1 Mobile Edge Computing

Ben Liang

1.1 Introduction

The ongoing development of the fifth-generation (5G) wireless technologies takes place in a unique landscape of recent advancement in information processing, marked by the emerging prevalence of cloud-based computing and smart mobile devices. These two technologies complement each other by design, with cloud servers providing the engine for computing and smart mobile devices naturally serving as human interface and untethered sensory inputs. Together, they are transforming a wide array of important applications such as telecommunication, industrial production, education, e-commerce, mobile healthcare, and environmental monitoring. We are entering a world where computation is ubiquitously accessible on local devices, global servers, and processors everywhere in between. Future wireless networks will provide communication infrastructure support to this ubiquitous computing paradigm, but at the same time they can also utilize the new-found computing power to drastically improve communication efficiency, expand service variety, shorten service delay, and reduce operation expenses.

Previous generations of wireless networks are passive systems. Residing near the edge of the Internet, they serve only as communication access pathways for mobile devices to reach the Internet core and the public switched telephone network (PSTN). Improvements to these wireless networks have focused on the communication hardware and software, such as advanced electronics and signal processing in the transmitters and receivers. Even for 5G, substantial research effort has been devoted to densification techniques, such as small cells, device-to-device (D2D), and massive multi-input multi-output (MIMO). Successes of this communication-only wireless evolution reflect the classical view of an information age centered at information consumption through the Internet.

Yet, in many emerging applications, communication and computation are no longer separated, but interactive and unified. For example, in an augmented-reality application, which might be displayed on smart eye-glasses, the user's mobile device continuously records its current view, computes its own location, and streams the combined information to the cloud server, while the cloud server performs pattern recognition and information retrieval and sends back to the mo-

Ben Liang is with the University of Toronto, Canada.

bile device contextual augmentation labels, to be seamlessly displayed overlaying the actual scenery. As seen from this example, there is a high level of interactivity between the communicating and computing functions, and a low tolerance for the total delay in information transmission and information processing. A fitting analogue of this may be found in a biological system, where the computation by neurons and the communication among them are inseparable. Indeed, we are moving toward an info-computation age characterized by tight coupling between communication and computation. With an explosion of available data and the consequent need for enormous data-processing capabilities in the emerging big data movement and the Internet-of-Things (IoT) environment, both communication and computation are paramount to future wireless applications and services.

Therefore, 5G and future wireless systems are expected to transition away from the model of passive information conduits, into active providers and creators of info-computation resources in an integrated communication-computation paradigm. To this end, one major characteristic of future-generation wireless systems will be the seamless integration between hardware and software. For example, the 5G Infrastructure Public Private Partnership (5G-PPP) has acknowledged the central role of software and recognized several key software-driven components for 5G standardization, including software-defined networking (SDN), network functions virtualization (NFV), and mobile edge computing (MEC) [1]. Through these technologies, it is envisioned that network functions will be provided over multiple points of presence, especially near the edge of the Internet, in order to achieve 5G targets on performance, scalability, and agility.

In this chapter, we focus on MEC in 5G and beyond systems. Here, the term *mobile edge* refers to the radio access network (RAN) side of the Internet. It signifies the position of mobile devices with respect to the core computing servers residing in cloud centers. This is in contrast to the network-centric view where all equipment attached to the Internet, including mobile devices and cloud computing servers, are considered as the edge. We will first describe the MEC functionalities and architecture, then present some example use cases, and finally discuss relevant research challenges.

1.2 Mobile edge computing

The concept of MEC is built on recent advances in mobile cloud computing (MCC). In MCC, cloud computing servers produce shared pools of always-on computing resources (e.g., processors, software, storage), while mobile devices consume these external resources through RANs and the Internet. Cloud computing resources can be rapidly provisioned without the usually prohibitive capital and management costs incurred by users of traditional, self-managed computing servers. MCC aims to make computation a ubiquitously accessible utility, similar to but more agile than physical utilities such as electricity and water.

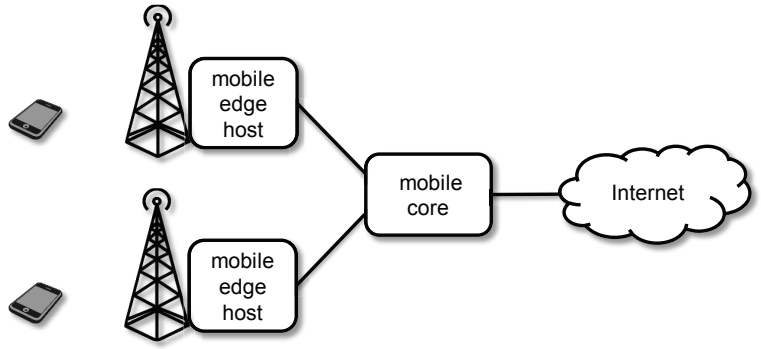

Figure 1.1 Illustration of the MEC system.

The computing resources in MCC may be centralized or distributed. In the conventional centralized form of MCC, computing resources are provided to mobile users from large remote cloud centers such as the Amazon Elastic Compute Cloud and Microsoft Azure. These cloud centers provide virtually unlimited computation capacity to augment the processors in mobile devices. However, the communication between mobile users and remote cloud centers is often over a long distance, adding to the latency in cloud computation. Therefore, alternate forms of MCC have been proposed, where computing resources may be accessed in a distributed manner, from smaller local servers such as computing-augmented base stations and WiFi access points, or from nearby mobile devices with excess computing capacity. The latter two scenarios of MCC are sometimes named *micro cloud centers* [22], *cloudlets* [35], or *fog computing* [4]. They supplement centralized cloud centers by offering lower access delay and more local awareness.

