Mobile Edge Computing: Recent Efforts and Five Key Research Directions

Tuyen X. Tran, Mohammad-Parsa Hosseini, and Dario Pompili
Department of Electrical and Computer Engineering
Rutgers University–New Brunswick, NJ, USA
{tuyen.tran, parsa, pompili}@cac.rutgers.edu

1. Introduction

In the past decade, we have witnessed Cloud Computing play as significant role for massive data storage, control, and computation offloading. However, the rapid proliferation of mobile applications and the Internet of Things (IoT) over the last few years has posed severe demands on cloud infrastructure and wireless access networks. Stringent requirements such as ultra-low latency, user experience continuity, and high reliability are driving the need for highly localized intelligence in close proximity to the end users. In light of this, Mobile Edge Computing (MEC) has been envisioned as the key technology to assist wireless networks with cloud computing-like capabilities in order to provide low-latency and context-aware services directly from the network edge.

Differently from traditional cloud computing systems where remote public clouds are utilized, the MEC paradigm is realized via the deployment of commodity servers, referred to as the MEC servers, at the edge of the wireless access network. Depending on different functional splitting and density of the Base Stations (BSs), a MEC server can be deployed per BS or at an aggregation point serving several BSs. With the strategic deployment of these computing servers, MEC allows for data transfer and application execution in close proximity to the end users, substantially reducing end-to-end (e2e) delay and releasing the burden on backhaul network [1]. Additionally, MEC has the potential to empower the network with various benefits, including: (i) optimization of mobile resources by hosting compute-intensive applications at the network edge, (ii) pre-processing of large data before sending it (or some extracted features) to the cloud, and (iii) context-aware services with the help of Radio Access Network (RAN) information such as cell load, user locations, and radio resource allocation.

In this letter, as a backdrop to identifying research questions, we briefly review recent research efforts on enabling MEC technologies and then discuss five key research directions. Specifically, the goals of this letter are: (i) to raise awareness of relevant and cutting-edge work being performed from various literature, and (ii) to identify a number of important research needs for future MEC systems.

2. Recent Efforts in Enabling MEC Technologies

Fueled by the promising capabilities and business opportunities, the MEC paradigm has been attracting considerable attention from both academia and industry. A number of deployment scenarios, service use cases, and related algorithms design has been proposed to exploit the potential benefits of MEC and to justify its implementation and deployment from both a technical and business point of view. In this section, we briefly review the recent efforts from both standardization and research perspectives towards enabling MEC technologies in wireless networks.

2.1 Proofs of Concepts and Standardization Efforts

In 2013, Nokia Networks introduced the very first real-world MEC platform [2], in which the computing platform–Radio Applications Cloud Servers (RACS)–is fully integrated with the Flexi Multiradio BS. Saguna also introduced their fully virtualized MEC platform, so called Open-RAN [3], that can provide an open environment for running third-party MEC applications. Besides these solutions, MEC standardization is being specified by the European Telecommunications Standards Institute (ETSI), which recently formed a MEC Industry Specifications Group (ISG) to standardize and moderate the adoption of MEC within the RAN. In the introductory white paper [4], four typical service scenarios and their relationship to MEC have been discussed, ranging from Augmented Reality (AR) and intelligent video acceleration to connected cars and IoT gateway. In the MEC World Congress 2016, ETSI has announced six Proofs of Concept (PoCs) that were accepted by the MEC ISG, including:

- Radio Aware Video Optimization in a Fully Virtualized Network (RAVEN);
- Flexible IP-based Services (FLIPS);
- Enterprise Services;
- Healthcare–Dynamic Hospital User, IoT, and Alert Status Management;
- Multi-Service MEC Platform for Advanced Service Delivery;

- Video Analytics.

These PoCs strengthen the strategic planning and decision making of organizations, helping them identify which MEC solutions may be viable in the network. Also in this Congress, ETSI MEC ISG has renamed Mobile Edge Computing as Multi-access Edge Computing in order to reflect the growing interest in MEC from non-cellular operators, which takes effect starting from 2017 [5]. The technical requirements for MEC are specified in [6] to guarantee interoperability and to promote MEC deployment. These requirements are divided into generic requirements, service requirements, requirements on operation and management, and finally security, regulations and charging requirements. Most recently, the 3GPP has shown a growing interest in incorporating MEC into its 5G standard and has identified functionality supports for edge computing in a recent technical specification contribution [7].

