Chapter 1
Introduction

Abstract This chapter first introduces the fundamental concepts of mobile cloud computing. The differences between mobile edge computing and mobile cloud computing are then discussed in detail. The European Telecommunications Standards Institute's concept of mobile edge computing is introduced with respect to mobile edge computing's definition, architecture, advantages, and potential applications.

1.1 Mobile Cloud Computing (MCC)

MCC integrates cloud computing with mobile devices to enhance mobile device capabilities such as computing and storage. The user experience is improved by the execution of computation- and storage-sensitive applications through cloud computing and the delivery of related services. The architecture of MCC and mobile edge computing (MEC) is illustrated in Fig. 1.1. Mobile devices connect to the web services through nearby base stations. Web services act as the application programming interface (API) between mobile devices and the cloud and deliver cloud applications to the mobile devices. In current architecture, mobile devices can access cloud services through base stations in mobile network or Wi-Fi access points. MCC enables resource-limited mobile devices to run applications that are latency insensitive but computation intensive.

However, the inherent limitation of MCC is the long transmission distance between mobile devices and the cloud, which incurs long execution latencies and cannot satisfy the time constraints of latency-critical applications. There exist significant differences between MEC systems and MCC systems. MEC integrates cloud computing into mobile networks to provide computing and storage capabilities to end users at the edge. The main differences between MCC and MEC are summarized in Table 1.1.

- *Physical server*: In MCC systems, the physical servers are located in large-scale data centers. The data centers are large specific buildings. The buildings are equipped with adequate power supply and cooling equipment. The MCC servers are equipped with high computing and storage capabilities. In MEC systems, however, the servers are colocated with small-scale data centers, such as wireless

Y. Zhang, *Mobile Edge Computing*, Simula SpringerBriefs on Computing 9,
https://doi.org/10.1007/978-3-030-83944-4_1

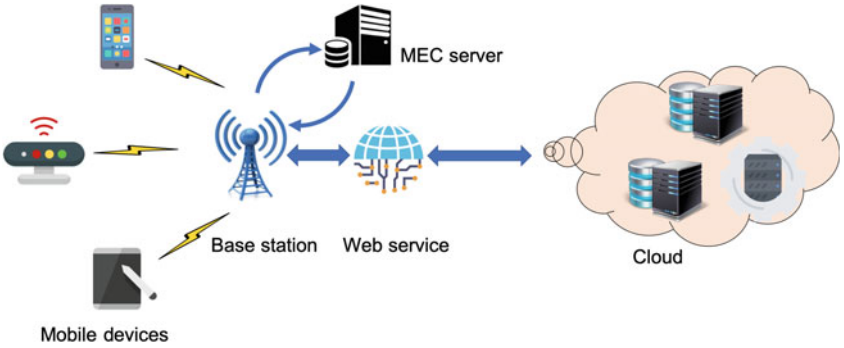

Fig. 1.1 The architecture of MCC and MEC

Table 1.1 Comparison of MCC and MEC

	MCC	MEC
Physical server	High computing and storage capabilities, located in large-scale data centers	Limited capabilities, colocated with base stations and gateways
Transmission distance	Usually far from users, from kilometers to thousands of kilometers	Quite close to users, from tens to hundreds of meters
System architecture	Sophisticated configuration, highly centralized	Simple configuration, densely distributed
Application characteristics	Delay tolerant, computation intensive, e.g., Facebook, Twitter	Latency sensitive, computation intensive, e.g., autonomous driving, online gaming

routers, base stations, and gateways. The MEC servers are equipped with a limited amount of computing and storage resources.

- *Transmission distances*: The distances between MCC servers and users can vary greatly, from kilometers to thousands of kilometers, even encompassing different countries, whereas the distances between MEC servers and end users are usually short, from tens to hundreds of meters.
- *System architectures*: The MCC systems are usually deployed by giant information technology (IT) companies such as Google and Amazon. The architectures of MCC systems are usually very sophisticated and highly centralized. The servers are controlled and maintained by specialized technical individuals. In MEC systems, the servers are usually deployed by telecommunications operators, enterprises, and communities. Theses servers are densely distributed in the network, with a simple configuration. MEC systems are hierarchically controlled in a centralized or distributed manner.
- *Application characteristics*: The applications in MCC systems can usually tolerate a certain degree of latency but require large amounts of computational resources. The computation data can thus be transmitted from end users to the MCC servers for computation. Typical examples of MCC applications are online social networking,

such as Facebook and Twitter. MEC applications are usually latency sensitive and computation intensive, such as image recognition in autonomous driving and online gaming. The computation of MEC applications requires execution at the network edge to mitigate long transmission delays between end users and the cloud.

