Collaborative Mobile Edge Computing in 5G Networks: New Paradigms, Scenarios, and Challenges

Tuyen X. Tran, Abolfazl Hajisami, Parul Pandey, and Dario Pompili

The authors envision a real-time, context-aware collaboration framework that lies at the edge of the RAN, comprising MEC servers and mobile devices, and amalgamates the heterogeneous resources at the edge. Specifically, they introduce and study three representative use cases ranging from mobile edge orchestration, collaborative caching and processing, and multi-layer interference cancellation.

Abstract

MEC is an emerging paradigm that provides computing, storage, and networking resources within the edge of the mobile RAN. MEC servers are deployed on a generic computing platform within the RAN, and allow for delay-sensitive and context-aware applications to be executed in close proximity to end users. This paradigm alleviates the backhaul and core network and is crucial for enabling low-latency, high-bandwidth, and agile mobile services. This article envisions a real-time, context-aware collaboration framework that lies at the edge of the RAN, comprising MEC servers and mobile devices, and amalgamates the heterogeneous resources at the edge. Specifically, we introduce and study three representative use cases ranging from mobile edge orchestration, collaborative caching and processing, and multi-layer interference cancellation. We demonstrate the promising benefits of the proposed approaches in facilitating the evolution to 5G networks. Finally, we discuss the key technical challenges and open research issues that need to be addressed in order to efficiently integrate MEC into the 5G ecosystem.

Introduction

Over the last few years, our daily lifestyle is increasingly exposed to a plethora of mobile applications for entertainment, business, education, health care, social networking, and so on. At the same time, mobile data traffic is predicted to continue doubling each year. To keep up with these surging demands, network operators have to spend enormous efforts to improve users' experience while maintaining healthy revenue growth. To overcome the limitations of current radio access networks (RANs), two emerging paradigms have been proposed:

- Cloud RAN (C-RAN), which aims at the centralization of base station (BS) functions via virtualization
- Mobile edge computing (MEC), which proposes to empower the network edge

While the two technologies propose to move computing capabilities in a different direction (to the cloud vs. to the edge), they are complementary, and each has a unique position in the fifth generation (5G) ecosystem.

As depicted in Fig. 1, MEC servers are implemented directly at the BSs using a generic computing platform, allowing the execution of applications in close proximity to end users. With this position, MEC can help fulfill the stringent low-latency requirement of 5G networks. Additionally, MEC offers various network improvements, including:

- Optimization of mobile resources by hosting compute-intensive applications at the network edge
- Pre-processing of large data before sending it (or some extracted features) to the cloud
- Context-aware services with the help of RAN information such as cell load, user location, and allocated bandwidth

Although the MEC principle also aligns with the concept of *fog computing* [1], and the two are often referred to interchangeably, they slightly differ from each other. While fog computing is a general term that opposes cloud computing in bringing the processing and storage resources to the lower layers, MEC specifically aims at extending these capabilities to the edge of the RAN with new function splitting and a new interface between the BSs and the upper layer. Fog computing is most commonly seen in enterprise-owned gateway devices, whereas MEC infrastructure is implemented and owned by the network operators.

Fueled by the potential capabilities of MEC, we propose a real-time context-aware collaboration framework that lies at the edge of the cellular network and works side by side with the underlying communication network. In particular, we aim at exploring the synergies among connected entities in the MEC network to form a heterogeneous computing and storage resource pool. To illustrate the benefits and applicability of MEC collaboration in 5G networks, we present three use cases including mobile edge orchestration, collaborative video caching and processing, and multi-layer interference cancellation. These initial target scenarios can be used as the basis for the formulation of a number of specific applications.

The remainder of this article is organized as follows. In the following section, we present the state of the art on MEC. Then we provide a comparison between MEC and C-RAN in various

Digital Object Identifier: 10.1109/MCOM.2017.1600863

The authors are with Rutgers University.

