Hence, in such situations, it is feasible to empower end devices with enough resources offered by transferring the resource-intensive tasks to external platforms like cloud and MEC server, which is known as task offloading. One of the most prominent and widely discussed features of MEC is computation offloading, which is able to break the bottlenecks of end devices. By computation offloading, MEC can extend cloud computing services from the centralized cloud to the edges of networks, and enable end devices to offload workloads to nearby MEC servers directly by leveraging BSs or APs. Such a pattern meets the expansion requirements of the computation capabilities of end devices, and improves the QoS of mobile applications with considerably reduced latency and energy consumption [11]. It is worth noting that there are two models to offload the task. One is entirely offloading; in this scenario, the whole process or task migrates to a MEC server. The other is called partial offloading; in this scenario, the task or process is divided into different parts, and some parts may migrate to the server. Moreover, with the recent advances of smart devices and wireless access technologies, more computation resources can be leveraged by using device-to-device (D2D) technology. The computation tasks can not only be offloaded to servers but also to devices by utilizing D2D communication. Thus, in order to make the optimal computation offloading decision, various aspects need to be considered, such as whether the task is offloaded, and where and how to offload the task.

Resource Management: Resource allocation is the technique of scheduling the resources to find the optimal association that will improve the utilization of the available resources as well as satisfy all the requirements of the users, of which various resources in the system may be coordinated basing on the time-varying system conditions, the number of users, and the heterogeneity, priority, and maximum tolerable delay of each computation task. As the computation and communication resources of edge servers are still limited [12], allocating resources on demand is also a crucial factor for exploiting the merits of MEC. In MEC networks, the communication and computation resource allocation are supported to be jointly optimized to enhance the system performance. In practice, however, most of the previous studies addressing the problem of resource allocation for the offloaded computation task to MEC seldom consider the dynamic property of the network. To be more precise, they only focus on the performance in quasi-static systems. Hence, dynamic allocation of the computing resources during the processing of computation offloading in the MEC is an interesting research challenge to be addressed.

Edge Caching: The flourishing emerging applications require users to access huge amounts of contents, especially for some delay-sensitive contents, which leads to significant redundant traffic loads and considerable delay for content delivery in the network. Nevertheless, due to the centralized characteristic of current cellular network architecture, long transmission distance and limited channel bandwidth pose unprecedented challenges for supporting massive content delivery while also satisfying the tight QoS requirements. Stemming from the recent studies, different popular contents require different levels of priority. Explicitly, only a few popular contents are repeatedly

Nowadays, effective inter-symbol interference suppression techniques, such as equalization and spread spectrum, are proposed for reliable transmissions in the MEC system. Moreover, interference alignment is also a potential technique for interference management of small cell networks, which can simplify the topology of interference networks and make interference management easier.

downloaded upon request from the majority of users, while the remaining large portion of the contents impose rather infrequent access demands. This pattern has promoted the implementation of the edge caching technique. Specifically, edge caching has emerged as a promising solution to alleviate the redundant traffic and reduce the content access latency in future MEC networks, which caches popular contents in close proximity to users by utilizing the storage resources at intermediate low-cost caching units, like APs. Hence, it will be possible for users to access popular contents from the caching-enabled nearby APs directly, instead of downloading them from the remote server via backhaul networks. Furthermore, advances in D2D communication will enable the storage unit of the user equipment to share content based on social relationships between other users.

AI-Empowered MEC: In general, a critical issue in the MEC system is user association and its joint optimization with other aspects. However, the joint optimization of user association, offloading decision, cache replacement, and resource allocation is typically an NP-hard problem, of which the complexity is further exacerbated in time-varying MEC systems. To this end, AI-empowered approaches can be used to build intelligent edge for real-time edge management and maintenance. Specifically, AI techniques, such as reinforcement learning and deep learning, have exhibited powerful learning and reasoning ability in solving various problems in MEC, so as to optimize strategies of computation offloading, edge caching, and resource allocation according to the dynamics of the system. On one hand, the integration of different AI algorithms can enhance data analytics, make adaptive decisions, predict arriving tasks, and preserve network security. On the other hand, adopting AI techniques for an MEC system can sense the change of the network conditions in different time domains, adaptively coordinate the configuration of the available resources, as well as realize real-time state collection and efficient workload processing, based on which the intelligent applications for different users can be optimized. In addition, AI-empowered approaches deployed at the MEC servers can also detect contaminated data and improve data privacy. The emerging federated learning is especially suitable to preserving data privacy in large-scale MEC systems, as it can provide distributed offloading decisions and learn the behaviors based on its local downloaded master model [13].