2.2 MEC Architecture and Virtualization

In recent years, the concept of integrating cloud computing-capabilities into the wireless network edge has been considered in the literature under different terminologies, including Small Cell Cloud (SCC), Mobile Micro Cloud (MMC), Follow Me Cloud (FMC), and CONCERT [8]. The basic idea of SCC is to enhance the small cells, such as microcells, picocells or femtocells, with additional computation and storage capabilities so as to support edge computing [9]. By exploiting the Network Function Virtualization (NFV) paradigm, the cloud-enabled small cells can pool their computation power to provide users with services/applications having stringent latency requirements. Similarly, the concept of MMC introduced in [10] allows users to have instantaneous access to the cloud services with low latency. Differently from the SCC where the computation/storage resources are provided by interworking clusters of enhanced small cells, the User Equipment (UE) exploits the computation resources of a single MMC, which is typically connected directly to a BS. The FMC concept [11] proposes to move computing resources a bit further from the UEs, compared to SCC and MMC, to the core network. It aims at having the cloud services running at distributed data centers so as to be able to follow the UEs as they roam throughout the network. In all these described MEC concepts, the computing/storage resources have been fully distributed; conversely, the CONCERT concept proposes hierarchically placement of the resources within the network in order to flexibly and elastically manage the network and cloud services.

2.3 Computation Offloading

The benefits of computation offloading have been investigated widely in conventional Mobile Cloud Computing (MCC) systems. However, a large body of existing works on MCC assumed an infinite amount of computing resources available in a cloudlet, where offloaded tasks can be executed in negligible delay [12], [13]. Recently, several works have focused on exploiting the benefits of computation offloading in MEC network [14]. The problem of offloading scheduling was then reduced to radio resource allocation in [15], where the competition for radio resources is modeled as a congestion game of selfish mobile users. The problem of joint task offloading and resource allocation was studied in a single-user system with energy harvesting devices [16], and in a multi-cell, multi-user systems [17]; however, the congestion of computing resources at the MEC server was not taken into account. A similar problem is studied in [18] for single-server MEC systems, where the limited resources at the MEC server were factored in, and later on extended to multi-server MEC systems in [19].

2.4 Edge Caching

The increasing demand for massive multimedia services over mobile cellular network poses great challenges on network capacity and backhaul links. Distributed edge caching, which can well leverage MEC paradigm, has therefore been recognized as a promising solution to bring popular contents closer to the users, to reduce data traffic going through the backhaul links as well as the time required for content delivery, and to help smoothen/regulate the traffic during peak hours. In general, edge caching in wireless networks has been investigated in a number of works (cf. [20-22] and references therein). Recently, in [23], [24], we have proposed a cooperative hierarchical caching paradigm in a Cloud Radio Access Network (C-RAN) where the cloud-cache is introduced as a bridging layer between the edge-based and core-based caching schemes. Taking into account the heterogeneity of video transmissions in wireless networks in terms of video quality and device capabilities, our previous work in [25] proposes to utilize both caching and processing capabilities at the MEC servers to satisfy users' requests for videos with different bitrates. In this scheme, the collaborative caching paradigm has been extended to a new dimension where the MEC servers can assist each other to not only provide the requested video via backhaul links but also to transcode it to the desired bitrate version.

3. Five Key Research Directions for MEC in Wireless Networks

Research on MEC lies at the intersection of wireless communications and cloud computing, which has resulted in many interesting research opportunities and challenges. The spectrum of research required to achieve the promises of MEC requires significant investigation along many directions. In this section, we highlight and discuss the key open research issues and future directions, which are categorized into five main topics as follows.

3.1 Deployment Scenarios and Resource Management

The key concept of MEC is to shift the cloud computing-capabilities closer to the end users in order to reduce the service latency and to avoid congestion in the core network. However, there has been no formal definition on what the MEC servers would be and where they should be deployed within the network. Such decisions involve investigating the site-selection problem for MEC servers where their optimal placement is coupled with the computational resource provisioning as well as with the deployment budget. In addition, it is critical to determine the required server density to cope with the service demands, which is closely related to the infrastructure deployment cost and marketing strategies. Finally, the deployment of MEC servers also depends on the RAN architecture where different functional splitting options between the BSs and the centralized processing center (such as in C-RAN) are specified, depending on the delay requirement and fronthaul capacity.

