# Fog Computing Enabled Future Mobile Communication Networks: A Convergence of Communication and Computing

Yiqing Zhou, Lin Tian, Ling Liu, and Yanli Qi

The authors provide an overview on FogMNW, including network architecture, system capacity and resource management. First they analyze the heterogeneity of FogMNW with both advanced communication techniques and fog computing. Then they propose a heterogeneous communication and hierarchical fog computing network architecture.

## Abstract

The convergence of communication and computing (COM$^2$P) has been taken as a promising solution for the sustainable development of mobile communication systems. The introduction of fog computing in future mobile networks makes COM$^2$P possible. This article provides an overview on fog computing enabled mobile communication networks (FogMNW), including network architecture, system capacity and resource management. First, this article analyzes the heterogeneity of FogMNW with both advanced communication techniques and fog computing. Then a heterogeneous communication and hierarchical fog computing network architecture is proposed. With both communication and computing resources, FogMNW is enabled to achieve much higher capacity than conventional communication networks. This has been well demonstrated by the coded multicast scheme. Furthermore, a systematic management of communication and computing resources is necessary for FogMNW. By exploiting the communication load diversity in N cells, a communication load aware CLA scheme can achieve much higher computing resource efficiency than comparing schemes. The performance gap increases with N, and CLA can improve efficiency by more than 100 percent when there are 14 cells.

## Introduction

In the past few decades, wireless communication technologies have been developed rapidly. Starting from pure voice communication services provided by the first and second generations of mobile communication systems (1G and 2G), the current 4G supports various wireless multimedia services such as Internet surfing and video streaming. The future 5G will further extend the serving capability to three main application scenarios, that is, enhanced mobile broadband communication (eMBB), massive machine type communication (mMTC), and ultra-reliable and low latency communication (URLLC). However, radio communication resources are limited. How to satisfy the ever-increasing demands for mobile communication systems becomes a major concern of the industry and academia.

A convergence of communication and computing (COM$^2$P) may be a promising solution [1]. One point to be clarified is that COM$^2$P is not a new idea and has drawn a lot of attention. In mobile communication systems, COM$^2$P can be carried out in different levels, that is, in a single communication technique level, in a single communication device level, in a network architecture level, and in a system level. Communication and computing have been well converged in a single communication technique level (e.g., channel coding) and in a single device level (e.g., stored program control exchanger). However, for future mobile communication systems, COM$^2$P in these two levels are not sufficient to address the sustainable development problem. A systematic convergence is needed, with which problems should be investigated in a systematic way, by abstracting and inducing features of each level of the system, and analyzing and optimizing the overall system performance.

Among all the recent advances in this direction, fog computing is a basic enabling technique for future mobile communication networks with COM$^2$P. First proposed by Cisco in 2012 [2], fog computing is a non-trivial extension of cloud computing to the edge of networks. By providing computing, storage and communication services in the proximity of end users, fog computing aims to provide a new platform to meet the requirements of Internet of Things, such as low latency and mobility support. It is a highly virtualized platform enabling new applications and services. Meanwhile, a similar idea, mobile edge computing (MEC), was introduced in 2014 by ETSI [3]. MEC provides cloud computing capabilities within the radio access network (RAN). Although fog computing and MEC are proposed by different companies/organizations with different aims, their main idea, advances in performance, realization and benefits to the industry are the same. We would argue that there are no fundamental differences between fog computing and MEC, and the terminologies can be used interchangeably in this article.

Various research has been carried out on fog computing enabled future mobile communication networks (FogMNW). This article focuses on the impact of COM$^2$P on FogMNW in three aspects, that is, *network architectures, system capacity and resource management*. First, network architectures of FogMNW have drawn a lot of attention in both industry and academia. A fog-based service

This work was supported by the National Natural Science Foundation of China under Grant 61571425.

Digital Object Identifier: 10.1109/MCOM.2019.1800235

*The authors are with Key Laboratory of Mobile Computing and Pervasive Devices, Institute of Computing Technology, University of Chinese Academy of Sciences.*

0163-6804/19/$25.00 © 2019 IEEE

enablement architecture was proposed in [4] to underpin a system with multi-level fog computing. In addition, a high-level system architecture has been presented in [5], which could exploit edge computing and communication resources to enable the 5G ecosystem. Moreover, future mobile networks with centralized architectures like CRAN will also be quite suitable to support fog computing, since their central units (CUs) have plenty of computing resources. Hence, MEC has been standardized in 3GPP as an essential part of 5G [6]. IEEE also released the Open Fog reference architecture (OpenFog RA) [7]. All these advances in network architectures render future mobile communication networks multi-level fog computing capabilities and make COM$^2$P possible. This article will analyze different levels of communication and computing capabilities in FogMNW and propose a heterogeneous communication and hierarchical computing network architecture.