Feasible Wireless Communication Techniques in MEC

Combining efficient wireless communication techniques with MEC would change the design philosophy of communication networks greatly, which further makes certain inter- and intra-domain application cases feasible.

Interference Management: Unlike the wired connections in conventional cloud computing, interference is a crucial challenge in MEC. The wireless links between end devices and the MEC server may be highly time-varying and unstable due to the reflection and refraction from scattering objects in the space. Multipath fading can cause severe inter-symbol interference. In addition, the broadcast nature of wireless transmissions leads to a signal being interfered by other signals occupying the same spectrum, which reduces their respective signal-to-interference-plus-noise ratios and thereby results in the probability of error in detection. Thus, the inherent challenges with interference management should always be considered for the design of MEC systems to seamlessly integrate computation offloading and resource allocation. Nowadays, effective inter-symbol interference suppression techniques, such as equalization and spread spectrum, are proposed for reliable transmission in the MEC system. Moreover, interference alignment is also a potential technique for interference management of small cell networks, which can simplify the topology of interference networks and make interference management easier.

Millimeter-Wave (mmWave) Communication: Due to its enormous potential in 5G systems, the mmWave spectrum (30 GHz to 300 GHz) has been considered as an important communication pillar for the envisioned MEC architecture. The integrations of mmWave technology into the MEC system can achieve higher data rate transmission to perform the MEC functionalities with lower latency. Meanwhile, the coexistence of MEC with mmWave technology can also optimize the performance of mmWave communications by providing useful local computation capacities. In addition, the market deployment of mmWave technology is going to exploit new fields in the MEC architectures. Recently, various infrastructures that integrate MEC with mmWave technology have been proposed [14]. For example, the narrow beams of mmWave APs can effectively improve the channel quality and save communication resources.

Non-Orthogonal Multiple Access (NOMA): NOMA has been viewed as an essential principle for the design of new generation radio access techniques due to its significant spectral efficiency. The combination of NOMA and MEC is another vital research issue, which can improve the user satisfaction and extend the application scenarios of MEC in reality. The key idea of NOMA is that multiple users are encouraged to share the same time-frequency resource to achieve higher spectral efficiency. Meanwhile, NOMA can deal with the multiple access interference by enabling APs and users to perform advanced transceiver designs, such as superposition coding techniques and interference cancellation techniques, respectively. Compared to the conventional orthogonal multiple access (OMA), there are many benefits of applying NOMA to MEC, including achieving higher spectral efficiency, supporting massive wireless connectivity, reducing the transmission latency, and providing relaxed channel feedback. Meanwhile, it is worth noting that NOMA and OMA are compatible in the MEC system.

D2D Communication: D2D communication enables direct transmission between end devices in proximity. It has been recognized as one of the efficient technologies to promote low-latency communication in MEC. The benefits of D2D communication include one-hop communication, higher spectral efficiency, low transmission power, and coverage extension. With D2D communication, massive computation tasks can be either offloaded to APs directly or delivered to other neighboring end devices, which enhances overall network capacity and relieves the backhaul traffic load. Meanwhile, as the computation and storage resources on a large number of end devices can be exploited, the utilization efficiency of these resources can be significantly improved by caching and executing tasks in end devices if the resource sharing and computation-load balancing strategies are carefully designed.

Existing Reliable Mechanisms in MEC

Various vulnerabilities and possible attacks may cause potential problems within the system. Therefore, it is necessary to implement appropriate mechanisms. In this subsection, we present some existing reliable mechanisms for MEC.

Virtualization Security Mechanisms: The virtualization supporting technology is one of the foundations for the MEC paradigm to provide more scalable and reliable, and high-performing services. However, some serious attacks (distributed denial of servce attacks, wireless local area networking application conflicts) would be even more threatening for MEC applications, due to these applications relying heavily on virtualization. More seriously, malicious elements that get access to virtual servers may hijack the entire edge data center. Therefore, it is of paramount importance to secure the virtualization technology. In the existing MEC systems, the virtualization infrastructures will not be controlled by only one entity. Meanwhile, some useful mechanisms such as protection by network abstractions, hypervisor hardening, and isolation policies are used to protect the deployed virtual machines and their hosted physical servers.