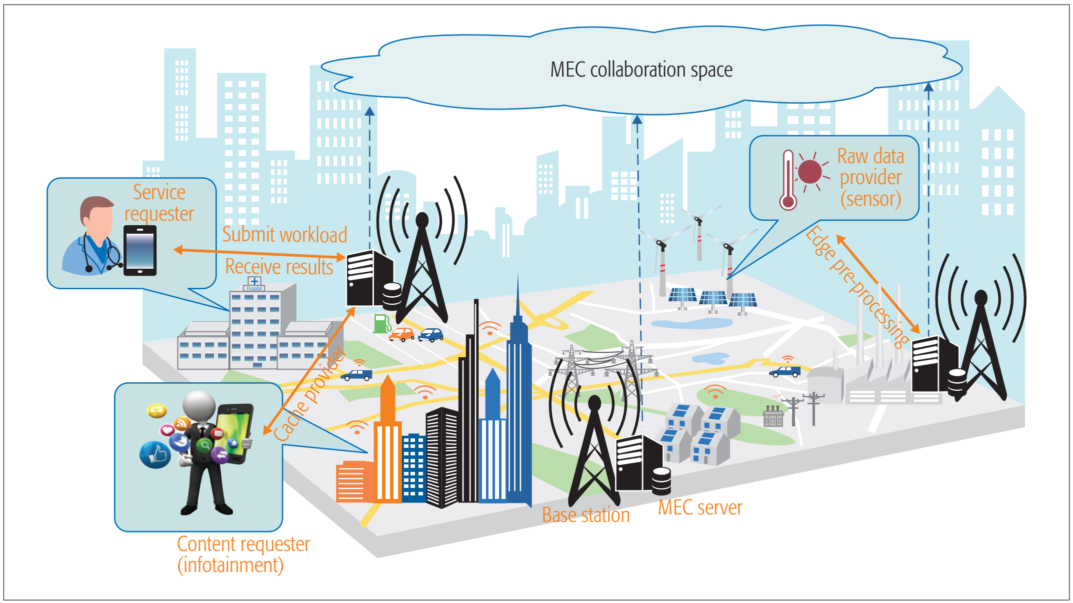

Figure 1. Illustration of Mobile Edge Computing Network.

In contrast to existing works on MEC, which do not explore the synergies among the MEC entities, this article takes one step further by proposing a collaborative MEC paradigm and presents three strong use cases to efficiently leverage this collaboration space.

features. Following that, we describe the three case studies to illustrate the applicability and benefits of the collaborative MEC paradigm. Then we highlight some key challenges and open research issues that need to be tackled. Finally, we draw our conclusions in the final section.

State of the Art

In 2013, Nokia Networks introduced a very first real-world MEC platform [2] in which the computing platform — radio applications cloud servers (RACS) — is fully integrated with the Flexi Multiradio base station. Under a scenario of a "smarter city" [3], IBM discusses how operators can leverage the capabilities of mobile edge network virtualization to deploy disruptive services for consumers and enterprises. Saguna also introduces their fully virtualized MEC platform, Open-RAN [4], which can provide an open environment for running third-party MEC applications. Recently, the European Telecommunications Standards Institute (ETSI) formed a MEC Industry Specifications Group (ISG) in order to standardize and moderate the adoption of MEC within the RAN [5].

From the theoretical perspective, the authors in [6] consider the computation offloading problem in a multi-cell MEC network, where a dense deployment of radio access points facilitates proximity high-bandwidth access to computational resources but also increases inter-cell interference. The authors in [7] provide a collective overview of the opportunities and challenges of "fog computing" in the networking context of the Internet of Things (IoT). Several case studies are presented to highlight the potential and challenges of the fog control plane such as interference, control, configuration, and management of networks, and so on (http://Fogresearch.org).

In summary, prior works on MEC focused on feasibility of MEC-RAN integration, deployment scenarios, and potential services and applications. *In contrast to existing works on MEC, which do not explore the synergies among the MEC entities, this article takes one step further by proposing a collaborative MEC paradigm and presents three strong use cases to efficiently leverage this collaboration space.*

MEC vs. C-RAN

A redesigned centralization of RAN is proposed as C-RAN, where the physical layer communication functionalities are decoupled from the distributed BSs and are consolidated in a virtualized central processing center. With its centralized nature, it can be leveraged to address the capacity fluctuation problem and to increase system energy efficiency in mobile networks [8]. Besides an approach to 5G standardization, C-RAN can provide new opportunities for IoT, opening up a new horizon of ubiquitous sensing, interconnection of devices, service sharing, and provisioning to support better communication and collaboration among people and things in a more distributed and dynamic manner. The integration of cloud provider, edge gateways, and end devices can support powerful processing and storage facilities to massive IoT data streams (big data) beyond the capability of individual "things" as well as provide automated decision making in real time. Thus, the C-RAN and IoT convergence can enable the development of new innovative applications in various emerging areas such as smart cities, smart grids, smart healthcare, and others aimed at improving all aspects of human life.

The full centralization principle of C-RAN, however, entails the exchange of radio signals between the radio heads and cloud processing unit, which imposes stringent requirement to the fronthaul connections in terms of throughput and latency. On the other hand, the MEC paradigm is useful in reducing latency and improving localized user experience, but the amount of processing power and storage is orders of magnitude below that of the centralized cloud in C-RAN. In Table 1, we summarize the comparison between MEC and C-RAN in various aspects. One important note is that MEC does not contradict with C-RANs but rather complement them. For example, an appli-

The proposed novel resource management framework lies at the intermediate edge layer and orchestrates both the horizontal collaboration at the end-user layer and the MEC layer as well as the vertical collaboration between end users, edge nodes, and cloud nodes.

	MEC	C-RAN
Location	Co-located with base stations or aggregation points.	Centralized, remote data centers.
Deployment planning	Minimal planning with possible ad hoc deployments.	Sophisticated.
Hardware	Small, heterogeneous nodes with moderate computing resources.	Highly capable computing servers.
Fronthaul requirements	Fronthaul network bandwidth requirements grow with the total amount of data that need to be sent to the core network after being filtered/processed by MEC servers.	Fronthaul network bandwidth requirements grow with the total aggregated amount of data generated by all users.
Scalability	High	Average, mostly due to expensive fronthaul deployment.
Application delay	Support time-critical applications that require latencies less than tens of milliseconds.	Support applications that can tolerate round-trip delays on the order of a few seconds or longer.
Location awareness	Yes	N/A
Real-time mobility	Yes	N/A

Table 1. Comparison of features: MEC vs. C-RAN.

cation that needs to support very low end-to-end delay can have one component running in the MEC cloud and other components running in the distant cloud.