Due to the different deployment architectures, the performance of MEC outweighs that of MCC in terms of latency, energy consumption, context-aware computing, and security and privacy. The benefits of MEC over MCC can be summarized as follows.

- *Latency performance*: The latency of mobile applications is composed of two parts: communication latency and computation latency. Compared with MCC, the propagation distances of MEC systems are much shorter. Generally, the distances of MEC systems are no longer than a kilometer. However, the distances between the cloud center and end users in MCC can be hundreds of kilometers and even span countries or continents. For example, the transmission distances for end users who want to use Google MCC applications in different parts of the world can vary from several kilometers to thousands of kilometers. Moreover, the transmission of MCC data usually requires passage through several networks, including a radio access network and the Internet, which can lead to additional delays in communication for MCC applications. In MEC systems, however, the computation data are transmitted through edge mobile networks or device to device, which are much simpler transmissions than in MCC systems. In terms of computation latency, although the cloud has large amounts of computational resources, they are shared with massive numbers of MCC users. In MEC systems, on the other hand, the computational capabilities of the servers are allocated to limited numbers of end users within their coverage. The gap in available computation capabilities for end users is thus mitigated. With short transmission distances and simple transmission schemes, MEC systems achieve better latency performance than MCC systems. In MEC systems, the latency is usually less than tens of milliseconds, whereas in MCC systems, the latency can be longer than hundreds of milliseconds.
- *Energy consumption*: Energy-consuming computation tasks can be offloaded from end devices to MEC servers, thus reducing the energy consumption of end devices. Specifically, in the Internet of Things (IoT), such offloading prolongs the battery life of IoT devices. In MCC systems, however, the long communication distances of computation data require end devices to maintain high transmission power, which will increase their energy consumption. By offloading computation-intensive tasks to nearby MEC servers, energy consumption is significantly reduced in MEC systems.
- *Context-aware computing*: Since MEC servers are much closer to the end devices, they can interact with end devices in real time by tracking their running states and making instantaneous decisions for them. The real-time interactions between MEC servers and end devices enable users to access context-aware services [1], such as real-time traffic updates and live video feeds, based on users' locations. For example, in autonomous driving, the MEC server leverages the information

from vehicles, such as locations and traffic conditions, to determine the vehicle's driving actions.

- *Security and privacy*: With increasing concerns about data security and privacy, the protection of user data has become a critical issue in mobile applications. With the development of end devices, the data collected can contain much of users' sensitive information. In MCC applications, the user data are transmitted to a remote cloud center over long distances. The data are then managed and processed by the cloud providers, such as Amazon and Microsoft. The risks of data leakage are extremely high during long-distance transmissions and remote management in the cloud. Cloud centers are more prone to become the targets of economically motivated attacks. On the other hand, MEC servers are deployed in distributed architectures of small scale. Many MEC servers can be privately operated and owned by the users in environments such as home cloudlets. Thus the risks of data leakage are considerably mitigated. MEC systems enhance user security and privacy for applications that might need to collect and process private user information.

Although MEC and MCC have different architectures and characteristics, they can sometimes also cooperate together, to enhance the computing capability and latency performance of the system. A series of works have explored combining MCC with MEC to improve application performance. For example, in the application scenario of online gaming, MEC provides users with cached data and local computation, while MCC provides users with new data and intensive computation. Thus the user's experience of image quality and delay performance can be considerably improved.

1.2 Overview of MEC

MEC provides a distributed computing environment by placing compute and storage resources closer to the consumer or enterprise end user. The term *MEC* was first introduced in 2013, when Nokia Siemens Networks and IBM developed a platform called Application Service Platform for Networks to allow mobile operators to deploy, run, and integrate applications at the edge of the network [2]. In 2014, the MEC technical white paper was developed by the European Telecommunications Standards Institute (ETSI) [3], and a new Industry Specification Group was established in ETSI to produce specifications. The Industry Specification Group has delivered several specifications on service scenarios, requirements, architecture, and APIs that will allow for the efficient and seamless integration of applications from vendors, service providers, and third parties across multi-vendor MEC platforms. In 2016, ETSI dropped the word *mobile* from MEC, renaming the technology multi-access edge computing (with the same acronym, *MEC*), to extend its scope to heterogeneous access technologies (e.g., LTE, 5G, Wi-Fi, and fixed access technologies).

