1

Introduction

1.1 Focus of the Book

Besides the latest research on computing design and standardization efforts made by the European Telecommunications Standards Institute (ETSI) Industry Specification Group (ISG), the driving forces behind edge computing and its technology foundations deserve a thorough illustration and analysis. Along with the advancement of Internet-of-Things (IoT) and mobile communications, the rapid development on "Edge" is calling for a handbook that covers the service/application perspectives and how to integrate "Edge" in the future computing infrastructures.

Given the "Edge" domain is in a fast growing phase, in this book, we deliberately use the term **MEC**^{2} in a more generic manner to enclose the essential technologies for mobile edge computing and communications.

One of the key reasons for a seemingly old idea to take off after many decades is that other enabling technologies make it possible. This holds especially for *MEC*^{2}. For the rapid advancement of *MEC*^{2}, this book is not intended to be encyclopedic and likely to be an evolving "piece" for the foreseeable future. In this respect, readers' suggestions, feedback are more than welcome. By serving as a reference material and introductory guide on mobile edge computing and communications, this book shall appeal to a wide intended audience including academic researchers, service developers, computing professionals, curious university students, network operators and industrial experts.

In this book, we take the system and incremental deployment perspective, which is not to be mixed with incremental research. The *MEC*^{2} research can be disruptive or ground-breaking (i.e., non-incremental) but still lend itself to incremental deployment and integration to existing mobile and cloud infrastructures. The research can hence go beyond the constraints of existing deployment context by envisioning what could be possible. Ultimately, it comes down to the deployment of *MEC*^{2} that leads to long-term impact. Overall, this book shall offer practical deployment insights for actors and stakeholders in the future mobile and cloud infrastructures.

Mobile Edge Computing and Communications: Driving Forces, Technology Foundation, and Application Areas, First Edition. Aaron Yi Ding, Chamitha De Alwis, and Madhusanka Liyanage.

1.2 The Vision of Edge

The vision for edge computing is brought out as early as 2005 [1] where the importance of network edge is highlighted as a new source of creative energy for system and applications, especially given the growing demand in mobile access networks. The initial concept is further formulated in recent years, along with a visible trend for data analytics moving toward this "edge." Gartner predicted that by 2023, over 50% of the primary responsibility of data analytics stakeholders will comprise data that is created, managed and analyzed in edge environments. The advantages may include greater data management flexibility, speed, governance, and resilience. In addition, the edge capabilities can support use cases ranging from real-time event analytics to autonomous driving services [2].

Albeit its great potential, the fundamental concept of edge computing, referred as MEC^2 in this book, is not new. By looking at the history of computing, it is not the first time for the network edge to receive tremendous attention. For instance, we have seen numerous computing design before the hype of MEC^2, including the peer-to-peer (P2P) networks, content delivery networks (CDN), and mobile cloud computing (MCC). It is hence appropriate to remind our readers that one of the key reasons for a seemingly "old" idea, such as edge computing, to take off after many years of undergoing is that other enabling technologies are making it possible. Those enablers and technology foundation are exactly what we intend to illustrate throughout this book.

In fact, the vision is largely built upon its promises to the rising demands for high bandwidth, low latency, better privacy and reliability. Those promises have been argued recently by experts as to whether it is merely an academic enthusiastic pursuit [3]. At the same time, as the Internet is transformed into a more ethical system, MEC^2 is gradually showing its potential in enhancing data privacy and sovereignty [4].

Regardless of the controversial opinions [3, 5, 6], MEC^2 is clearly far from reaching its "Hay Day." Many uncertainties and challenges are intertwined with the opportunities along its advancement. From a pragmatic perspective, this book conveys our vision to facilitate future development of MEC^2 towards a coherent edge-to-cloud continuum. This continuum path needs more collaborative efforts with experts from multiple disciplines even outside the technical sphere (e.g., law, ethics and public policies), for achieving the full potential of MEC^2.