3.2 Computation Caching and Offloading

The combination of computation and storage resources at the MEC servers offers unique opportunities for caching of computation tasks. In this technique, the MEC server can cache several application services and their related database, and handle the offloaded computation from multiple users so as to enhance the user experience. Computation caching can help decrease the load on the access link by providing computing results to the end users without the need to fetch their tasks beforehand. Unlike content caching, computation caching presents several new challenges. First, computing tasks can be of diverse types and depend on the computing environment; while some of the content is cacheable for reuse by other devices, personal computing data is not cacheable and must often be executed in real time. Second, it is not practical to build popularity patterns locally at each server; instead, studying popularity distributions over larger sets of servers can provide a broader view on the popularity patterns of computing tasks.

3.3 IoT Applications and Big Data Analytics

The emerging IoT and Big Data services have changed the traditional networking paradigm where the network infrastructure, instead of being the dump pipe, can now process the data and generate insights. MEC resources can be utilized for pre-processing of massive IoT data so as to reduce bandwidth consumption, to provide network scalability, and to ensure a fast response to the user requests. A MEC platform can also encompass a local IoT gateway functionality capable of performing data aggregation and big data analytics for event reporting, smart grid, e-health, and smart cities. For instance, our previous work in [26] describes an autonomic edge-computing platform that supports deep learning for localization of epileptogenicity using multimodal rs-fMRI and EEG big data. To fully exploit the benefits of MEC for IoT, there needs to be significant research on how to efficiently distribute and manage data storage and computing, how to make edge computing collaborate with cloud computing for more scalable services, and how to secure the whole system.

3.4 Mobility Management

Mobility management is an essential feature for MEC to ensure service continuity for highly dynamic mobile users. For vehicular communications and automotive, integrating MEC with mobile cloud computing or vehicular cloud, wherein mobile or vehicle resources are utilized for communication and computation services, is a highly challenging issue from the service orchestration perspective. For many applications, estimating and predicting the movement and trajectory of users as well as personal preference information can help the MEC servers improve the user experience. For example, mobility prediction can be integrated with edge caching to enhance the content migration at the edges and caching efficiency. In addition, to achieve better user computation experience, existing offloading techniques can be jointly considered with mobility-aware scheduling policies at the MEC servers. This approach introduces a set of interesting research problems including mobility-aware online prefetching of user computation data, server scheduling, and fault-tolerance computation. For instance, in our previous works [27], [28], multi-tier distributed computing infrastructures based on MEC and Mobile Device Cloud (MDC) are proposed to link mobility management and pervasive computing with medical applications.

3.5 Security and Privacy

Security issues might hinder the success of the MEC paradigm if not carefully considered. Unlike traditional cloud computing, MEC infrastructure is vulnerable to site attacks due to its distributed deployment. In addition, MEC requires more stringent security policies as third-party stakeholders can gain access to the platform and derive information regarding user proximity and network analytics. Existing centralized authentication protocols might not be applicable for some parts of the infrastructure that have limited connectivity to the central authentication server. It is also important to implement trust-management systems that are able to exchange compatible trust information with each other, even if they belong to different trust domains. Furthermore, as service providers want to acquire user information to tailor their services, there is a great challenge to the development of privacy-protection mechanisms that can efficiently protect users' locations and service usage.

4. Conclusion

Mobile Edge Computing (MEC) is an emerging technology to cope with the unprecedented growth of user demands for access to low-latency computation and content data. This paradigm, which aims at bringing the computing and storage resources to the edge of mobile network, allows for the execution of delay-sensitive and context-aware applications in close proximity to the end users while alleviating backhaul utilization and computation at the core network. While research on MEC has gained its momentum, as reflected in the recent efforts reviewed in this letter, MEC itself is still in its nascent stage and there is a myriad of technical challenges that need to be addressed. In this regard, we discussed five key open research directions that we consider to be among the most important and challenging issues of future MEC systems.

References

[1] T. X. Tran, A. Hajisami, P. Pandey, and D. Pompili, "Collaborative Mobile Edge Computing in 5G Networks: New Paradigms, Scenarios, and Challenges," IEEE Communications Magazine, vol. 55, no. 4, pp. 54-61, 2017.