In the following sections, we present our case studies where we propose novel scenarios and techniques to take advantage of the collaborative MEC systems.

Case Study I: Mobile Edge Orchestration

In spite of the limited resources (e.g., battery, CPU, memory) on mobile devices, many computation-intensive applications from various domains such as computer vision, machine learning, and artificial intelligence are expected to work seamlessly with *real-time* responses. However, the traditional way of offloading computation to the remote cloud often leads to unacceptable delay (e.g., hundreds of milliseconds [9]) and heavy backhaul usage. Due to its distributed computing environment, MEC can be leveraged to deploy applications and services as well as to store and process content in close proximity to mobile users. This would enable applications to be split into small tasks with some of the tasks performed at the local or regional clouds as long as the latency and accuracy are preserved.

In this case study, we envision a collaborative distributed computing framework where resource-constrained end-user devices outsource their computation to the upper-layer computing resources at the edge and cloud layers. Our framework extends the standard MEC originally formulated by ETSI, which only focuses on individual MEC entities and on the vertical interaction between end users and a single MEC node. Conversely, our proposed collaborative framework will bring many individual entities and infrastructures to collaborate with each other in a distributed system. In particular, our framework oversees a hierarchical architecture consisting of:

- *End user*, which implies both mobile and static end-user devices such as smartphones, sensors, and actuators
- *Edge nodes*, which are the MEC servers co-located with the BSs
- *Cloud node*, which is the traditional cloud-computing server in a remote data center

Our novel resource management framework lies at the intermediate edge layer and orchestrates both the *horizontal* collaboration at the end-user layer and the MEC layer as well as the *vertical* collaboration between end users, edge nodes, and cloud nodes. The framework will make dynamic decisions on "what" and "where" the tasks in an application should be executed based on the execution deadline, network conditions, and device battery capacity.

There have been a number of works in the mobile computing domain where data from the local device is uploaded to the cloud for further processing [10] or executed locally via approximate computing [11] to combat the problem of limited resources. In [12] we focused on the "extreme" scenario in which the resource pool was composed purely of proximal mobile devices. In contrast, MEC introduces a new stage of processing such that the edge nodes can analyze the data from nearby end users and notify the cloud node for further processing only when there is a significant change in data or accuracy of results. In addition, sending raw sensor values from end users to the edge layer can overwhelm the fronthaul links; hence, depending on the storage and compute capabilities of user devices and the network conditions, the MEC orchestrator can direct the end users to extract features from the raw data before sending to the edge nodes.

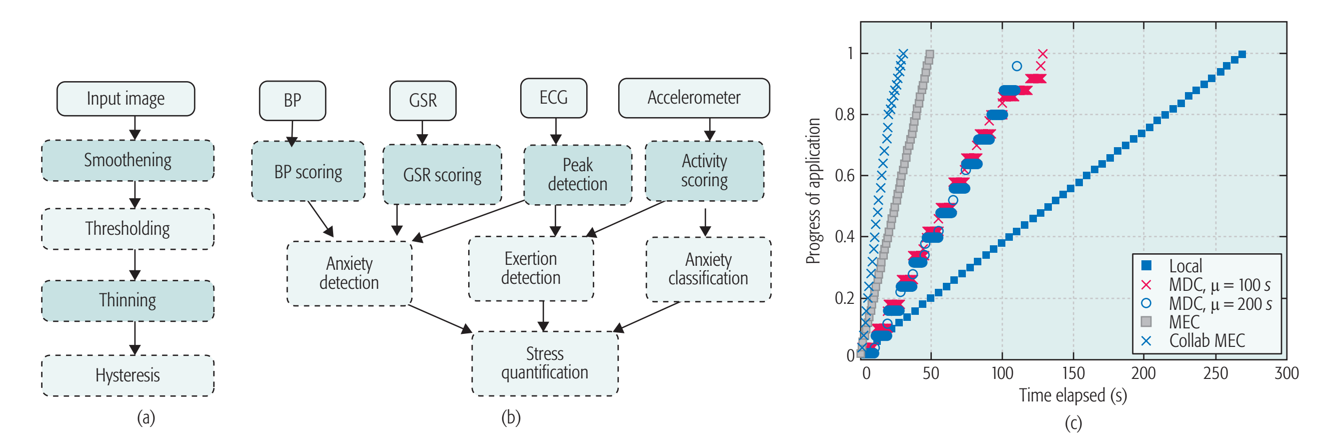

Figure 2. Block diagram showing tasks of different mobile applications in a) image processing domain (canny edge detection); b) ubiquitous health-care domain (stress quantification). The darker blocks in these applications represent the computationally intensive tasks of the applications that can be offloaded to the remote resources (edge and cloud); c) comparison of different startegies to execute computationally intensive mobile applications.