Given FogMNW, a fundamental problem is to get the capacity of the system with communication and computing resources. Research has been carried out to reveal the impact of three resources, that is, communication, caching and computing, on system capacity [1]. Based on the definition of the three resource vectors, the capacity gains of caching and computing over a normalized communication-only system could be derived. These results can be verified by a typical COM$^2$P scheme, *coded multicast* [8], which was designed for file delivery in mobile networks. It could collaboratively exploit the communication and computing functionalities provided by the mobile network with fog computing. By designing a smart storage scheme, performing logical computing among information streams, and creating multicast opportunities, coded multicast can considerably reduce transmission bandwidth, thus improve system capacity equivalently.

Moreover, in FogMNW, a systematic management of communication and computing resources is necessary. One motivation is that the introduction of fog computing brings new mobile services demanding high computing capability. The computing task can be split and offloaded to different fog nodes in mobile networks, which should be jointly optimized with communication resources such as bandwidth and power [9-11]. Different from previous works, [12] aims to improve the computing resource efficiency of CU (a fog node in FogMNW) by exploiting the features of communication load in mobile networks. A communication load aware computing resource allocation scheme (CLA) was proposed. Based on the 24-hour network load statistics provided by China Mobile, the performance of CLA has been verified via simulations. By exploiting the communication load diversity in N cells, CLA achieves much higher computing resource efficiency than comparable schemes. The performance gap increases with N, and CLA can improve the efficiency by more than 100 percent when there are 14 cells.

Although existing research has partially verified the potential of COM$^2$P to enhance the performance of FogMNW, there are still many open problems regarding how COM$^2$P would support the sustainable development of mobile communication systems, and they will be discussed in the section on open issues.

## Mobile Communication Network Architectures with Fog Computing

The development of future mobile communication network architectures can be perceived from two aspects. One is from the communication point of view. Mobile communication networks will evolve into a heterogeneous network with both distributed and centralized architectures, satisfying different coverage and transmission requirements of various services. The other is from the computing point of view. With the introduction of fog computing at different locations, future mobile communication networks will also provide multi-level computing capabilities. *It is the introduction of fog computing in mobile communications that makes COM$^2$P possible*.

### Heterogeneous Mobile Communications

Mobile communication systems were originally designed to provide one main service, that is, voice communications, and networks were constructed by pseudo-regularly distributed macro cells with uniform coverage. With the emergence of mobile multimedia services, various transmission and coverage demands occur. The current mobile network is already a heterogeneous one, constructed by macro cells, micro cells, pico cells, femto cells, and so on, providing reduced coverage with increased broadband communication capabilities in sequence. The future mobile network will be even more heterogeneous with the possible deployment of new technologies like millimeter-wave (mmWave) communications, ultra dense cellular networks (UDNs) [13] and centralized network architectures.

*MmWave communications* were proposed to exploit the undeveloped frequency spectrum in the mmWave band (30–300 GHz). Although spectrum is abundant in this band, the radio signals suffer high atmospheric attenuation and rain fade, and may be blocked by buildings and trees. MmWave may be possibly employed to provide ultra-broadband data transmission for a small area, but it is not suitable for control signal transmission which needs uniform coverage for all users over a wide area. Hence, data and control signals may be transmitted separately by different cells using different spectrums. This demands user and control plane separation (CUPS) in the core network, which facilitates the deployment of fog computing servers in future mobile networks, as shown in the next subsection.

By deploying a large number of irregularly deployed small cells, *UDNs* are widely taken as a promising way to improve system capacity. This deployment breaks the classic paradigm of controlled cellular planning with pseudo-regular placement of cells. It should be modeled by new network topologies such as random topology, where base stations (BSs) are assumed to be located randomly and independently in the space [14]. With the increased number of deployed cells in UDN, the mobile stations (MSs) are getting closer and closer to BSs, which creates favorable wireless transmission conditions and makes broadband communications easier.