MEC, as defined by the European Telecommunications Standards Institute (ETSI) [17], refers to a distributed MCC system where computing resources are installed within the RANs, close to the mobile-device end of the Internet. An illustration of the MEC system is given in Figure 1.1, where the *mobile edge hosts* are computing equipment installed at or near base stations. Unlike centralized cloud servers or peer-to-peer mobile devices, MEC is managed locally by the network operator. The generic computing resources within the mobile edge hosts are virtualized and are exposed via application program interfaces (APIs), so that they are accessible by both user and operator applications.

The mobile edge hosts provide local virtual machines (VMs) to serve the computation needs of mobile devices, often with much lower latency than remote cloud centers. They also serve some functions of the traditional mobile core, such as user content caching and traffic monitoring, as well as new functions such as local information aggregation and user location services. Thus, the MEC system may be viewed as a natural outcome of the evolution of mobile base stations from

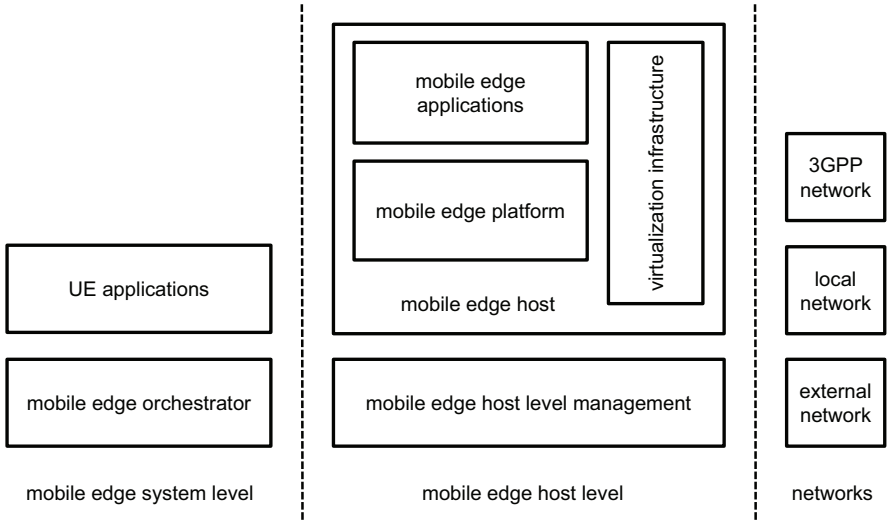

Figure 1.2 Simplified ETSI MEC reference architecture.

passively serving purely communication functions to becoming an integral part of the new communication-computation paradigm. It is both a midway stop for mobile access to information and computation in the Internet, and a cross-layer bridge that promotes more efficient integration between mobile devices and the mobile core, facilitating the operations of both.

1.3 Reference architecture

A simplified illustration of the MEC reference architecture proposed by ETSI is given in Figure 1.2. The MEC system resides between the user equipment (UE) and mobile core networks. It consists of management and functional blocks at the mobile edge host level and the mobile edge system level.

At the mobile edge host level, *mobile edge applications* run VMs supported by the *virtualization infrastructure* within the mobile edge host. They provide services such as computational job execution, radio network information, bandwidth management, and UE location information. The *mobile edge platform* hosts mobile edge services. It interacts with mobile edge applications, so that they can advertise, discover, offer, and consume mobile edge services. The *mobile edge platform manager* provides element management functions to the mobile edge platform and administers application essentials such as life cycle, service requirements, operational rules, domain name system (DNS) configuration, and security.

At the mobile edge system level, the *mobile edge orchestrator* serves the cen-

tral role of coordinating among the UEs, the mobile edge hosts, and the network operator. It records accounting and topological information about the deployed mobile edge hosts, available resources, and available mobile edge services. It interfaces with the virtualization infrastructure and maintains authentication and validation of application packages before their on-boarding. It is also in charge of triggering application instantiation, termination, and relocation, based on its choice of appropriate mobile edge hosts to satisfy the application's requirements and constraints. The MEC-capable *UE applications* run within the UE and interact with the mobile edge system to request the on-boarding, instantiation, termination, and relocation of mobile edge applications. An important service provided by the mobile edge orchestrator to an MEC-capable UE application is the timely migration of mobile edge applications between mobile edge hosts, to support UE handoff between different network attachment points.

1.4 Benefits and application scenarios

Because of its proximity to the mobile devices, the MEC system offers low-latency and high-bandwidth mobile access to both information and computation resources. At the same time, by locally absorbing some communication traffic and computation functions of the mobile core, it reduces the resource demand on the mobile backhaul. Furthermore, because of its unique location within the RAN, it is also capable of monitoring and reporting the local network condition to the mobile core, which promotes awareness of the communication environment at the edge for improved operational efficiency.

Developing the MEC system will benefit a wide range of concerned parties in various application scenarios. MEC directly serves the end users by placing information and computation resources in close proximity to them. To network operators, MEC also serves important roles in improving wireless system performance and reducing the cost of operation. To hardware and software developers, the availability of mobile edge platforms and virtualization infrastructure promotes the creation of new applications and consumer products. The following are some example use cases suggested by ETSI for the users and operators of MEC [16].