In Figs. 2a and 2b, we illustrate two mobile applications from different domains that are good candidates for being executed at the edge. The darker blocks in these applications represent the computation-intensive tasks of the applications that can be offloaded to the upper-level resources (edge and cloud). In Fig. 2c we compare the time taken for execution of the mobile application represented in Fig. 2a (canny edge detection) by using different strategies:

- Executing the application locally on the mobile device (Local)
- Distributing tasks to proximal mobile devices forming a mobile device cloud (MDC) [12]
- Offloading the tasks to a single MEC server (MEC)
- To two collaborating MEC servers (collab MEC), respectively.

For execution in an MDC we model the mobility patterns of devices in the proximity as a normal distribution with mean availability duration of devices varying with $\mu = \{100, 200\}$ s and $\sigma = 5$ s. We assume that the local mobile devices connect with the MEC server on a 1 Mb/s link. The mobile devices involved in the experiment include two Samsung Galaxy Tabs and four smartphones (two ZTE Avid N9120s and two Huawei M931s). For MEC servers we used two desktops with Intel Core i7 CPU at 3.40 GHz and 16 GB RAM. We execute the application in Fig. 2a by using input data from the Berkeley image segmentation and benchmark dataset. Resolution of each image is 481 × 321 pixels. A task consists of finding edges of 20 images from the dataset. For the current simulation, we use a round-robin technique for the MDC where all the devices are given equal tasks. Sophisticated task allocation algorithms can be run at the arbitrator to decide how many tasks to run at each service provider based on the computational capabilities of different service providers. After execution of the tasks, the service provider returns the task to the service requester. In Fig. 2c we see that the performance of execution on a single MEC server is significantly better than the execution on a local device and MDC. The gain in terms of execution time on using collaborative MEC over execution of the application on a single MEC server is around 40 percent.

The example above illustrates the benefit of the collaborative MEC framework in reducing execution time of the two image processing tasks. The extension of this strategy will greatly benefit the service requesters, which are health analytics providers in this case, as they see lower latency in the execution of the application as the MEC servers are at the BS rather than at the cloud. These service requesters require processing of large data, and the MEC servers expedite the processing time by dividing the processing between MEC servers (extracting features from the raw data) and cloud resources (running computation-intensive applications using extracted features as input data). This leads to faster availability results for the data analytics expert and also gives faster results to patients requesting results.

Currently, to present preliminary results we use a simple image processing application. However, we believe that a compute-intensive application (e.g., real-time activity detection with significant variations in execution time of tasks) or a data-intensive application (e.g., real-time face detection in a video with a large volume of input data) will require a powerful computing environment like ours to make dynamic decisions on what and where are the tasks to be executed based on real-time conditions, which will make application execution via collaborative MEC even more challenging.

Case Study II: Collaborative Video Caching and Processing

Mobile video streaming traffic is predicted to account for 72 percent of the overall mobile data traffic by 2019 [13], posing immense pressure on network operators. To overcome this challenge, edge caching has been recognized as a promising solution, by which popular videos are cached in the BSs or access points so that demands from users for the same content can be accommodated easily without duplicate transmission from remote servers. This approach helps substantially reduce backhaul usage and content access delay. While

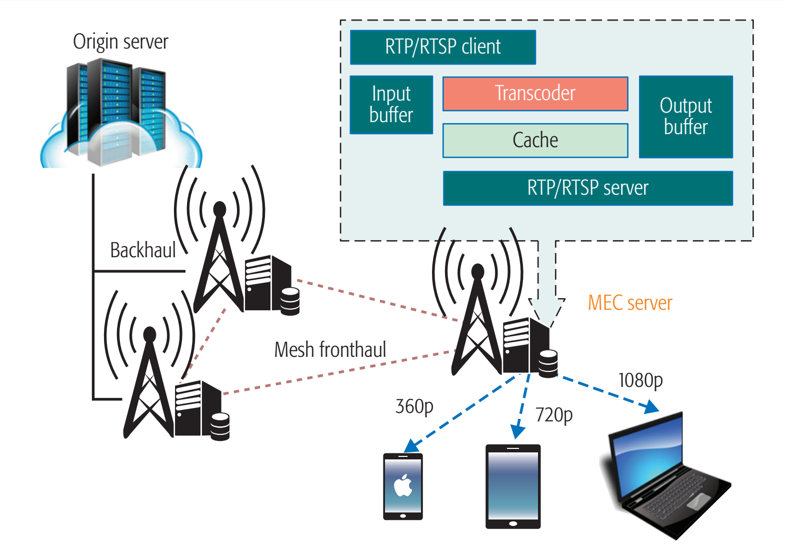

Figure 3. Illustration of collaborative video caching and processing framework deployed on an MEC network.

content caching and delivery techniques in wireless networks have been studied extensively (e.g., [14, references therein]), existing approaches rarely exploit the synergy of caching and computing at the cache nodes. Due to the limited cache storage at individual BSs, the cache hit rate is still moderate. Several solutions have considered collaborative caching, in which a video request can be served using not only the local BS's cache, but also the cached copy at neighboring BSs via the backhaul links [15].