MEC is a new paradigm that provides an IT service environment and cloud-computing capabilities at the edge of the network, within the radio access network, and in close proximity to mobile subscribers. The main purpose of MEC is to

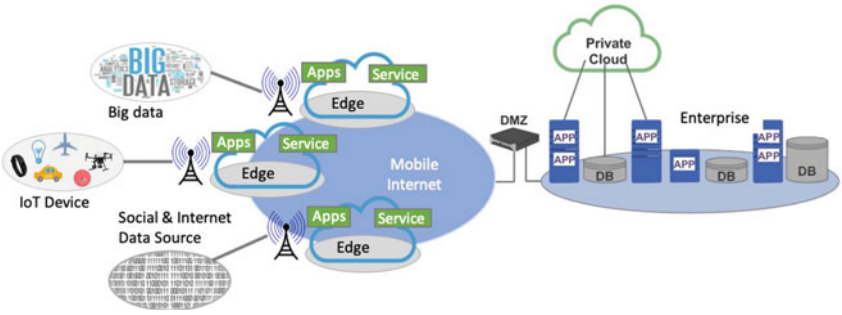

Fig. 1.2 The MEC framework

reduce backhaul network congestion, support low-latency applications, and offer an improved user experience. The general framework of MEC is shown in Fig. 1.2. Different types of big data applications, IoT devices, and social and Internet services are connected to distributed mobile edge networks. The mobile edge networks are connected to the private cloud network via a demilitarized zone for security. The private cloud network is equipped with sufficient databases and applications to provide centralized processing, storage, and computing service. Since cloud services and applications are far from mobile users, MEC deploys the distributed edge services and applications at wireless network infrastructures (i.e., base stations, Wi-Fi access points, or femto access points) to form distributed mobile edge networks. Users can easily access nearby the wireless network infrastructure to enjoy real-time and high-quality service applications. Additionally, MEC not only benefits end users, but also improves resource utility and network efficiency with network optimization, such as computation and caching resource allocation.

According to the ETSI white paper, MEC can be characterized by features such as on premises, proximity, low latency, location awareness, and network context information. These features can be briefly explained as follows.

- *On premises*: MEC platforms can run isolated from the rest of the network while maintaining access to local resources. This is very important for machine-to-machine scenarios, such as security or safety systems that need high levels of resilience.
- *Proximity*: MEC servers are usually deployed close to mobile users. MEC is thus particularly useful in capturing key information for analytics and big data. It is also beneficial for compute-hungry devices, such as augmented reality (AR) and video analytics.
- *Lower latency*: Since MEC services run close to end devices, latency can be considerably reduced, which can be utilized to react faster, improve user experience, or minimize congestion in other parts of the network.
- *Location awareness*: Due to proximity, MEC can leverage signaling information received from edge devices to determine the location of each connected device.

This feature leads to an entire family of business-oriented use cases, including location-based services and analytics.

- *Network context information*: Applications providing network information and real-time network data services can benefit businesses and events by implementing MEC for their business model. Based on real-time radio network conditions and local contextual information, these applications can estimate the radio cell and network bandwidth congestion. This can help in the future to make smart decisions to improve customer service delivery.

MEC not only enhances the performance of existing applications, but also provides tremendous potential for developing a wide range of new and innovative applications. In the following, we introduce several typical use cases in MEC.