1.4.1 User-oriented use cases

Application computation offloading. Computationally intensive jobs can be processed by the mobile edge host instead of mobile devices. Examples of such jobs include high-speed browser, 3D rendering, video analysis, sensor data processing, and language translation. They tend to require many CPU cycles and are major sources of drainage on the mobile on-board battery. With the option to cheaply offload heavy computation to nearby mobile edge hosts for accelerated processing, the mobile device's computational capability and energy consumption will

no longer be the bottleneck in delivering rich applications. This is particularly appealing in the IoT environment, where the mobile devices are likely to consist mostly of small sensors and other equipment with minuscule processors and limited energy supply.

Gaming, virtual reality, and augmented reality. These applications all require low latency. On one hand, the user could implement the rendering pipeline on the mobile device itself, but the heavy computation requirements of jobs such as physical simulation and artificial intelligence might overwhelm the limited processing capability of the mobile device. On the other hand, offloading these jobs to a remote cloud server might incur too much latency. Instead, with MEC, part of the computational load can be offloaded to some mobile edge application running on a mobile edge host. Thus, MEC provides an appropriate balance between computation power and proximity.

Edge video orchestration. In a local event with a densely populated audience, such as sports game or concert, a huge number of mobile devices simultaneously access videos of the same event. The videos often are rich in content and include multiple streams, since they may include both real-time and on-demand versions from multiple camera angles. However, they are all locally generated. Instead of sending these videos back and forth through the mobile core and the Internet to and from a video content server, they can be processed and delivered out of the mobile edge hosts. In addition to relieving the traffic demand on the backhaul, edge video orchestration also provides a more convenient framework to control the service quality, for example, through better matching between the hardware capability of mobile devices and scalable video coding.

Vehicle-to-infrastructure communication. Vehicle-to-infrastructure communication services, such those between cars and roadside units, are important to the safety and efficiency of the transportation system. MEC can enable the application of 5G wireless to serve such communication functionalities, which more readily provides global coverage than dedicated short-range communications (DSRC). Applications that provide these functionalities can be loaded in the mobile edge hosts, which are installed inside the RANs along the roadway. Due to their proximity to vehicles and roadside equipment, hazard warning and other latency-sensitive messages can reach targeted vehicles within their extremely stringent time scale requirements.

1.4.2 Operator-oriented use cases

Local content caching at the mobile edge. The prevalence of widely shared large files in social media, such as high-definition videos, stresses the backhaul network capacity. However, these viral contents tend to be consumed by many users within the same geographical area and the same time period. Therefore, they can be cached locally at the mobile edge hosts to drastically reduce the traffic demands on the backhaul. Furthermore, with a mobile edge application that proactively moves the cached content between mobile edge hosts in anticipation

of the user movement, the service quality can be improved. Similar benefits extend also to broadcast videos, such as mobile TV, where the same video stream is viewed by many mobile users near a mobile edge host.

Data aggregation and analytics. Many services provided by the operator or third-party vendors, such as security monitoring and massive sensor information processing, depend on the large amount of data collected from the sensors and mobile devices. Often it is inefficient to send all collected raw data from each sensor and mobile device to backend servers. In particular, the massive influx of IoT devices may overwhelm the core network. Instead, a distributed mobile edge application running on multiple mobile edge hosts can process the data first to extract the meta data of interest, before forwarding the meta data to backend servers.

Mobile media streaming with bandwidth feedback. Hypertext transfer protocol (HTTP) over transmission control protocol (TCP) is ubiquitous in media streaming. Yet, TCP is highly inefficient when applied over conventional wireless networks. TCP is designed to view packet loss as a signal for network congestion, but in wireless networks, packet loss is commonplace and often is due to drastic fluctuation in the radio channel condition. Thus, streaming performance can suffer because of the miscalculation of the available bandwidth by TCP's congestion control algorithm. With MEC, a radio analytics application running at the mobile edge host can monitor the available wireless bandwidth and forward it to an MEC-capable backend video server. Such information will then be used by TCP at the video server to properly adjust its sending rate.

Mobile backhaul optimization. Conventional wireless networks lack coordination between the RANs and the backhaul, which prevents effective utilization of the backhaul capacity shared by multiple RANs. With MEC, the traffic and performance at the RAN level can be monitored and processed by mobile edge applications. Such localized real-time information, along with other pertinent information such as RAN scheduling and user application profiles, can then be made available to the backhaul network. Thus, the backhaul can be optimized through techniques such as application traffic shaping, traffic routing, and capacity provisioning.

Location-based services. MEC provides more effective location-based services in two ways. First, it allows user location tracking using advanced analytical techniques beyond the received signal strength. Second, mobile edge applications can provide more appropriate location-based service recommendations. A mobile edge application can take into account its knowledge of the context of the user location, such as shopping mall or museum, as well as the user behavior pattern, to give recommendations. It may also utilize advanced machine-learning techniques and interface with big data analysis at backend servers to further improve the accuracy and usefulness of its recommendations.