With the emergence of MEC, it is possible to not only perform edge caching but also edge processing. Our approach will leverage edge processing capability to improve caching performance/efficiency. Such a joint caching and processing solution will trade off storage and computing resources with backhaul bandwidth consumption, which directly translates into sizable network cost saving. Due to the heterogeneity of users' processing capabilities and the variance of network connections, user preference and demand toward a specific video might be different. For example, users with highly capable devices and fast network connections usually prefer high-resolution videos, whereas users with low processing capabilities or low bandwidth connections may not enjoy high-quality videos because the delay is large and the video may not fit within the device's display. Leveraging such behavior, adaptive bit rate (ABR — https://en.wikipedia.org/wiki/Adaptive_bitrate_streaming) streaming techniques have been developed to improve the quality of delivered video on the Internet as well as wireless networks. Examples of such techniques include Apple HTTP Live Streaming (HLS), Microsoft Smooth Streaming, and Adobe Systems HTTP Dynamic Streaming. In ABR streaming, the quality of the streaming video is adjusted according to the user device's capabilities, network connection, and specific request. Existing video caching systems often treat each request for a video version equally and independently, without considering their transcoding relationship, resulting in moderate benefits.

In this case study, we exploit both ABR streaming and collaborative caching to improve the caching benefits beyond what can be achieved by traditional approaches. The proposed collaborative video caching and processing framework deployed on a MEC network [16] is illustrated in Fig. 3. Given the storage and computing capabilities, each MEC server acts as a cache server as well as a transcoding server. These servers collaborate with each other to not only provide the requested video but also transcode it to an appropriate variant. Each variant is a bit rate version of the video, and a higher bit rate version can be transcoded into lower bit rate ones. The potential benefits of this strategy are three-fold:

- The content origin servers need not generate all variants of the same video.
- Users with various capabilities and network conditions will receive videos that are suited for their capabilities, as content adaptation is more appropriately done at the network edge.
- Collaboration among the MEC servers enhances cache hit ratio and balance processing load in the network.

In our proposed joint collaborative caching and processing strategy, referred to as *CoPro-CoCache*, we distribute the most popular videos in the serving cell of each BS to the corresponding cache server of that BS until the cache storage is full. When a user requests a video that requires transcoding from a different version in the cache, the transcoding task is assigned to the MEC server having lower load, which could be the MEC server storing the original video version (data provider node) or the serving MEC server (delivery node). This helps balance the processing load in the network.

To illustrate the potential benefits of the proposed approach, we perform numerical simulation on a representative RAN with five BSs, each equipped with a MEC server that performs caching and transcoding. We assume a library of 1000 videos is available for download. The video popularity requested at each BS follows a Zipf distribution with parameter 0.8, that is, the probability that an incoming request is for the ith most popular video is proportional to $1/i^{0.8}$. In order to obtain a scenario where the same video can have different popularities at different locations, we randomly shuffle the distributions at different BSs. Video request arrival follows a Poisson distribution with same rate at each BS. In Figs. 4a and 4b we compare the performance of four caching strategies in terms of backhaul traffic reduction. It can be seen that utilizing processing capabilities significantly helps reduce the backhaul traffic load. In addition, our proposed *CoPro-CoCache* strategy explores the synergies of processing capabilities among the MEC servers, rendering additional performance gain. Figure 4c illustrates the processing resource utilization of the *CoPro-CoCache* scheme vs. different video request arrival rates and cache capacity. We observe that the processing utilization increases with arrival rate and moderate cache capacity; however, it decreases at high cache capacity. This is because with high cache capacity, we can store almost all the popular videos and their variants, and thus there are fewer requests requiring transcoding.

While choosing the optimal bit rate for video streaming can enhance instant download throughput, existing client-based bit rate selection may not be able to adapt fast enough to the rapidly varying conditions, leading to underutilization of radio resources and suboptimal user experience. A promising solution is to use a RAN analytic agent at the MEC server to inform the video server of the optimal bit rate to use given the radio conditions for a particular video request from an end user. Designing an efficient solution to address bit rate adaption with respect to channel conditions is still an open problem.

Case Study III: Two-Layer Interference Cancellation

Deploying more small cell BSs can improve spectral efficiency in cellular networks, however, making inter-cell interference become more prominent. To mitigate such interference, a promising approach is to employ coordinated multipoint (CoMP) transmission and reception techniques. In CoMP, a set of neighboring cells are divided into clusters; within each cluster, the BSs are connected to each other via a fixed backhaul processing unit (BPU) and exchange channel state information (CSI) as well as mobile station (MS) signals to cancel the intra-cluster interference. However, CoMP does not take into account the inter-cluster interference, resulting in moderate improvement in system capacity. Furthermore, the additional processing required for multi-site reception/transmission, CSI acquisition, and signaling exchanges among different BSs could add considerable delay and thus limit the cluster size in order to comply with the stringent delay requirement in 5G networks. In addition, applying CoMP for all users might be unnecessary as certain users, especially those at the cell centers, often have high levels of signal-to-interference-plus-noise ratio (SINR) and do not cause intense interference to the neighboring BSs.