Authentication Mechanisms: Different from well-managed cloud service providers, providers in MEC may be hosted by several providers depending on their choices. Therefore, it is inapplicable for MEC nodes under a multi-management domain to authenticate themselves by administrative cloud servers in an open environment. For example, cloud service providers may extend their services to the edge using fixed infrastructures. Multiple providers may deploy different types of MEC servers, and the end users may demand to lease and use the limited resources depending on their own budget. To meet all of these services, every entity with which the system is interacting should be mutually identified to each other. Proper authentication mechanisms are required to evaluate the reliability of MEC servers. At present, some feasible solutions have been proposed to solve the authentication problems, such as Diffie-Hellman key-exchange-based methods and public-key-infrastructure-based methods.

Network Security Mechanisms: A comprehensive network security mechanism is of prime concern for MEC due to the preponderance of various network infrastructures. Specifically, intrusion detection and prevention is an important prerequisite before designing the MEC system. Several network entities should be monitored from internal or external threats as they are vulnerable to any security threats. In such cases, the MEC infrastructures should monitor the overall network and equally coordinate with the surrounding and core networks. To this end, intrusion detection systems can be employed in the MEC data center to analyze the system logs for any unauthorized access. Meanwhile, it can also be employed at the MEC network side to detect and prevent the network from attacks, such as man-in-the-middle attack, denial of service, and port scanning. A MEC server that is located one hop away from end devices can efficiently be meshed to form a security framework to detect any intrusion. It can serve as a proxy for distant cloud servers in case of any unavailability issue due to certain attacks. Moreover, it is easy to reduce network cost and scale network resources by implementing SDN in the MEC system, as well as isolate network traffic and segregate malicious data.

Standardization and Application Scenarios of MEC

Standardization Efforts

Standardization is an indispensable step for the successful promotion of a new technology, which has defined voluntary characteristics and rules in a specific industry. Just like other standardized technologies, the MEC standard helps to innovate cutting-edge techniques and disseminate workable solutions. More importantly, MEC standardization is an essential fingerpost to build the open ecosystem, which will open up a flat path toward favorable market conditions.

Nowadays, MEC has been acknowledged by the European 5G Infrastructure Public Private Partnership (5G PPP) as one of the prime technologies for 5G networks. Various members in the value chain have spent sustained efforts to develop MEC specifications based on industry consensus. Specifically, ETSI is the forerunner of MEC standardization, which published an introductory technical white paper on MEC in September 2014. In this handbook, the concept, referenced platform, technical requirements and challenges, as well as typical application scenarios of MEC are formally proposed and discussed. In December 2014, Huawei, Intel, Nokia Networks, NTT DOCOMO, IBM, and Vodafone in ETSI launched a MEC Industry Specification Group (MEC ISG) to promote the technical standardization development of MEC. This group aimed to build up a standardized and open MEC environment, which will enable the proficient and seamless integration of innovative applications from different vendors or service providers across the MEC platforms. MEC ISG has delivered several specifications for technical requirements, system architecture, server framework, application scenarios, application programming interfaces (APIs), and more. All these efforts have further accelerated the rapid development of edge applications and ultimately expanded the market size. As of early 2019, MEC ISG had completed the first two phases of standardization and was conducting the third phase of standard maintenance and standard additions. Meanwhile, ETSI has also published several technical manuals about MEC. Some noteworthy aspects in the earlier white paper have been rediscussed and documented in the ETSI specifications in 2015. At MEC World Congress 2016, ETSI announced six proofs of concept that are accepted by the MEC ISG, which will assist the strategic planning and decision making of organizations, and help to identify which MEC solutions may be viable in the network. This provides the community with confidence in MEC and will lead the pace of the standardization. It should be noted that the MEC ISG officially extended the terminology of MEC from mobile edge computing

Standardization is an indispensable step for the successful promotion of a new technology, which has defined voluntary characteristics and rules in a specific industry. Just like other standardized technologies, the MEC standard helps to innovate cutting-edge techniques and disseminate workable solutions. More importantly, MEC standardization is an essential fingerpost to build the open ecosystem, which will open up a flat path toward favorable market conditions.