1.5 Research challenges

MEC is a complex system that creates a new framework to support tight integration between wireless networking and cloud computing. On the networking side, it addresses all layers of the network protocol stack. On the cloud computing side, it involves all three cloud service modes: Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS). Because of the communication and computation synergistic nature of MEC, there are wide ranging challenges in its research and development.

1.5.1 Computation offloading

Offloading computation-intensive and resource-hungry applications to resource-rich servers can augment the capabilities of mobile devices. The reduction in energy consumption by mobile devices through computation offloading has been demonstrated in experimental tests and analytical studies [26]. However, energy is only one of many important factors in mobile resource optimization [13][11]. In MEC, there are multiple mobile devices sharing the same mobile edge host, so its limited capacity in communication and computation is a major concern. The usage cost of the backend cloud servers and the MEC system should also be considered.

For each job, a decision needs to be made on whether to execute it locally on the mobile device, send it to the mobile edge host, or further forward it to remote cloud servers. Even assuming the processing times and costs of each job at different locations are known in advanced, this leads to an integer programming problem that often is extremely difficult to solve [15]. Moreover, the processing times and costs are likely unknown before a job has been executed, due to the lack of exact information about the job's required number of operations, as well as the randomness in each processor's available computing cycles [38][6]. New analytical tools are needed to handle such uncertainty toward designing MEC systems with both average and worst-case performance guarantees.

Yet another challenge is in the placement constraints imposed on each job. With the co-existence of multiple mobile operating systems and vast heterogeneity in hardware and software requirements, not all mobile edge hosts are usable to execute every job. The patchwork of such placement constraints and user utility preferences will substantially complicate the optimization of computation offloading [37].

1.5.2 Communication access to computation resources

The benefit of computation offloading cannot be enjoyed without efficient access to the mobile edge hosts and the remote cloud servers. In conventional grid computing and cloud computing over wired links, a common assumption is that the communication pathway between the user and the cloud center is unimpeded

[38], so that much research effort has been focused instead on ensuring wired connectivity between grid computers or cloud nodes. However, a main vulnerability of mobile cloud computing is the instability of the wireless access links between mobile devices and cloud computing resources [27]. The access links may be interrupted without notice, leading to random service outages. Therefore, the joint allocation of communication resources and computation resources is paramount to the successful deployment of MEC.

There are three important communication concerns in MEC: wireless access while offloading to the mobile edge host; backhaul access while offloading to a remote cloud server; and the communication among mobile devices, mobile edge hosts, and remote cloud servers when they collaboratively execute multiple jobs. Both the wireless and backhaul access links have limited capacity and must be properly shared among multiple mobile devices, in similar ways as the computing resources of the mobile edge host are shared [52][10]. Therefore, a joint communication-computation resource sharing framework is required in MEC. Furthermore, such resource sharing decisions are strongly influenced by the requirement of data exchange between jobs that are executed at different locations. This data requirement induces a dependency relation among jobs, where the execution of one job must wait for the completion of some other jobs and their output data. Thus, the jobs are equivalent to vertices in a directed graph, with edges representing dependency and their weights modeling data communication costs, such as the number of bits, the price per bit, or transmission delay [44][42]. Hence, joint communication-computation resource sharing in MEC engenders a complicated graph partitioning problem on a directed graph where the edge weights also need to be suitably optimized.

The design of computation offloading may be further challenged by the need to satisfy job execution deadlines, e.g., when the common application supported by these jobs is delay sensitive. The problem of minimizing the makespan of executing multiple jobs in parallel, even in the simple case where there is no communication requirement among the jobs, is NP-hard [5]. Therefore, the general case of scheduling jobs with dependency and execution deadlines is among the hardest problems in MEC.

1.5.3 Multi-resource scheduling

The fairness and efficiency of resource allocation are fundamental problems in distributed computing systems. MEC requires the allocation of multiple communication and computation resources. Within each mobile edge host, the communication bandwidth, processor cycles, memory bandwidth, and on-board storage are shared among multiple users and applications. Scheduling jobs that require multiple resources is a challenging problem due to the combinatorial nature in allocating the blocks of different resources. Known results from the classical flow-shop problem suggest optimal scheduling is often intractable when three or more resources are considered [33].

The multi-resource fairness problem has been studied in the field of operations research [50], and seminal works in the networking and computing context have appeared in [20]. Most existing works focus on either a single machine or an isolated group of servers. In contrast, MEC incurs a unique scheduling environment across multiple access networks, mobile edge hosts, and remote servers. New multi-resource scheduling designs for MEC must account for this complexity. Furthermore, multi-resource scheduling necessarily leads to the tradeoff between fairness among jobs and system efficiency [25][46], which does not exist in traditional single-resource scheduling with work conservation. This is further exacerbated by the heterogeneous computation and communication capabilities of local and remote equipment [18][48]. It is particularly challenging to distribute jobs across multiple heterogeneous mobile edge hosts, since these hosts are shared by other users and applications with uncertain workload, and the access quality to these hosts are also random due to wireless channel fluctuation, spectrum sharing, and user mobility.

1.5.4 Mobility management

One main character of MEC is device mobility. When a mobile user is handed off from one base station to another, the association with all of its active applications running on the mobile edge hosts and the remote servers must remain intact. Furthermore, when the user moves too far away from its serving mobile edge host, the VMs created on that host for this user must migrate to a more suitable new host. Thus, mobility management in MEC involves both communication handoff and computation handoff. VM migration is costly in general [12], and in the case of MEC, it is further complicated by the heterogeneity in device software, mobile edge host capabilities, and the stringent delay constraints of some applications. From the operator's point of view, device mobility brings substantial challenges to a wide range of functionalities such as content caching, data analytics, and backhaul optimization.