To overcome the existing challenges of CoMP, and reduce the latency and bandwidth between the BSs and the BPU, we advocate a two-layer interference cancellation strategy for an uplink MEC-assisted RAN. In particular, based on the channel quality indicator (CQI) of each user, our solution identifies "where" to process its uplink signal so as to reduce complexity, delay, and bandwidth usage. In a MEC-assisted RAN, we have access to the computational processing at the BSs, and the signal demodulation of the cell center MSs can be done in local BSs (layer 1). This means that the system performance for cell center MSs relies on a simple single transmitter and receiver. On the other hand, since the SINRs of cell edge MSs are often low, their signals should be transmitted to the BPU (layer 2) for further processing. In this case, the BPU has access to all the celledge MSs from different cells and is able to improve their SINRs via coordinated processing.

As illustrated in Fig. 5, each red dotted circle indicates the interference region of the corresponding cell, which is defined as a region within which if MSs from other cells moved in, they could render "intense" interference at the BS serving the cell. Since MS #1 is a cell center MS and is outside the interference region of BSs #2 and #3, its interference at those BSs is low due to the high path loss; hence, there is no need to employ coordinated interference cancellation for MS #1, and thus its signal demodulation can be performed at the edge layer. Conversely, since MS #2 is a cell

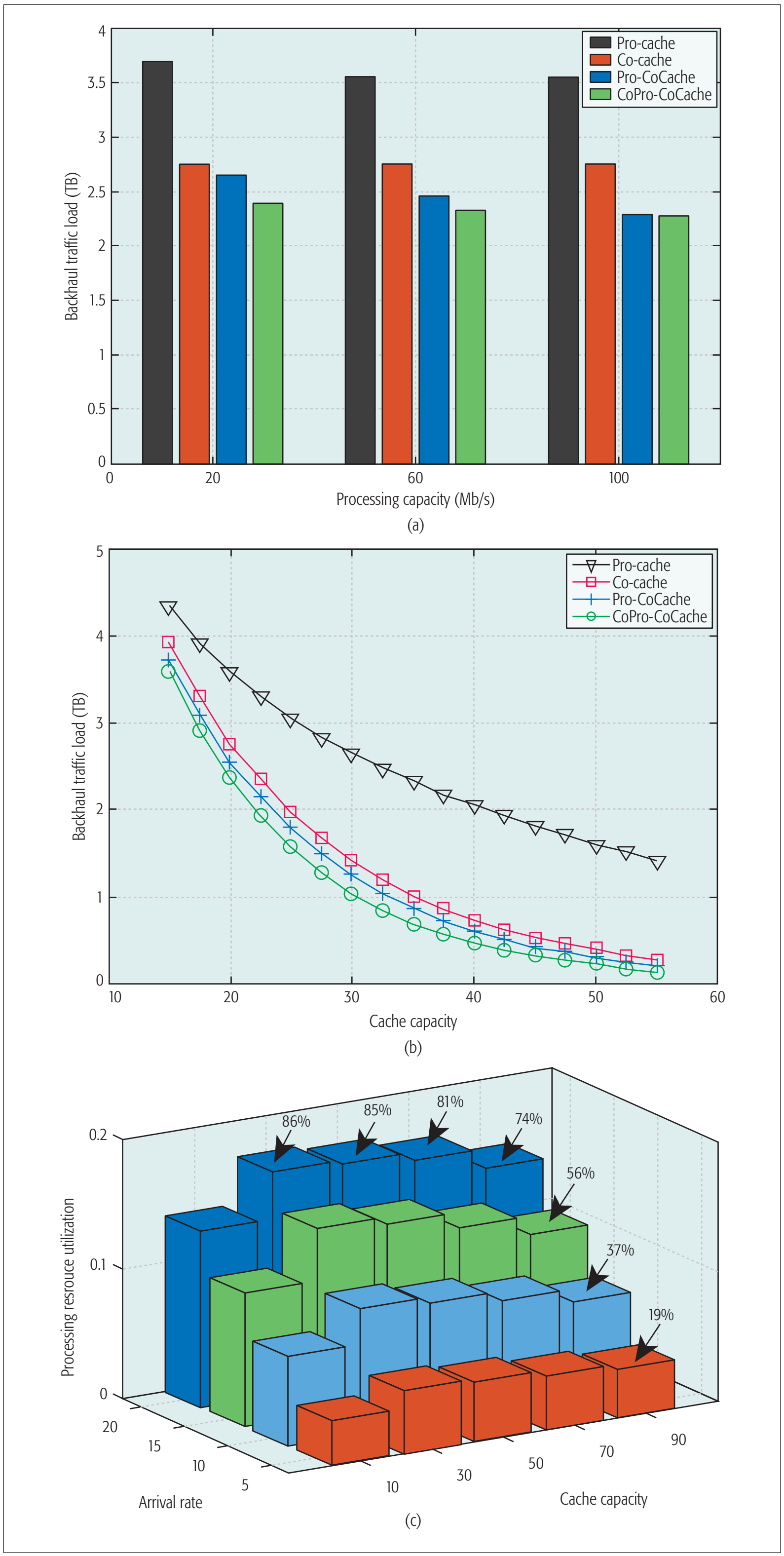

Figure 4. Considered caching strategies: Pro-Cache—non-collaborative caching with processing; Co-Cache—collaborative caching without processing; Pro-CoCache—collaborative caching with processing; and CoPro-CoCache—collaborative caching with collaborative processing (proposed). Video duration is set to 10 min, and each video has four variants with relative bit rates of 0.82, 0.67, 0.55, and 0.45 of the original video bit rate (2 Mb/s): a) backhaul traffic load vs. processing capacity (Mb/s) with cache capacity = 30 percent library size; b) backhaul traffic load vs. cache capacity; c) processing resource utilization vs. arrival rate (request/BS/min) and cache capacity. In b and c, we set processing capacity = 40 Mb/s.

Mobile-Edge Computing enables a capillary distribution of cloud computing capabilities to the edge of the radio access network. This emerging paradigm allows for 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.

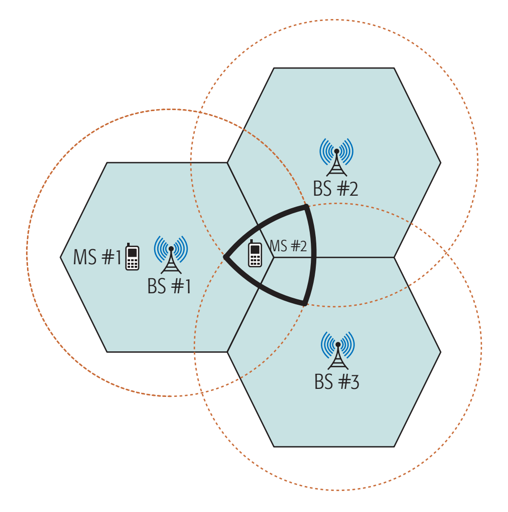

Figure 5. MSs #1 and #2 are located at cell center and cell edge regions, respectively. Since MS #1 is far from the neighboring BSs, signal demodulation can be performed at the edge (layer 1). However, MS #2 is located at the cell edge region, and its interference to the neighbouring BSs should be canceled at the upper layer (layer 2).

edge MS and is located in the interference region of BSs #2 and #3, there may be an intense interference from MS #2 to BSs #2 and #3; thus, coordinated interference cancellation at the upper layer is needed to cancel this interference, and the BS should transmit the raw data to the upper layer for further processing.

Challenges and Open Research Issues

The decentralization of cloud computing infrastructure to the edge brings various benefits that contribute to the 5G evolution, and at the same time introduces new challenges and open research issues, highlighted in the following.