Mobile communication systems were originally designed to provide one main service, that is, voice communications, and networks were constructed by pseudo-regularly distributed macro cells with uniform coverage. With the emergence of mobile multimedia services, various transmission and coverage demands occur.

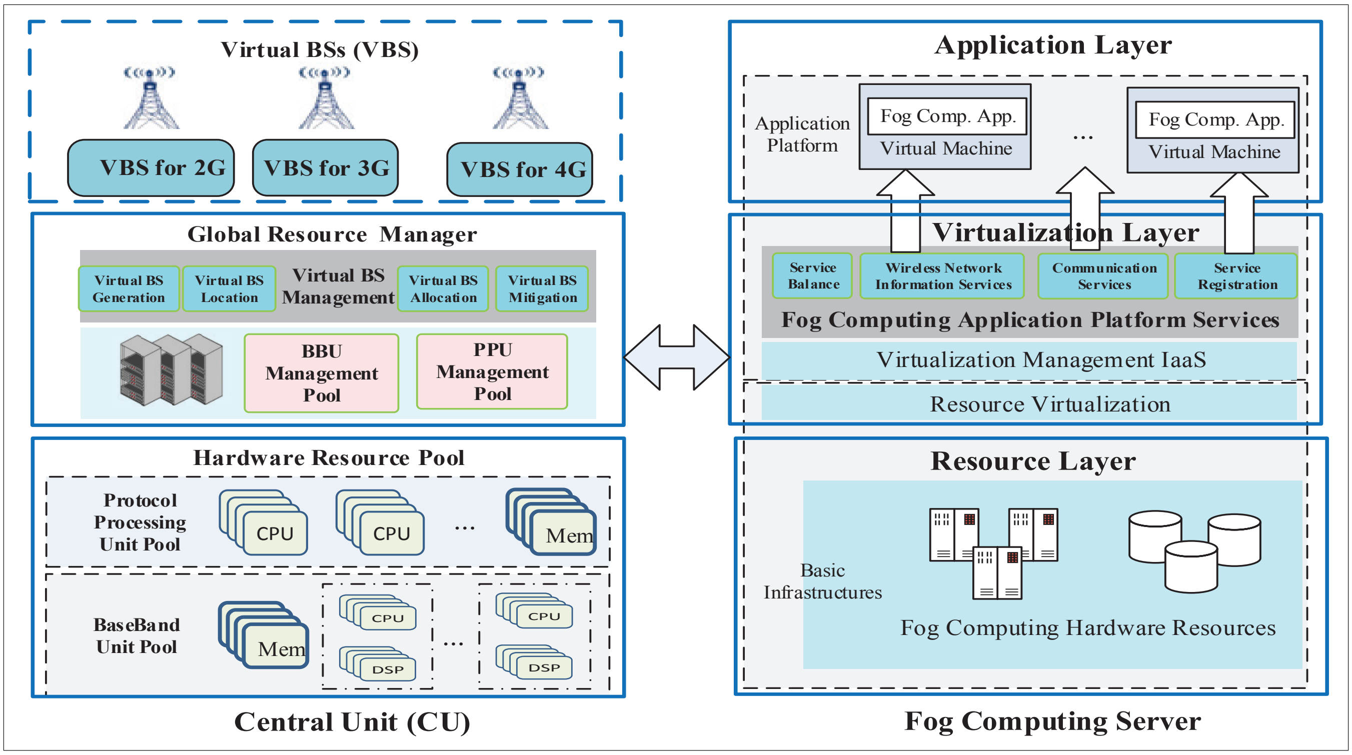

Figure 1. Block diagram comparison of CU and fog computing server.