For optimal system design, it will be necessary to accurately model the impact of mobile handoff. Furthermore, because of the high cost of handoff in MEC, it may be prudent to reduce the handoff frequency of a mobile device as it moves through the system. This can be achieved by disregarding opportunities for stronger connection with nearer base stations while the device's weaker connection with its present base station remains useful [40][3], and by delaying the migration of VMs even when the device connects with a new base station [43]. The former will be facilitated by the high density of base stations in 5G and beyond systems, with a tradeoff between handoff frequency and data rate. The latter will be supported by virtual network connections, with a tradeoff between handoff frequency and application latency. Thus, a main challenge in MEC is in optimally balancing these tradeoffs in a vastly complex system with many mobile devices and complicated connections among the mobile edge hosts.

1.5.5 Resource allocation and pricing

MEC offers unprecedented opportunities for innovative applications and services. These applications and services have diverse resource requirement profiles. Some have real-time demands with strict timeliness requirements, while others can tolerate some delay, and yet others may benefit from reserving resources for future use. Any resource allocation scheme will need to balance these diverse needs of different applications [9]. In particular, separate pools of resources may need to be reserved for steady long-term usage contracts and for the unpredictable arrival of urgent demands [51][45]. Moreover, the mobile edge hosts serve both user application jobs and MEC service jobs, and the allocation of communication and computation resources between these two types of jobs reflect the relative importance that the operator places on them. This adds another dimension of difficulty to the resource allocation problem.

Resource allocation is also tightly coupled with resource pricing. The mobile operator may charge the users a price for offloading user applications to mobile edge hosts. This may follow the pricing schemes used by current cloud computing service providers, e.g., charging higher for on-demand VMs and giving discounts for VMs that are reserved ahead of time. However, joint optimization of resource pricing and resource allocation remains an open problem in general cloud computing systems. Furthermore, since the offloaded user application in MEC may be further forwarded to remote cloud servers, which can be paid for directly by the users themselves or brokered by the mobile operator [39][47], the problem is substantially more complicated in MEC.

1.5.6 Network functions virtualization

As a core enabling technology, NFV is employed by the mobile edge hosts to create a wide variety of network appliances, such as routers, packet gateways, and Internet protocol (IP) multimedia subsystems, using generic hardware. By separating software from hardware, it allows dynamic provisioning of services and flexible deployment of network functions. However, the performance of current NFV implementations often falls well below that of dedicated hardware network equipment [24]. This is a particularly acute issue for small cells, where the base stations and their associated mobile edge hosts need to have a light footprint for flexible installation. Therefore, one difficult challenge in MEC is to improve the performance of virtualized services inside moderately endowed mobile edge hosts [29].

Furthermore, because of the performance limitation of NFV, it is essential to make judicious decisions on whether to keep certain network functions within the mobile core, or to virtualize them and move them to the mobile edge hosts. This decision should balance the tradeoff between hardware-equipment performance and NFV service flexibility [31]. An additional dimension to this issue is service latency, as the proximity of services to mobile devices is of paramount impor-

tance to the overall characteristic of MEC. Finally, since the mobile edge hosts are shared by user applications and NFV appliances, the dynamics of user communication traffic and computation offloading demand also have a high impact on the NFV decisions.

1.5.7 Security and privacy

MEC is a complex amalgamation of diverse technologies, including wireless networking, distributed computing, and the virtualization of networking equipment and computing servers. This opens up multiple fronts for malicious attacks, and a wide range of security measures are needed to thwart them [36][41]. In particular, due to the user-friendly location and limited size of mobile edge hosts, they cannot enjoy the physical protection afforded to large data centers. Furthermore, even if each component technology is individually secured, because of the complicated interactions and dependencies among them, there is no guarantee that the entire MEC system is secure.

Meanwhile, the MEC objective of ubiquitous and low-latency availability of information and computation resources, and the consequent requirement for high density and diversity of connections everywhere, make it difficult to set up a global secure perimeter. Vulnerabilities in a single mobile edge host can provide the means to launch an attack vector to the whole system. Furthermore, the existence of many VMs, spread across multiple mobile edge hosts, increases the chances of multiple compromised VMs coordinating in large-scale attacks such as distributed denial of service (DDoS). These security risks do not exist in previous generations of wireless networks, so there are scarce studies in the literature on countering measures.

Furthermore, the ever present conflict between system security and service agility is amplified in MEC. On one hand, MEC aims for efficient and responsive services, so that the resource overheads on security, such as authentication, access control, and intrusion detection, should be minimal. On the other hand, the unique MEC architecture, with its amalgamation of numerous heterogeneous components, requires strong security protection through complex multidimensional strategies. Adding to this challenge, the software nature of NFV-based implementation of security protocols can present a severe performance bottleneck in MEC [32]. We require efficient security strategies that match the unique characteristics of MEC, e.g., distributed authentication services implemented in mobile edge hosts [49].