Why Do We Need Edge?

Regarding a frequently asked question "Why do we need edge?," the general motives and potential benefits include:

- **Latency and Bandwidth:** MEC^2 brings the latency benefit by placing computation units closer to the data sources and end users. Given the bandwidth bottleneck to access a distant cloud via the Internet, MEC^2 can conserve bandwidth by processing data locally, hence avoiding transferring excessive amount of data to the Internet. This helps alleviate network congestion. The latency and bandwidth benefits are crucial for real-time services that also demand higher bandwidth, such as virtual reality (VR) gaming and autonomous driving.
- **Proliferation and Personalisation:** MEC^2 is envisioned to capture high quality data in a distributed manner from the massive deployment of IoT, augmented reality (AR)/VR

devices, and smart vehicles. *MEC*^{2} can achieve higher quality by filtering out the data "noise," labeling the data with more context and with better sampling. The high-quality data with locality context is demanded by analytic services for personalisation (e.g., for end users/clients).

- **Privacy and Sovereignty:** The edge-native computing and analytics (often referred as Edge AI) can keep the data ownership and control closer to the end users. This can when computation is managed and task distribution controlled from the user's own devices, and suitable security and privacy protection methods are in use. By utilizing local context, *MEC*^{2} can strike a balance between privacy and usability, while allowing ethical data management.
- **Energy Saving and Sustainability:** *MEC*^{2} avoids transmitting redundant data traffic to consume network resources. The energy required for transmitting data is proportional to the distance that data travels from the source (end user devices) to the Internet or centralized cloud sites. By cutting down the data volume to be transmitted and processed by the network units along the path, *MEC*^{2} can promote the energy saving of the Internet. In addition, the data flowing through the Internet is becoming one the primary drivers for CO_2 emissions. *MEC*^{2} can help reduce the carbon footprint.
- **Mobility and Collaboration:** *MEC*^{2} can support end user devices continue functioning even when they are disconnected from the cloud. This is particularly valuable for mobile scenarios where network connectivity is intermittent or unreliable. In addition, *MEC*^{2} is complementing the corporate data centers that facilitate computing, storage, networking, and data analytic functions at locations such as collaboratively training large-scale machine learning models.

What and Where is the Edge?

To approach the question "**What is the Edge?**," it can often trigger heated discussions. Can it be a lightweight computing server, a layer of networking devices, or a set of IoT oriented services? For some of you who are studying *MEC*^{2} for the first time, this can be rather confusing.

One major reason for such blurring interpretation is the fact that *MEC*^{2} is in rapid development. The concept of "edge" is hence enriched and evolved along the way, sometimes mixed with different flavours. For example, a general expectation for edge is to extend the cloud computing capability to the access networks. For this prospect, *MEC*^{2} can take the form of lightweight computing servers that can offload the tasks from data centers (cloud) for performance considerations. Meanwhile, for mobile network operators that are enhancing their access infrastructure, *MEC*^{2} can also be regarded as a layer of network devices (e.g., routers) in the proximity to mobile devices and data sources. In addition, there is an increasing pressure to offer low-latency and high-bandwidth applications such as augmented reality (AR), virtual reality (VR), Internet-of-Things (IoT), and autonomous driving. From the service offering angle, *MEC*^{2} can be regarded as a "meta" service that enables numerous advanced applications demanded by the growing amount of connected things.

Closely related to the "What" question, "**Where is the Edge?**" also depends on which stakeholder we are asking. For an Internet service provider (ISP) that owns and operates the network access infrastructure, *MEC*^{2} can be a layer that consists of last hop gateways to connect end users to ISP's infrastructure. For instance, mobile operators can utilize *MEC*^{2}

to construct a computing layer on their base stations, so the *MEC*^{2} enabled base station is regarded as where the edge is "physically" located. For cloud service providers such as Amazon and Google, *MEC*^{2} is the extended computing and storage infrastructure closer to their customers for supporting real-time services, analytics, and content delivery. For instance, Content Delivery Networks (CDNs) can be regarded as where the edge is placed, for reducing the distance and latency of content delivery. In the future, *MEC*^{2} may come in different form factors for different scenarios. The devices/entities with computational capability to perform computing near the data source could be considered as part of edge infrastructure This is in line with the motive of edge computing.