Although the MEC architecture is a new revenue stream for all stakeholders that has not matured sufficiently, application cases are the most direct way to verify whether emerging technologies are valuable. Nowadays, we have witnessed that the key value propositions of MEC technology are exemplified in numerous applications. In this section, we present several typical application scenarios of MEC.

to multi-access edge computing at this congress in order to reap additional benefits of MEC by accommodating more wireless access technologies. After this scope expansion, MEC servers can be deployed by network operators at various locations within RANs and/or collocated with different units of the network edge. This transformation not only pushes intelligence toward the edge so that communication, computation, caching, and control services can be better facilitated, but also conveniently retains the acronym of MEC, which has become widely recognized among stakeholders in the industry. Most recently, the 3rd Generation Partnership Project (3GPP) added MEC into its 5G standardization in 3GPP TS 23.501, and a recent technical specification contribution in 3GPP has clarified and reported how to seamlessly deploy and integrate MEC function into emerging 5G architecture.

More specific details in the evolution of the MEC standardization are exhibited in Fig. 2. Overall, the MEC standardization is still in its infancy and demands constant efforts from all aspects. Not only academic theories, but also more realistic challenges and requirements need to be considered.

Some Application Scenarios

Although the MEC architecture is a new revenue stream for all stakeholders that has not matured sufficiently, application cases are the most direct way to verify whether emerging technologies are valuable. Nowadays, we have witnessed that the key value propositions of MEC technology are exemplified in numerous applications. In this section, we present several typical application scenarios of MEC [15].

Augmented Reality (AR) and Virtual Reality (VR): The AR and VR technologies have been anticipated to be the most hopeful applications that will change the way we live; they can enable users to interact with the virtual scenarios by using virtualization. Nevertheless, these applications usually have intense demand on computation capacity and real time. Considering the resource limitation and latency constraints, there are many inevitable bottlenecks in the spreading of AR and VR applications, as the currently adopted devices are not ready to allow these types of scenes. However, the adoption of MEC may turn the tables. Indeed, MEC is very suitable to be applied in the AR and VR domains. The MEC server can exploit local context-aware information and always pose powerful processing abilities to deal with latency-sensitive applications. Specifically, MEC is capable of performing more optimized and efficient low-latency services for AR and VR applications. When we offload the computation-intensive tasks of AR and VR applications to the nearest MEC server, the server will be able to provide a significant amount of computation resources and accurately analyze the input data, then transmit the outcomes back to the end users. In this way, the applications can reduce delay and give a faster response to the user, while further sustaining more positive and immersive user experiences.

Connected Vehicles: Along with the recent advances in wireless access technologies and the steady growth of vehicles on the roads, more and more vehicles are facilitated to connect with each other or the roadside units (RSUs) on the road, which have paved a path toward data-driven intelligent transportation, as well as enhanced driving safety, relieved traffic congestion, and sensed vehicles' behaviors, and provided opportunities for numerous value-added services. However, the maturity of such technology is not yet to come as the latency requirement cannot be satisfied with the existing connected cloud server. Therefore, MEC has become a key enabling technology for connected vehicles by adding computation and geo-distributed services to RSUs. Deploying MEC environments alongside the road can play an important role in connecting moving vehicles, vehicle-to-everything communication, and automotive safety services. On one hand, the applications run on MEC servers in close proximity to the vehicles and can provide low-latency, high-reliability communication and fast-response services for vehicles. 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. On the other hand, AI-empowered vehicular edge networks have exhibited fascinating potential for handling various smart vehicle applications. Further, they are expected to optimize the complicated Internet of Vehicles systems and make automatic driving possible. In addition, MEC enables scalable, reliable, and distributed environments that are synced with the local sensors.

Big Data Analytics: Big data consists of large and complex datasets that are generated by applications, sensors, devices, video channels, and social media. These datasets are heterogeneous, and it may not be possible to integrate them with existing facilities, interoperability, and so on. Big data analytics is a process of extracting meaningful information from raw data, which could be helpful for various aspects of daily life and work. MEC is a distributed intelligent platform that provides distributed networking, ample computation, and caching resources at the client layer. Implementing the MEC server near users allows big data analytics at the more capable edge platforms rather than at the source producing data, which can elevate data analysis capabilities with the help of high bandwidth/computation and low power consumption. Meanwhile, edge caching could be leveraged for avoiding numerous redundant traffic loads for content delivery in the networks. In addition, the MEC server can collect information from multiple sources, which helps those devices perform multiple modularized tasks. Compared to big data analytics performed at the centralized core network, doing big data analytics at the network edge will reduce power consumption and network latency.