Privacy is an important component of the overall system security. In MEC, with increased hardware and software connections between a mobile device and the network operator, there is higher risk for information leakage. MEC applications such as computation offloading, content caching, and augmented reality bring the frontline of privacy out of the relatively safe core mobile network. At the same time, they require a large amount of user data and user interaction, which adds to the difficulty in privacy protection. Whether it is information gathering

by the operator or privacy breaches by malicious attackers, in MEC there are more opportunities for security mis-steps. Networking and computing security functionalities should be installed in mobile edge hosts to balance the needs in both privacy protection and service performance, e.g., establishing trusted local proxies for anonymous access to information and computation [21], and replacing deep packet inspection with locally-aware machine-learning techniques for traffic classification and anomaly detection [30].

1.5.8 Integration with emerging technologies

MEC joins other advanced technologies in the 5G ecosystem, such as D2D communication [14] and cloud RAN (C-RAN) [8]. It is also expected to co-exist with other emerging information and computation technologies such as information-centric networking (ICN) [2], intelligent vehicular systems [19], and hybrid private-public cloud computing [23]. The successful deployment of MEC will depend on its integration with these new technologies.

Although the traffic of D2D communication does not go through base stations, the mobile edge hosts can serve important coordination functions. With their unique ability for monitoring and processing near the D2D nodes, and their inter-connectedness through the backhaul, the mobile edge hosts are well positioned to assist traffic scheduling and interference management in D2D communication. The main concern here is scalability as the D2D communication group increases in size, particularly in the IoT environment. The integration between MEC and C-RAN is also challenging, since C-RAN promotes light-weight remote radio heads. The mobile edge hosts may reside at the remote radio heads or the baseband units of C-RAN. In either case, the corresponding MEC platform and virtualization designs will need to account for the unique characteristics of C-RAN.

The mobile edge hosts provide ICN with convenient hardware and software resources, allowing content caching directly within the RANs. However, the current MEC framework is described under the assumption of a traditional TCP/IP network. It needs to evolve as the role of ICN becomes more prominent in the future Internet [34]. For intelligent vehicular systems, the ETSI MEC example use case given earlier covers only vehicle-to-infrastructure communication, while in vehicle-to-vehicle communication based on either D2D links or DSRC, the mobile edge hosts will serve only coordination functionalities. Proper operation of the vehicular system requires seamless integration of these two communication modes [28]. Finally, in hybrid cloud computing, an enterprise maintains local private cloud centers at the same time that it employs remote public cloud services [7]. Application scheduling and resource allocation in MEC will depending on whether the mobile edge hosts or the VMs residing within them have access to the private cloud.

1.6 Conclusion

MEC is a natural outcome of the emerging convergence between communication and computation. It requires multi-dimensional study into the complex amalgamation of diverse subjects in communication system, distributed computing, software engineering, and system optimization. Its standardization alongside 5G and future wireless technologies is poised to bring drastic changes on how wireless systems are designed, operated, and utilized. In this early stage of MEC development, there remain many challenging open problems and ample opportunities for future innovation.

Bibliography

[1] The 5G Infrastructure Public Private Partnership, "5G Vision: The 5G Infrastructure Public Private Partnership: The Next Generation of Communication Networks and Services," Feb. 2015.

[2] B. Ahlgren, C. Dannewitz, C. Imbrenda, D. Kutscher, and B. Ohlman, "A Survey of Information-Centric Networking," *IEEE Communications Magazine*, vol. 50, no. 7, pp. 26-36, Jul. 2012.

[3] W. Bao and B. Liang, "Stochastic Geometric Analysis of User Mobility in Heterogeneous Wireless Networks," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 10, pp. 2212-2225, Oct. 2015.

[4] F. Bonomi, R. Milito, J. Zhu, and S. Addepalli, "Fog Computing and Its Role in the Internet of Things," in *Proceedings of the ACM SIGCOMM Workshop on Mobile Cloud Computing*, Helsinki, Finland, Aug. 2012.

[5] D. G. Cattrysse and L. N. Van Wassenhove, "A Survey of Algorithms for the Generalized Assignment Problem," *European Journal of Operational Research*, vol. 60, no. 3, pp. 260-272, 1992.

[6] J. P. Champati and B. Liang, "Semi-online Task Partitioning and Communication between Local and Remote Processors," in *Proceedings of the IEEE International Conference on Cloud Networking (CLOUDNET)*, Niagara Falls, Canada, Oct. 2015.

[7] J. P. Champati and B. Liang, "One-Restart Algorithm for Scheduling and Offloading in a Hybrid Cloud," in *Proceedings of the IEEE/ACM International Symposium on Qualify of Service (IWQoS)*, Portland, USA, Jun. 2015.

[8] A. Checko, H. L. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. S. Berger, and L. Dittmann, "Cloud RAN for Mobile Networks - A Technology Overview," *IEEE Communications Surveys and Tutorials*, vol. 17, no. 1, pp. 405-426, First Quarter 2015.

[9] Y. Chen, A. Das, W. Qin, A. Sivasubramaniam, Q. Wang, and N. Gautam, "Managing Server Energy and Operational Costs in Hosting Centers," in *Proceedings of the ACM International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS)*, Banff, Canada, Jun. 2005.

[10] M.-H. Chen, M. Dong, and B. Liang, "Joint Offloading Decision and Resource Allocation for Mobile Cloud with Computing Access Point," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, Mar. 2016.