Resource Management: The computing and storage resources in an individual MEC platform are expected to be limited and may be able to support a constrained number of applications with moderate needs for such resources. Currently, network providers often race for extensively standalone infrastructures to keep up with the demand while struggling with lower return on investment. An alternative approach such as MEC as a service may need to be considered, whereby operators' resources can be opened up for interested service providers to request or relinquish based on service demand.

Interoperability: MEC infrastructures owned by different network providers should be able to collaborate with each other as well. This necessitates the specification of common collaboration protocols, also allowing for service providers to access network and context information regardless of their deployment place.

Service Discovery: Exploiting the synergies of distributed resources and various entities, as envisioned in our mobile edge orchestration framework, requires discovery mechanisms to find appropriate nodes that can be leveraged in a decentralized setup. Automatic monitoring of the heterogeneous resources and accurate synchronization across multiple devices are also of great importance.

Mobility Support: In a small cell network, the range of each individual cell is limited. Mobility support becomes more important, and a solution for fast process migration may become necessary.

Fairness: Ensuring fair resource sharing and load balancing is also an essential problem. There is potential that a small number of nodes could carry the burden of processing, while a large number of nodes would contribute little to the efficiency of the distributed network.

Security: Security issues might hinder the success of the MEC paradigm if not carefully considered. 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 (e.g., content providers want to know users' preferences and mobility patterns to proactively cache their contents, as discussed in case study II), there is a great challenge to the development of privacy protection mechanisms that can efficiently protect users' location and service usage.

Conclusions

Mobile edge computing enables a capillary distribution of cloud computing capabilities to the edge of the radio access network. This emerging paradigm allows for execution of delay-sensitive and context-aware applications in close proximity to end users while alleviating backhaul utilization and computation at the core network. This article proposes to explore the synergies among connected entities in the MEC network to form a heterogeneous resource pool. We present three representative use cases to illustrate the benefits of MEC collaboration in 5G networks. Technical challenges and open research issues are highlighted to give a glimpse of the development and standardization roadmap of the mobile edge ecosystem.

Acknowledgment

This work was supported in part by the National Science Foundation (NSF) under Grant No. CNS-1319945.

References

[1] F. Bonomi *et al.*, "Fog Computing and Its Role in the Internet of Things," *Proc. 1st ACM Wksp. Mobile Cloud Computing*, 2012, pp. 13–16.

[2] Intel and Nokia Siemens Networks, "Increasing Mobile Operators' Value Proposition with Edge Computing," technical brief, 2013.