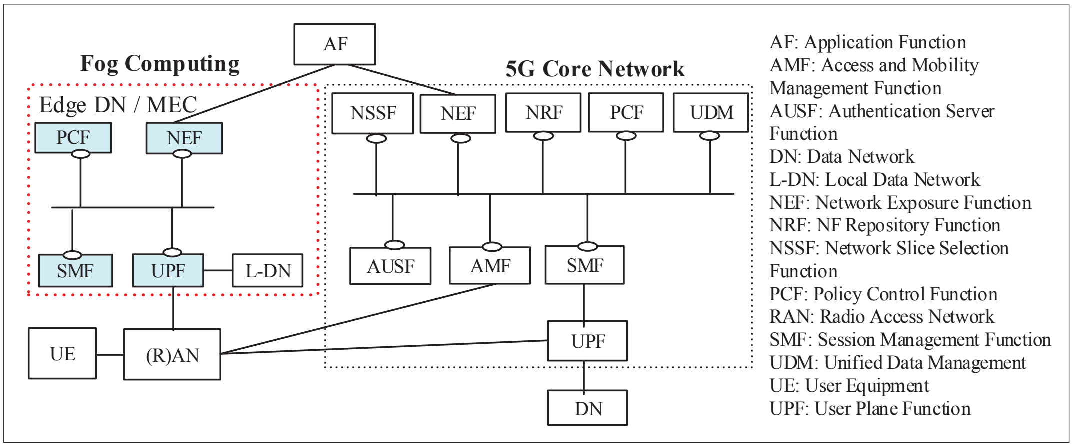
AF
Fog Computing
5G Core Network
Edge DN / MEC
PCF
NEF
SMF
UPF
L-DN
NSSF
NEF
NRF
PCF
UDM
AUSF
AMF
SMF
UE
(R)AN
UPF
DN
AF: Application Function
AMF: Access and Mobility Management Function
AUSF: Authentication Server Function
DN: Data Network
L-DN: Local Data Network
NEF: Network Exposure Function
NRF: NF Repository Function
NSSF: Network Slice Selection Function
PCF: Policy Control Function
RAN: Radio Access Network
SMF: Session Management Function
UDM: Unified Data Management
UE: User Equipment
UPF: User Plane Function

Figure 2. 5G core network architecture with fog computing.

The deployment of mmWave and UDN would result in a large number of small cells which bring MSs close to BSs. *Due to the small coverage and small number of serving MSs, the BSs in these cells must be with limited size and power. If fog computing is deployed within these BSs in FogMNW, their computing capabilities are also limited.* On the other hand, even though these BSs are of small size, due to the huge volume, it would be difficult to find sites for them, and their maintenance and upgrading would be a big problem. One promising solution is to use centralized network architectures.

Different from traditional distributed architectures where signal processing and transmission are coupled at every BS, centralized architectures decouple the signal processing part of each BS and concentrate them in a CU, leaving distributed antenna units (DUs) at BS sites for signal transmission and receiving. It would be easier to find sites for DUs and carry out maintenance and upgrading in one CU. An additional advantage is the statistical multiplexing gain achieved via the resource pooling in CUs. Moreover, with plenty of heterogeneous computing resources like FPGA, DSP, CPU and memory, the CU can be flexibly managed to realize different BS functions using virtualization techniques.

The block diagram of an example CU is shown in Fig. 1, which is composed of a hardware resource pool, global resource manager and virtual BSs (VBSs). The hardware resource pool includes a base band unit (BBU) pool and a protocol processing unit (PPU) pool, supporting various processing in communications. On top of the hardware resource pool, there are baseband and protocol resource management softwares, responsible for resource allocation and BS virtualization. The CU's structure is similar to that of the fog computing server shown in the same figure, which is composed of the resource layer, virtualization layer and application layer. Hence, *it*