[11] B. Chun, S. Ihm, P. Maniatis, M. Naik, and A. Patti, "CloneCloud: Elastic Execution between Mobile Device and Cloud," in *Proceedings of the European Conference on Computer Systems (EuroSys)*, Salzburg, Austria, Apr. 2011.

[12] C. Clark, K. Fraser, S. Hand, J. G. Hansen, E. Jul, C. Limpach, I. Pratt, and A. Warfield, "Live Migration of Virtual Machines," in *Proceedings of*

the USENIX Symposium on Networked Systems Design and Implementation (NSDI), Berkeley, USA, May 2005.

[13] E. Cuervo, A. Balasubramanian, D. Cho, A. Wolman, S. Saroiu, R. Chandra, and P. Bahl, "MAUI: Making Smartphones Last Longer with Code Offload," in *Proceedings of the ACM International Conference on Mobile Systems, Applications, and Services (MobiSys)*, San Francisco, USA, Jun. 2010.

[14] K. Doppler, M. Rinne, C. Wijting, C. B. Ribeiro, and K. Hugl, "Device-to-Device Communication as an Underlay to LTE-Advanced Networks," *IEEE Communications Magazine*, vol. 47, no. 12, pp. 42-49, Dec. 2009.

[15] M. Drozdowski, *Scheduling for Parallel Processing*, Springer, 2009.

[16] ETSI Group Specification, "Mobile Edge Computing (MEC); Technical Requirements," ETSI GS MEC 002 V1.1.1 (2016-03), Mar. 2016.

[17] ETSI Group Specification, "Mobile Edge Computing (MEC); Framework and Reference Architecture," ETSI GS MEC 003 V1.1.1 (2016-03), Mar. 2016.

[18] E. Friedman, A. Ghodsi, and C.-A. Psomas, "Strategyproof Allocation of Discrete Jobs on Multiple Machines," in *Proceedings of the ACM Conference on Economics and Computation (EC)*, Palo Alto, USA, Jun. 2014.

[19] M. Gerla and L. Kleinrock, "Vehicular Networks and the Future of the Mobile Internet," *Computer Networks*, vol. 55, no. 2, pp. 457-469, Feb. 2011.

[20] A. Ghodsi, M. Zaharia, B. Hindman, A. Konwinski, S. Shenker, and I. Stoica, "Dominant Resource Fairness: Fair Allocation of Multiple Resource Types," in *Proceedings of the USENIX Symposium on Networked Systems Design and Implemenation (NSDI)*, Boston, USA, Mar. 2011.

[21] G. Ghinita, P. Kalnis, and S. Skiadopoulos, "PRIVE: Anonymous Location-Based Queries in Distributed Mobile Systems," in *Proceedings of the ACM International Conference on World Wide Web (WWW)*, Banff, Canada, May 2007.

[22] A. Greenberg, J. Hamilton, D. A. Maltz, and P. Patel, "The Cost of a Cloud: Research Problems in Data Center Networks," *ACM SIGCOMM Computer Communication Review*, vol. 39, no. 1, pp. 68-73, Dec. 2008.

[23] T. Guo, U. Sharma, P. Shenoy, T. Wood, and S. Sahu, "Cost-Aware Cloud Bursting for Enterprise Applications," *ACM Transactions on Internet Technology*, vol. 13, no. 3, pp. 10:110:24, May 2014.

[24] B. Han, V. Gopalakrishnan, L. Ji, and S. Lee, "Network Function Virtualization: Challenges and Opportunities for Innovations," *IEEE Communications Magazine*, vol. 53, no. 2, pp. 90-97, Feb. 2015.

[25] C. Joe-Wong, S. Sen, T. Lan, and M. Chiang, "Multi-Resource Allocation: Fairness-Efficiency Tradeoffs in a Unifying Framework," in *Proceedings of the IEEE International Conference on Computer Communications (INFOCOM)*, Orlando, USA, Mar. 2012.

[26] K. Kumar and Y.-H. Lu, "Cloud Computing for Mobile Users: Can Offloading Computation Save Energy?" *IEEE Computer*, vol. 43, no. 4, pp. 51-56, Apr. 2010.

[27] K. Kumar, J. Liu, Y.-H. Lu, and B. Bhargava, "A Survey of Computation Offloading for Mobile Systems," *Mobile Networks and Applications*, vol. 18, no. 1, pp. 129-140, Feb. 2013.

[28] E. Lee, E. K. Lee, M. Gerla, and S. Y. Oh, "Vehicular Cloud Networking: Architecture and Design Principles," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 148-155, Feb. 2014.

[29] A. Manzalini, R. Minerva, F. Callegati, W. Cerroni, and A. Campi, "Clouds of Virtual Machines in Edge Networks," *IEEE Communications Magazine*, vol. 51, no. 7, pp. 63-70, Jul. 2013.

[30] T. T. T. Nguyen and G. Armitage, "A Survey of Techniques for Internet Traffic Classification Using Machine Learning," *IEEE Communications Surveys and Tutorials*, vol. 10, no. 4, pp. 56-76, Fourth Quarter 2008.

[31] D. Oppenheimer, B. Chun, D. Patterson, A. C. Snoeren, and A. Vahdat, "Service Placement in a Shared Wide-Area Platform," in *Proceedings of the USENIX Annual Technical Conference*, Boston, USA, Jun. 2006.