1.3 Stakeholders and Related Paradigms

Beyond the "What" and "Where" questions, it is important to understand the major stakeholders and related paradigms of *MEC*^{2}. The major stakeholders and actors involved in *MEC*^{2} include:

- **Mobile Network Operators (MNOs):** MNOs are currently looking to adopt edge computing to offer responsive mobile services to their customers. For telecom's next generation mobile infrastructure, MNOs can harness the advantage of edge to create an integrated computing-communication hub in the "last-mile" network for efficient data processing, analytics, and communications.
- **Cloud Service Providers:** To meet the demands from a wide-range of latency-sensitive and data-intensive applications that will emerge in the near future, cloud service providers are adopting edge designs to enhance their service performance. This is in line with the new industry trend of computing continuum, which could lead to the convergence of cloud and edge.
- **Academic Researchers and Industrial Experts:** *MEC*^{2} has created active research and practical solutions for academic researchers and industrial experts, respectively, to tackle challenges in computing sustainability and data sovereignty.
- **ICT Service/Application Developers:** As computing requirements and user demands can outpace the traditional tools, ICT service developers are constantly seeking new technologies such as *MEC*^{2} that can facilitate them to create innovative ICT applications that cater to future demands from users, government regulations, and international standards.

Since *MEC*^{2} is in its making phase, there are related computing designs that have been explored and inspired the trend of edge. The related computing concepts include P2P, CDN, MCC, Cloudlets, Fog and Mist:

- **P2P**: The concept of Peer-to-Peer (P2P) is driven by decentralization. It is intended to address the limitations of classic client–server design in terms of performance bottleneck and single point of failure on the server side. As a decentralized networking design, P2P can enable direct communication between distributed devices for better scalability, reliability and resource utilization. P2P has provided inspiration to *MEC*^{2} from decentralization perspective.

- **CDN**: To optimize the latency in content-oriented ICT/web services, Content Delivery Network (CDN) forms a distributed proxy layer that can cache content and serve such content via CDN servers geographically closest to the end users. The latency benefit of CDN has inspired *MEC*^{2} to achieve low latency, load balancing and service availability in a distributed environment.
- **MCC**: The design of Mobile Cloud Computing (MCC) integrates cloud computing with mobile computing to support efficient computational offloading. By harnessing the computing and storage capacity in the remote cloud, MCC can enable devices with limited processing power and storage to run resource-intensive applications. MCC has inspired *MEC*^{2} to enhance end user experience in terms of resource offloading and battery saving on the end devices.
- **Cloudlets, Fog and Mist**: The idea of Cloudlets [7] is to deploy small-scale data processing units close to the edge of access network. Cloudlets aims to tackle latency and bandwidth challenges for applications that often need to exchange data with remote cloud servers. In line with the goal of improving latency and bandwidth, Fog computing [8] aims to transform the network edge to a distributed computing infrastructure for the rapid growth of IoT. Similar to Fog computing, the term Mist is geared towards embedded computing devices deployed in the access networks. In general, Cloudlets, Fog and Mist intend to extend the Cloud computing paradigm by allowing data to be processed and analyzed across various levels of the network hierarchy. The concepts of Cloudlets, Fog and Mist formed the groundwork for computing continuum that inspired *MEC*^{2} to complement both centralized cloud and distributed end devices by offering an intermediate layer of edge resources to balance the benefits of local processing with the advantages of cloud services.

In the following chapters of Part 1 we will illustrate the applications of *MEC*^{2} to help readers comprehend the practical usage of edge technologies.