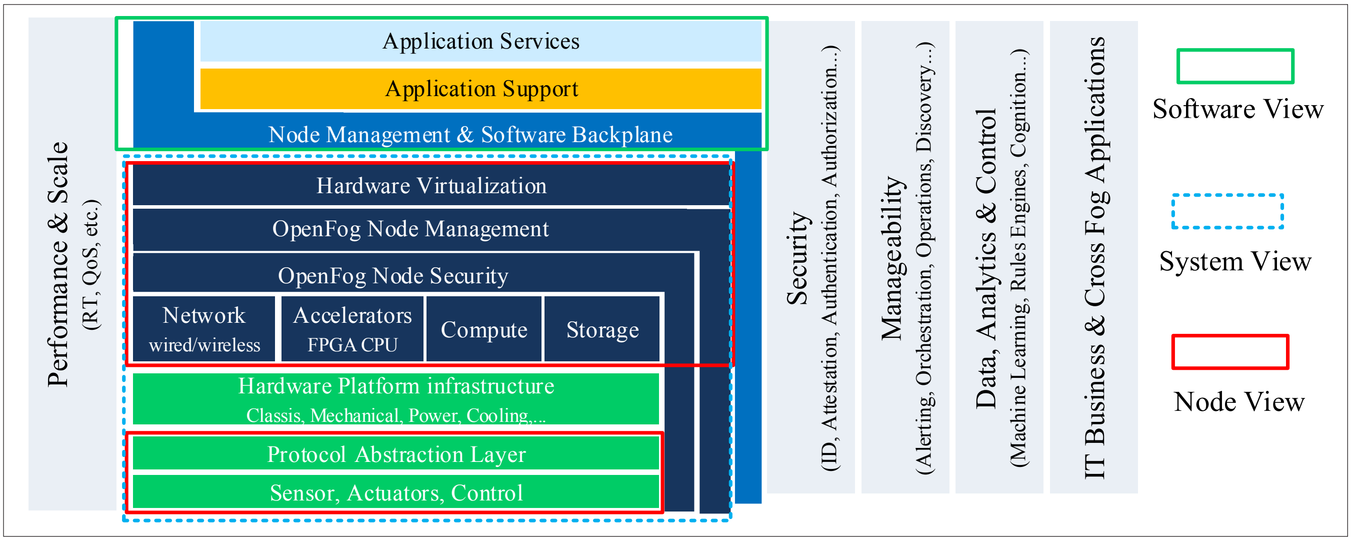
Application Services
Application Support
Node Management & Software Backplane
Hardware Virtualization
OpenFog Node Management
OpenFog Node Security
Network
wired/wireless
Accelerators
FPGA CPU
Compute
Storage
Hardware Platform infrastructure
Classis, Mechanical, Power, Cooling,...
Protocol Abstraction Layer
Sensor, Actuators, Control
Performance & Scale
(RT, QoS, etc.)
Security
(ID, Attestation, Authentication, Authorization...)
Manageability
(Alerting, Orchestration, Operations, Discovery...)
Data, Analytics & Control
(Machine Learning, Rules Engines, Cognition...)
IT Business & Cross Fog Applications
Software View
System View
Node View

**Figure 3.** Architecture description with perspectives in OpenFog RA.

*would be natural for CUs to support fog computing, and CUs can also be taken as a type of special fog computing server.*

Although the centralized architecture could effectively reduce high network maintenance costs, it has one main disadvantage, that is, a huge bandwidth requirement on the fronthaul. Since the signals from all MSs in one cell should be delivered over the fronthaul, the bandwidth of the fronthaul could be as high as Tb/s, which is difficult to support even with the most advanced fiber techniques. Thus, next generation fronthaul interface (NGFI) has been proposed [15], aiming to alleviate the bandwidth requirement on fronthauls. It suggests moving some baseband processing to DUs and making the transmission over fronthauls proportional to the system's communication load. Currently, one focus of 5G network architecture standardization is how to split processing tasks between the CU and DU. Obviously, *the more tasks allocated in CUs, the more computing resources are possessed by CUs and the more powerful fog computing can be supported by the CUs.*

## Standardization of Fog Computing in 5G and IEEE

For mobile communication networks, fog computing can bring cloud computing capability to the edge of the network. The stringent requirement on backhaul bandwidth can be alleviated and the service delay can be significantly reduced from 100 ms to 1 ms, considering the 5G air interface. Thus, fog computing enhances the serving capability of mobile networks to support advanced applications such as augmented reality/virtual reality (AR/VR) [3]. Moreover, with computing and storage capability introduced by fog computing, future FogMNW like 5G is enabled to adopt new techniques. For example, wireless big data can be stored in nodes and intelligent algorithms like deep learning could be supported. Thus, new techniques based on big data and deep learning can be developed for 5G, such as spectrum sensing and network self organizing. The deployment of fog computing in future mobile communication networks has been widely accepted. Standardization has been carried out by ETSI [3], 3GPP [6] and IEEE [7].