Open Challenges

Although MEC can benefit us tremendously, some challenges still need to be resolved before its commercial deployment. In this section, we look at the open challenges in MEC.

Resource Allocation and Joint Optimization: By upgrading the cloud infrastructure to the edge of the network, MEC integrates fewer resources than the cloud. Applications and virtualized MEC servers support computation offloading. However, due to the heterogeneous processor architecture, computing tasks will bring additional loads. For example, mobile smartphones mainly have ARM and x86 architectures, so they need translation or emulation. Furthermore, optimized mechanisms still need to enhance the performance of inherently limited resources.

Security and Privacy Issues: Although MEC can improve security and privacy compared to MCC, MEC still faces its own security and privacy challenges. First, MEC can be collocated with different heterogeneous network elements, so the conventional privacy and security mechanisms already running in MCC are not suitable for MEC systems. Second, because wireless eavesdroppers may eavesdrop on computing tasks, it may not be safe to offload tasks through wireless channels. The transmission of application data can be protected by encrypting on the user side and decrypting on the target server side. However, this may increase transmission delay and execution delay, thereby reducing application performance. Finally, sharing the same storage and computing resources among multiple mobile users can lead to leakage and loss of private data.

The transmission of application data can be protected by encrypting on the user side and decrypting on the target server side. However, this may increase transmission delay and execution delay, thereby reducing application performance. Finally, sharing the same storage and computing resources among multiple mobile users can lead to the leakage and loss of private data.

Other Challenges: In addition, there are other challenges that are also critical to enhance the MEC framework:
- *Pricing Policy*: In MEC, the storage, computing, and communication resources are allocated dynamically according to users' demands. Thus, the optimal pricing policy is different from legacy systems. From a commercial perspective, the profits of all the stakeholders in the system should be balanced.
- *Transparent Workload Migration*: As mentioned earlier, the users' workloads can be offloaded to the MEC servers for execution. Migrating these workloads transparently is critical to the usability of latency-sensitive applications, such as real-time applications.
- *Openness of the Network*: The mobile network has full authorization to the network, but the network must be opened for third-party providers because of the possible security risks of MEC.

Conclusion

In this article, we have conducted a comprehensive survey of the recent research in MEC. Specifically, we first highlight the advantages of MEC by comparison with MCC, based on which we then offer a holistic overview of MEC, including its related architectures and key enablers. Subsequently, we introduce several specific themes and the existing techniques in MEC, where we are particularly concerned with the efficient communication techniques and reliable mechanisms in MEC. Furthermore, we elaborate on the evolution of MEC standardization and discuss several typical application scenarios. Finally, we strive to shed light on some challenges and potential research directions for MEC, which may facilitate the transformation of MEC from theory to practice.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grant No. 61872221. The correspoinding author is H. Zhou.

References

[1] N. Abbas, Y. Zhang, and A. Taherkordi, "Mobile Edge Computing: A Survey," *IEEE IoT J.*, vol. 5, no. 1, 2018, pp. 450–65.

[2] Q. Pham *et al.*, "A Survey of Multi-Access Edge Computing in 5G and Beyond: Fundamentals, Technology Integration, and State-of-the-Art," *IEEE Access*, vol. 8, 2020, pp. 116,974–117,017.

[3] G. Qiao *et al.*, "Deep Reinforcement Learning for Cooperative Content Caching in Vehicular Edge Computing and Networks," *IEEE IoT J.*, vol. 7, no. 1, 2020, pp. 247–57.

[4] T. Taleb *et al.*, "On Multi-Access Edge Computing: A Survey of the Emerging 5G Network Edge Cloud Architecture and Orchestration," *IEEE Commun. Surveys & Tutorials*, vol. 19, no. 3, 2017, pp. 1657–81.

[5] Y. Dai *et al.*, "Artificial Intelligence Empowered Edge Computing and Caching for Internet of Vehicles," *IEEE Wireless Commun.*, vol. 26, no. 3, June 2019, pp. 12–18.

[6] Y. Mao *et al.*, "A Survey on Mobile Edge Computing: The Communication Perspective," *IEEE Commun. Surveys & Tutorials*, vol. 19, no. 4, 2017, pp. 2322–58.