[3] IBM, "Smarter Wireless Networks; Add Intelligence to the Mobile Network Edge," Thought Leadership white paper, 2013.

[4] Saguna and Intel, "Using Mobile Edge Computing to Improve Mobile Network Performance and Profitability," white paper, 2016.

[5] Y. C. Hu *et al.*, "Mobile Edge Computing — A Key Technology Towards 5G," ETSI white paper, vol. 11, 2015.

[6] S. Sardellitti, G. Scutari, and S. Barbarossa, "Joint Optimization of Radio and Computational Resources for Multicell Mobile-Edge Computing," *IEEE Trans. Signal Info. Processing over Networks*, vol. 1, no. 2, 2011, pp. 89–1035.

[7] M. Chiang and T. Zhang, "Fog and IoT: An Overview of Research Opportunities," *IEEE Internet of Things J.*, vol. 3, no. 6, 2016, pp. 854–64.

[8] D. Pompili, A. Hajisami, and T. X. Tran, "Elastic Resource Utilization Framework for High Capacity and Energy Efficiency in Cloud RAN," *IEEE Commun. Mag.*, vol. 54, no. 1, Jan. 2016, pp. 26–32.

[9] E. Cuervo *et al.*, "MAUI: Making Smartphones Last Longer with Code Offload," *Proc. ACM Int'l. Conf. Mobile Systems Applications Services*, 2010, pp. 49–62.

[10] M. S. Gordon *et al.*, "COMET: Code Offload by Migrating Execution Transparently," *Proc. USENIX Conf. Operating Systems Design and Implementation*, Oct. 2012, pp. 93–106.

[11] P. Pandey and D. Pompili, "Exploiting the Untapped Potential of Mobile Distributed Computing via Approximation," *Pervasive and Mobile Computing*, 2017.

[12] H. Viswanathan, P. Pandey, and D. Pompili, "Maestro: Orchestrating Concurrent Application Workflows in Mobile Device Clouds," *Proc. Wksp. Distrib. Adaptive Systems, Int'l. Conf. Autonomic Computing*, July 2016, pp. 257–62.

[13] Global Mobile Data Traffic Forecast, Update 2014–2019, White Paper c11-520862, Cisco Visual Networking Index.

[14] E. Bastug, M. Bennis, and M. Debbah, "Living on the Edge: The Role of Proactive Caching in 5G Wireless Networks," *IEEE Commun. Mag.*, vol. 52, no. 8, Aug. 2014, pp. 82–89.

[15] T. X. Tran and D. Pompili, "Octopus: A Cooperative Hierarchical Caching Strategy for Cloud Radio Access Networks," *Proc. IEEE Int'l. Conf. Mobile Ad Hoc Sensor Systems*, Oct. 2016, pp. 154–62.

[16] T. X. Tran *et al.*, "Collaborative Multi-bitrate Video Caching and Processing in Mobile-Edge Computing Networks," *Proc. IEEE Annual Conf. Wireless On-demand Network Systems Services*, Feb. 2017.

Biographies

Tuyen X. Tran (tuyen.tran@cac.rutgers.edu) is working toward his Ph.D. degree in electrical and computer engineering (ECE) at Rutgers University under the guidance of Dr. Pompili. He received his M.Sc. degree in ECE from the University of Akron, Ohio, in 2013, and his B.Eng. degree (Honors Program) in electronics and telecommunications from Hanoi University of Technology, Vietnam, in 2011. His research interests are in the application of optimization, statistics, and game theory to wireless communications and cloud computing.

Abolfazl Hajisami (hajisamik@cac.rutgers.edu) started his Ph.D. program in ECE at Rutgers University in 2012. Currently, he is pursuing research in the fields of C-RAN, cellular networking, and mobility management under the guidance of Dr. Pompili. Previously, he received his M.S. and B.S. from Sharif University of Technology and Shahid Beheshti University, Tehran, Iran, in 2010 and 2008, respectively. His research interests are wireless communications, cloud radio access networks, statistical signal processing, and image processing.

Parul Pandey (parul_pandey@cac.rutgers.edu) is a Ph.D. candidate in the Department of ECE at Rutgers University. She is currently working on mobile and approximate computing, cloud-assisted robotics, and underwater acoustic communications under the guidance of Dr. Pompili as a member of the Cyber-Physical Systems Laboratory (CPS-Lab). Previously, she received her B.S. degree in electronics and communication engineering from Indira Gandhi Institute of Technology, Delhi, India, and her M.S. degree in ECE from the Univeristy of Utah in 2008 and 2011, respectively.

Dario Pompili [SM] (pompilig@cac.rutgers.edu) is an associate professor with the Department of ECE at Rutgers University, where he directs the CPS-Lab. He received his Ph.D. in ECE from the Georgia Institute of Technology in 2007. He previously received his Laurea (combined B.S. and M.S.) and doctorate degrees in telecommunications and systems engineering from the University of Rome "La Sapienza," Italy, in 2001 and 2004, respectively. He is a recipient of the NSF CAREER '11, ONR Young Investigator Program '12, and DARPA Young Faculty '12 awards. He is a Senior Member of the ACM.