The 5G core network architecture standardized by 3GPP is shown in Fig. 2 [6]. *To facilitate the deployment of fog computing in future mobile networks, new network functionalities have been defined*. For example, network exposure function (NEF) has been added to open capabilities like monitoring and charging to third parties. In addition, local data network (L-DN) is defined to enable local content visiting at fog computing servers. Moreover, CUPS will be employed by 5G, so fog computing servers can flexibly and adaptively deploy user plane function (UPF) and some control plane function like policy control function (PCF) and session management function (SMF), according to various requirements. It can be seen that *fog computing or MEC has been taken as an essential part of the 5G core network.*

Recently, IEEE adopted the OpenFog RA released by the OpenFog Consortium as an official standard [7]. The core of this RA is nine pillars (i.e., technical principles) which provide guidance to define the reference architecture, such as scalability and openness. The OpenFog RA provides an in-depth look at the reference architecture through description, views, viewpoints and perspectives. As shown in Fig. 3, the architecture description presents the composite requirements (i.e., views) for different stakeholders in the fog computing continuum. Moreover, five cross-cutting perspectives of the fog computing architecture are shown, such as performance and scale perspective. The OpenFog RA is the first step in creating industry standards for fog computing. Other topics like detailed guidance will be next established and developed.

## Future Heterogeneous Communication and Hierarchical Fog Computing Network Architecture

Fog computing servers can be deployed at various locations in mobile communication networks, such as BSs and cell aggregation sites [3]. Moreover, since MSs are getting more and more powerful in computing, they can also be taken as fog computing nodes. Combined with device-to-device (D2D) communications, neighboring MSs can work cooperatively to complete a computing task. Therefore, considering the fog computing

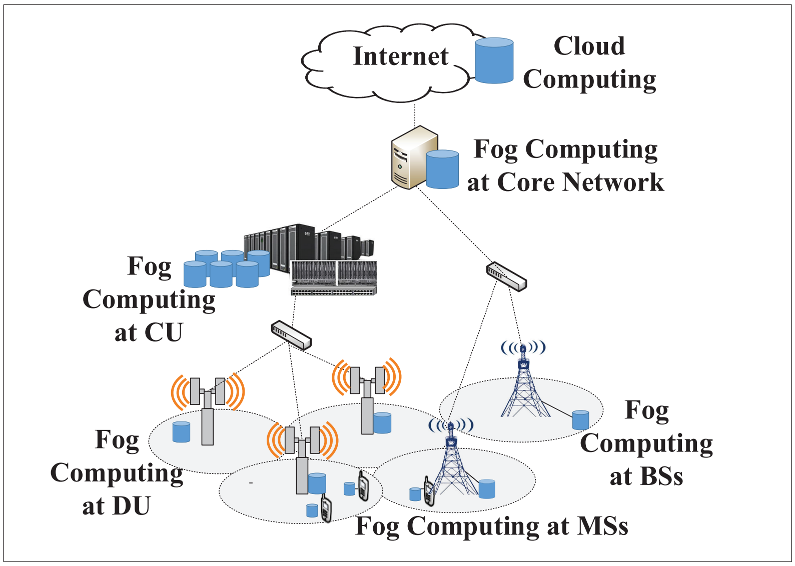

Figure 4. Future heterogeneous communication and hierarchical fog computing network architecture.

capabilities located at different locations, including MSs, BSs and CUs, future mobile networks are not only heterogeneous communication networks, but also hierarchical fog computing networks. As shown in Fig. 4, *in such a heterogeneous communication network with hierarchical fog computing, the closer the fog computing node/server is to a MS, the faster service response can be achieved, but the less powerful computing capability it can provide*. So there is a performance trade-off for different levels of fog computing in mobile networks in terms of computing capability and delay.

Compared to existing fog network architectures [4-5], the advantage of this architecture is to show the convergence of mobile communication networks and fog computing networks. It clearly demonstrates the heterogeneity of FogMNW and shows the relationship between computing capabilities, communication capabilities and performances like delay. Note that communication and computing resources available at different nodes are not only heterogeneous, but also dynamic, since applications running on nodes change from time to time. So one big challenge for FogMNW is how to maximize its capability by collaboratively using heterogeneous and dynamic communication and computing resources at different nodes, considering performance requirements. This architecture also provides a useful reference for us to set up appropriate models for computing and communication resources in FogMNW, and enable us to develop novel COM$^2$P schemes exploiting the heterogeneity of FogMNW. Yet it is not closely related to applications and services like [4], and it may not be the best to demonstrate the importance of COM$^2$P to the 5G ecosystem.

## Communication Capacity Gains from Fog Computing

### System Capacity with Communication and Computing Resources

Given FogMNW, a fundamental question is if there exists a capacity limit for the system with both communication and computing resources, similar to conventional communication systems. Some effort has been made in this direction.