[7] H. Zhou *et al.*, "Data Offloading Techniques through Vehicular Ad Hoc Networks: A Survey," *IEEE Access*, vol. 6, no. 1, 2018, pp. 65,250–59.

[8] A. Mukherjee *et al.*, "A Power and Latency Aware Cloudlet Selection Strategy for Multi-Cloudlet Environment," *IEEE Trans. Cloud Comp.*, vol. 7, no. 1, 2019, pp. 141–54.

[9] J. Wan *et al.*, "Fog Computing for Energy-Aware Load Balancing and Scheduling in Smart Factory," *IEEE Trans. Ind. Informat.*, vol. 14, no. 10, Oct. 2018, pp. 4548–56.

[10] Y. Sun *et al.*, "Hierarchical Radio Resource Allocation for Network Slicing in Fog Radio Access Networks," *IEEE Trans. Vehic. Tech.*, vol. 68, no. 4, Apr. 2019, pp. 3866–81.

[11] H. Zhou *et al.*, "DRAIM: A Novel Delay-Constraint and Reverse Auction-Based Incentive Mechanism for WiFi Offloading," *IEEE JSAC*, vol. 38, no. 4, 2020, pp. 711–22.

[12] C. Liu *et al.*, "Dynamic Task Offloading and Resource Allocation for Ultra-Reliable Low-Latency Edge Computing," *IEEE Trans. Commun.*, vol. 67, no. 6, 2019, pp. 4132–50.

[13] N. C. Luong *et al.*, "Applications of Deep Reinforcement Learning in Communications and Networking: A Survey," *IEEE Commun. Surveys & Tutorials*, vol. 21, no. 4, 2019, pp. 3133–74.

[14] R. Khan *et al.*, "A Survey on Security and Privacy of 5G Technologies: Potential Solutions, Recent Advancements, and Future Directions," *IEEE Commun. Surveys & Tutorials*, vol. 22, no. 1, 2020, pp. 196–248.

[15] K. Zhang *et al.*, "Mobile Edge Computing and Networking for Green and Low-Latency Internet of Things," *IEEE Commun. Mag.*, vol. 56, no. 5, May 2018, pp. 39–45.

Biographies

KAI JIANG [S'21] (jiangkai0112@gmail.com) received his B.Sc. degree in measurement and control technology and instrument from Yangtze University, Jingzhou, China, in 2018. He is currently pursuing an M.S. degree in computer technology at China Three Gorges University, Yichang. His research interests include mobile edge computing and reinforcement learning.

HUAN ZHOU [M'14] (zhouhuan117@gmail.com) received his Ph.D. degree from the Department of Control Science and Engineering at Zhejiang University. He was a visiting scholar at Temple University from November 2012 to May 2013, and a CSC supported postdoctoral fellow at the University of British Columbia, Canada, from November 2016 to November 2017. Currently, he is a full professor in the College of Computer and Information Technology, China Three Gorges University. He was a Lead Guest Editor of *Pervasive and Mobile Computing*, Special Session Chair of the 3rd International Conference on Internet of Vehicles, and a TPC member of IEEE WCSP '13 and '14, CCNC '14 and '15, ICNC '14 and '15, ANT' '15 and '16, IEEE GLOBECOM '17 and '18, ICC '18and '19, and others. He has published more than 50 research papers in some international journals and conferences, including *IEEE JSAC*, *TPDS*, *TVT*, and so on. His research interests include mobile social networks, VANETs, opportunistic mobile networks, and mobile data offloading. He received the Best Paper Award of I-SPAN 2014 and I-SPAN 2018, and is currently serving as an Associate Editor for *IEEE Access* and the *EURASIP Journal on Wireless Communications and Networking*.

XIN CHEN [S'21] (sexychenxin@gmail.com) received his B.S. degree from China Three Gorges University in 2018. He is currently pursuing an M.S. degree in computer science and technology at China Three Gorges University. His research interests include mobile data offloading and VANETs.

HAIJUN ZHANG [M'13, SM'17] (haijunzhang@ieee.org) is currently a full professor with the University of Science and Technology Beijing, China. He was a postdoctoral research fellow with the Department of Electrical and Computer Engineering, University of British Columbia. He serves as an editor for *IEEE Transactions on Communications* and *IEEE Communications Letters*.