[2] Intel and Nokia Siemens Networks, "Increasing mobile operators' value proposition with edge computing," Technical Brief, 2013.

[3] Saguna and Intel, "Using mobile edge computing to improve mobile network performance and profitability," White paper, 2016.

[4] Y. C. Hu, M. Patel, D. Sabella, N. Sprecher, and V. Young, "Mobile Edge Computing — A Key Technology Towards 5G," ETSI white paper, vol. 11, 2015.

[5] N. Sprecher, J. Friis, R. Dolby, and J. Reister, "Edge computing prepares for a multi-access future," Sep. 2016. [Online]. Available: http://www.telecomtv.com/articles/mec/edge-computing-prepares-for-a-multi-access-future-13986/

[6] ETSI GS MEC 002: Mobile Edge Computing (MEC); Technical Requirements V1.1.1, March 2016.

[7] 3GPP, "Technical specification group services and system aspects; system architecture for the 5g systems; stage 2 (release 15)," 3GPP TS 23.501 V0.4.0, Apr. 2017.

[8] J. Liu, T. Zhao, S. Zhou, Y. Cheng, and Z. Niu, "CONCERT: a cloud-based architecture for next-generation cellular systems", IEEE Wireless Communications, vol. 21, no. 6, pp. 14-22, Dec. 2014.

[9] FP7 European Project, "Distributed computing, storage and radio resource allocation over cooperative femtocells (TROPIC)," [Online]. Available: http://www.ict-tropic.eu/, 2012.

[10] S. Wang, et al., "Mobile Micro-Cloud: Application Classification, Mapping, and Deployment", Annual Fall Meeting of ITA (AMITA), 2013.

[11] T. Taleb, A. Ksentini, and P. A. Frangoudis, "Follow-Me Cloud: When Cloud Services Follow Mobile Users," IEEE Transactions on Cloud Computing, vol PP, no. 99, pp. 1-1, May 2017.

[12] W. Zhang, Y. Wen, and D. O. Wu, "Collaborative task execution in mobile cloud computing under a stochastic wireless channel," IEEE Trans. Wireless Commun., vol. 14, no. 1, pp. 81–93, 2015.

[13] Z. Cheng, P. Li, J. Wang, and S. Guo, "Just-in-time code offloading for wearable computing," IEEE Trans. Emerg. Topics Comput., vol. 3, no. 1, pp. 74–83, 2015.

[14] P. Mach and Z. Becvar, "Mobile edge computing: A survey on architecture and computation offloading." IEEE Communications Surveys & Tutorials, Mar. 2017.

[15] X. Chen, "Decentralized computation offloading game for mobile cloud computing," IEEE Trans. Parallel Distrib. Syst., vol. 26, no. 4, pp. 974–983, 2015.

[16] Y. Mao, J. Zhang, and K. B. Letaief, "Dynamic computation offloading for mobile-edge computing with energy harvesting devices," IEEE J. Sel. Areas in Commun., vol. 34, no. 12, pp. 3590–3605, 2016.

[17] S. Sardellitti, G. Scutari, and S. Barbarossa, "Joint optimization of radio and computational resources for multicell mobile-

edge computing," IEEE Trans. Signal Inf. Process. Over Netw., vol. 1, no. 2, pp. 89–103, 2015.

[18] X. Lyu, H. Tian, P. Zhang, and C. Sengul, "Multi-user joint task offloading and resources optimization in proximate clouds," IEEE Trans. Veh. Technol., vol. 66, no. 4, pp. 3435-3447, April 2017.

[19] T. X. Tran and D. Pompili. "Joint Task Offloading and Resource Allocation for Multi-Server Mobile-Edge Computing Networks." arXiv preprint arXiv:1705.00704, 2017.

[20] E. Bastug, M. Bennis, and M. Debbah, "Living on the edge: The role of proactive caching in 5G wireless networks," IEEE Communications Magazine, vol. 52, no. 8, pp. 82-89, 2014.

[21] N. Golrezaei, K. Shanmugam, A. G. Dimakis, A. F. Molisch, and G. Caire, "Femtocaching: Wireless video content delivery through distributed caching helpers," in Proc. IEEE INFOCOM, pp. 1107-1115, 2012.