- *Internet of Vehicles (IoV)*: The IoV aims to enhance safety, reduce traffic congestion, sense vehicles' behaviors, as well as provide opportunities for numerous vehicular services, such as smart navigation, traffic warnings, and real-time driving route planning. The communication model in IoV can either be vehicle to vehicle (V2V), vehicle to roadside infrastructure (V2R), or vehicle to Internet (V2I). However, resource-constrained vehicles can be strained by computation-intensive applications, resulting in bottlenecks and making it challenging for the vehicles to ensure the required quality of service level. MEC can alleviate the heavy computation requirement of vehicles by providing computation capabilities at the edge of the radio access network [7]. Due to the proximity of MEC servers to vehicles, the offloaded tasks can be accomplished with low latency and high efficiency.
- *Smart grids*: A smart grid offers transparent energy distribution where both consumers and utilities can monitor and control their pricing, production, and consumption in almost real time. A smart grid infrastructure is an electrical grid that consists of several components, such as smart appliances, renewable energy resources, and energy efficiency resources [8]. Smart meters are distributed throughout the network to receive and transmit measurements of energy consumption. All the data collected by the smart meter are supervised by supervisory control and data acquisition systems to maintain and stabilize the power grid. The analysis of the data from various smart meters in the smart grid environment is challenging, since it varies with respect to parameters such as size, volume, velocity, and variety. The usage of MEC can improve performance in throughput, response time, and transmission delay. Distributed smart meters and microgrids, integrated with MEC, have the ability to conduct nearby data management and analysis. For example, a three-tier fog-based smart grid architecture [9] is proposed to extend the capabilities of cloud-based smart grids in terms of latency, privacy, and locality.
- *Unmanned aerial vehicles (UAVs)*: With recent advancements in technology and reductions in manufacturing cost, UAVs have received growing interest in various applications, such as disaster rescue, cargo delivery, filming, as well as monitoring. To maintain UAVs' safe operation with real-time commands and enable the above computation-intensive applications, it is important to enhance the communication and computational capabilities of UAVs. With the help of MEC, edge computing resources can be deployed on UAVs to support computation-intensive and latency-

critical applications. On the other hand, the rapid growth of network traffic has made it difficult for static base stations to support the data demands of billions of devices. UAVs can act as flying base stations to support ubiquitous wireless communications and unprecedented IoT services, due to their high flexibility, easy deployability and operability. In UAV-aided MEC networks, UAVs can act as mobile computation servers or computation offloading routers to provide users better wireless connections and greater flexibility in the implementation of MEC.

- *AR/virtual reality (VR) services*: AR and VR allow users to interact more naturally with virtual worlds based on the data generated from sensory inputs, such as sound, video, graphics, or a global positional system. AR/VR applications need real-time information on users' status, such as their location and direction, and also require low latency as well as intensive data processing for a better user experience. MEC is an ideal solution for AR and VR applications, since MEC servers can exploit local context information and provide nearby data processing. Deploying a VR controller on a MEC server and utilizing wireless links to transmit VR images and audio can increase tracking accuracy, obtaining round-trip latencies of one millisecond and high reliability [4]. Caching parts of VR videos on MEC servers in advance and then performing computations on VR devices can save large amounts of communication bandwidth and fulfill low latency requirements [5]. Offloading computation-intensive tasks to edge servers can increase the computational capacity of AR devices and save their battery life [6].s
- *Video stream analytics*: Video streaming has a wide range of applications, such as vehicular license plate recognition, face recognition, and security surveillance. Video streaming has been observed to comprise more than half of the entire mobile data traffic in current networks, and the percentage is still increasing. The main video streaming operations are object detection and classification. These tasks usually have high computation complexity. If these tasks are processed in the central cloud, the video stream will be transmitted to the cloud network, which will consume a great amount of network bandwidth. MEC can offer ultra-low latency, which is required for live video streaming, by performing the video analysis in a place close to edge devices. A caching-based millimeter-wave framework is proposed to pre-cache content at the base station for hand-off users [11]. The proposed solution can provide consistent high-quality video streaming for high-mobility users with low latency.

1.3 Book Organization

This book aims to provide a comprehensive view of MEC. As a key enabling technology for achieving intelligence in wireless communications, MEC has been widely studied in a series of areas, including edge computing, edge caching, and the IoV. However, with the development of new technologies, such as blockchain, artificial intelligence, and beyond 5G/6G communications, new opportunities have opened up for the fulfillment of MEC and its applications. Motivated by these new changes, this

work provides comprehensive discussions on MEC in the new era. We first present the fundamental principles of MCC and MEC technologies. Next, we present applications of MEC in typical edge computing and edge caching scenarios. Furthermore, we discuss research opportunities in MEC in emerging scenarios such as the IoV, 6G, and UAVs. Finally, we provide potential directions of MEC for the future.

This book is organized as follows. Chapter 2 presents the models and policies of edge computing. Chapter 3 describes the architecture and performance metrics of mobile edge caching. A case study of deep reinforcement learning–empowered edge caching is further conducted in Chap. 4. Applications of MEC in the IoV for task and computation offloading are presented in Chap. 5. Chapter 6 describes details on the application of MEC to UAVs. Finally, Chap. 7 provides a comprehensive discussion of the future of MEC.