[32] G. Pek, L. Buttyan, and B. Bencsath, "A Survey of Security Issues in Hardware Virtualization," *ACM Computing Surveys*, vol. 45, no. 3, pp. 40:1-40:34, Jul. 2013.

[33] M. L. Pinedo, *Scheduling: Theory, Algorithms, and Systems*, Springer, 2012.

[34] R. Ravindran, X. Liu, A. Chakraborti, X. Zhang, and G. Wang, "Towards Software Defined ICN Based Edge-Cloud Services," in *Proceedings of the IEEE International Conference on Cloud Networking (CloudNet)*, San Francisco, USA, Nov. 2013.

[35] M. Satyanarayanan, P. Bahl, R. Caceres, and N. Davies, "The Case for VM-Based Cloudlets in Mobile Computing," *IEEE Pervasive Computing*, vol. 8, no. 4, pp. 14-23, Oct.-Dec. 2009.

[36] M. Satyanarayanan, G. Lewis, E. Morris, S. Simanta, J. Boleng, and K. Ha, "The Role of Cloudlets in Hostile Environments," *IEEE Pervasive Computing*, vol. 12, no. 4, pp. 40-49, Oct.-Dec. 2013.

[37] B. Sharma, V. Chudnovsky, J. L. Hellerstein, R. Rifaat, and C. R. Das, "Modeling and Synthesizing Task Placement Constraints in Google Compute Clusters," in *Proceedings of the ACM Symposium on Cloud Computing (SoCC)*, Cascais, Portugal, Oct. 2011.

[38] D. B. Shmoys, J. Wein, and D. P. Williamson, "Scheduling Parallel Machines On-Line," *SIAM Journal on Computing*, vol. 24, no. 6, pp. 1313-1331, Dec. 1995.

[39] Y. Song, M. Zafer, and K.-W. Lee, "Optimal Bidding in Spot Instance Market," in *Proceedings of the IEEE International Conference on Computer Communications (INFOCOM)*, Orlando, USA, Mar. 2012.

[40] E. Stevens-Navarro, Y. Lin, and V. W. S. Wong, "An MDP-Based Vertical Handoff Decision Algorithm for Heterogeneous Wireless Networks," *IEEE Transactions on Vehicular Technology*, vol. 57, no. 2, pp. 1243-1254, Mar. 2008.

[41] I. Stojmenovic, S. Wen, X. Huang, and H. Luan, "An Overview of Fog Computing and Its Security Issues," in press, to appear in *Concurrency and Computation: Practice and Experience*.

[42] S. Sundar and B. Liang, "Communication Augmented Latest Possible Scheduling for Cloud Computing with Delay Constraint and Task Dependency," in *Proceedings of the IEEE INFOCOM Workshop on Green and Sustainable Networking and Computing (GSNC)*, San Francisco, USA, Apr. 2016.

[43] R. Urgaonkara, S. Wang, T. He, M. Zafer, K. Chan, and K. K. Leung, "Dynamic Service Migration and Workload Scheduling in Edge-Clouds," *Performance Evaluation*, vol. 91, pp. 205-228, Sep. 2015.

[44] C. Wang and Z. Li, "Parametric Analysis for Adaptive Computation Offloading," in *Proceedings of the ACM SIGPLAN Conference on Programming Language Design and Implementation (PLDI)*, Washington, USA, Jun. 2004.

[45] W. Wang, B. Li, and B. Liang, "Towards Optimal Capacity Segmentation with Hybrid Cloud Pricing," in *Proceedings of the IEEE International Conference on Distributed Computing Systems (ICDCS)*, Macau, China, Jun. 2012.

[46] W. Wang, C. Feng, B. Li, and B. Liang, "On the Fairness-Efficiency Tradeoff for Packet Processing with Multiple Resources," in *Proceedings of the ACM SIGCOMM International Conference on Emerging Networking Experiments and Technologies (CoNEXT)*, Sydney, Australia, Dec. 2014.

[47] W. Wang, D. Niu, B. Liang, and B. Li, "Dynamic Cloud Resource Reservation via IaaS Cloud Brokerage," *IEEE Transactions on Parallel and Distributed Systems*, vol. 26, no. 6, pp. 1580-1593, Jun. 2015.

[48] W. Wang, B. Liang, and B. Li, "Multi-resource Fair Allocation in Heterogeneous Cloud Computing Systems," *IEEE Transactions on Parallel and Distributed Systems*, vol. 26, no. 10, pp. 2822-2835, Oct. 2015.

[49] R. Yahalom, B. Klein, and T. Beth, "Trust Relationships in Secure Systems - a Distributed Authentication Perspective," in *Proceedings of the IEEE Computer Society Symposium on Research in Security and Privacy*, Oakland, USA, May 1993.

[50] H. P. Young, *Equity: In Theory and Practice*, Princeton University Press, 1994.

[51] Q. Zhang, Q. Zhu, and R. Boutaba, "Dynamic Resource Allocation for Spot Markets in Cloud Computing Environments," in *Proceedings of the IEEE International Conference on Utility and Cloud Computing (UCC)*, Victoria, Australia, Dec. 2011.

[52] B. Zhou, A. V. Dastjerdi, R. N. Calheiros, S. N. Srirama and R. Buyya, "A Context Sensitive Offloading Scheme for Mobile Cloud Computing Service," in *Proceedings of the IEEE International Conference on Cloud Computing (CLOUD)*, New York, USA, Jun. 2015.