[22] H. Ahlehagh and S. Dey, "Video-aware scheduling and caching in the radio access network," IEEE/ACM Transactions on Networking, vol. 22, no. 5, pp. 1444-1462, 2014.

[23] T. X. Tran and D. Pompili, "Octopus: A Cooperative Hierarchical Caching Strategy for Cloud Radio Access Networks," in Proc. IEEE Int. Conf. on Mobile Ad hoc and Sensor Systems (MASS), pp. 154-162, Oct. 2016.

[24] T. X. Tran, A. Hajisami, and D. Pompili, "Cooperative Hierarchical Caching in 5G Cloud Radio Access Networks (C-RANs)," IEEE Network, July 2017.

[25] T. X. Tran, P. Pandey, A. Hajisami, and D. Pompili, "Collaborative Multi-Bitrate Video Caching and Processing in Mobile-Edge Computing Networks," in Proc. IEEE/IFIP Conference on Wireless On-demand Network Systems and Services (WONS), pp.165-172, 2017.

[26] M. P. Hosseini, T. X. Tran, D. Pompili, K. Elisevich, and H. Soltanian-Zadeh, "Deep Learning with Edge Computing for Localization of Epileptogenicity using Multimodal rs-fMRI and EEG Big Data," in Proc. IEEE Int'l Conf. Autonomic Computing (ICAC), July 2017, to appear.

[27] M. P. Hosseini, A. Hajisami, and D. Pompili, "Real-time Epileptic Seizure Detection from EEG Signals via Random Subspace Ensemble Learning," in Proc. IEEE Int'l Conf. Autonomic Computing (ICAC), Wurzburg, Germany, Jul. 2016.

[28] M. P. Hosseini, H. Soltanian-Zadeh, K. Elisevich, and D. Pompili, "Cloud-based Deep Learning of Big EEG Data for Epileptic Seizure Prediction," IEEE Global Conference on Signal and Information Processing (GlobalSIP), Washington, D.C., Dec. 2016.

Tuyen X. Tran is working towards his PhD degree in Electrical & Computer Engineering (ECE) at Rutgers University, NJ. He is pursuing research in the fields of wireless communications and mobile cloud computing, with emphasis on Cloud Radio Access Networks and Mobile-Edge Computing. He received the MSc degree in ECE from the University of Akron, USA, in 2013, and the BEng degree (Honors Program) in Electronics and Telecommunications from Hanoi University of Technology, Vietnam, in 2011. He was a research intern at Huawei Technologies R&D Center, Bridgewater, NJ, during the summers of 2015 and 2016. He was the recipient of the Best Paper Award at the IEEE/IFIP Wireless On-demand Network systems and Services Conference (WONS) 2017.

Mohammad-Parsa Hosseini is a Senior Member of IEEE. He is a PhD candidate and a research assistant in the Dept. of ECE at Rutgers University, NJ, USA and a member of the CPS Lab under the guidance of Prof. Pompili. He is collaborating with Medical Image Analysis Lab at Henry Ford Health System, MI, under the guidance of Prof. Soltanian-Zadeh specifically in neuroimaging and data science. He is collaborating with the Clinical Neurosciences Dept. Spectrum Health, MI, under the guidance of Prof. Elisevich in computational neuroscience. His research focuses on deep/machine learning, signal/image processing, big data, and cloud computing with future applications in the field of health care. He was a research intern at Apple Inc., Silicon Valley, CA, during the summer of 2017. Previously, he was a PhD student in the ECE Dept. of Wayne State University, MI, in 2013 and he has been teaching as an adjunct professor at several universities since 2009.

Dr. Dario Pompili is an Assoc. Prof. with the Dept. of ECE at Rutgers U. He is the director of the Cyber-Physical Systems Laboratory (CPS Lab), which focuses on mobile computing, wireless communications and networking, acoustic communications, sensor networks, and datacenter management. He received his PhD in ECE from the Georgia Institute of Technology in June 2007. He had previously received his 'Laurea' (combined BS and MS) and Doctorate degrees in Telecommunications and System Engineering from the U. of Rome "La Sapienza," Italy, in 2001 and 2004, respectively. He is a recipient of the NSF CAREER'11, ONR Young Investigator