Taking communication, computing and caching as three primary resources of a mobile system, their impact on the system capacity were investigated in [1]. To analytically characterize mobile systems with the three resources and quantify their impact, the metrics and mode of operations were defined for communication, computing and caching. In the proposed framework, the communication vector was defined as the capability to deliver information streams over wireless channels without altering. It is measured by the data rate. Next, the caching vector was defined as the capability to store information at network nodes. It is measured by the memory size. Caching can exploit the feature of certain services, for example, they can be delivered over longer time, thus enhancing the system's ability. Finally, *the computing vector was defined as the capability to perform logic operations across information streams*, that is, alter the information bits such that additional logic operations must be carried out at the receiver to recover the original information. It is measured by the number of information streams involved in the operations. This kind of definition differentiated the computing resources from those used in computer science measured by flops, since the focus is not the amount of computations, but how the computing affects the communication performance limits.

Based on these definitions and normalizing the value of the communication vector to one, the capacity gains obtained by caching and computing were analyzed. Taking the coded multicast scheme as an example, it was shown that *the capacity to deliver information was proportional to the values of the cache and computing vector.*

### Coded Multicast

A typical COM$^2$P scheme for FogMNW is *coded multicast* [8]. First, MSs can be taken as fog computing nodes, as shown in Fig. 4, so existing files can be delivered to and stored at MSs when there are abundant communication resources. When one MS needs a certain file, it is possible that the file is among the locally stored files. However, this hit ratio is usually negligibly low since the memory size of a MS is much smaller than that of a BS. Hence, coded multicast was proposed to exploit both the computing and storage capabilities of fog computing nodes. *The main idea is to design a file storage scheme among multiple MSs, such that when these MSs are requesting files, some combined files can be multicast to them, and each MS can recover the needed information via computing*. Due to the logical operation (i.e., XOR computing) used to combine files and the spectrum efficient multicast transmission, the data delivered over wireless channels is reduced considerably and thus the system capacity is increased equivalently.

For example, assume that MS$_1$ and MS$_2$ have the same memory size of *M* bits. There are two files of *M* bits at the BS, denoted by F$_1$ and F$_2$. Coded multiplexing would suggest MS$_1$ storing the first half of F$_1$ (F$_{1,1}$) and F$_2$ (F$_{2,1}$), while MS$_2$ stores the second half of F$_1$ (F$_{1,2}$) and F$_2$ (F$_{2,2}$). When MS$_1$ and MS$_2$ request F$_1$ and F$_2$ from the BS, respectively, the BS could XOR F$_{1,2}$ with F$_{2,1}$ and multicast the combined file F$_{3,\text{half}}$ to MS$_1$ and

$MS_2$. Since $MS_1$ already stores $F_{2,1}$, it can recover $F_{1,2}$ by XORing $F_{3,half}$ with $F_{2,1}$. Similarly, $MS_2$ can obtain $F_{2,1}$. Hence, using the storage and XOR-computing capabilities provided by fog computing nodes, only $M/2$ bits are transmitted when $MS_1$ and $MS_2$ request $F_1$ and $F_2$, respectively. For comparison, in traditional communication systems without fog computing, 2 $M$ bits should be delivered. Therefore, a well-designed COM²P scheme has the potential to improve system capacity significantly.

## Systematic Management of Communication and Computing Resources

In FogMNW, system performance depends on not only conventional communication resources, but also computing resources.

### Joint Optimization of Communication and Computing Resources

To exploit the hierarchical fog computing capabilities in FogMNW, one main problem is how to split the computing task of a mobile service to multiple fog computing nodes at different locations. *Different from traditional computing offloading which only concerns the amount of computations, the computing offloading in mobile systems must be jointly optimized with computing resources.* For example, AR/VR is widely taken as a killer application for FogMNW [3]. Yet the computing capability of MSs is insufficient to support AR/VR. Thus, the computing task must be split into several parts and offloaded to multiple fog computing nodes.

Compared with local computing, offloading subtasks from MSs to networks needs transmission bandwidth and power and will introduce extra delay. Hence, joint communication and computing resource optimization has been widely investigated, considering application requirements such as delay constraints, energy consumption of communication and computing, and wireless transmission bandwidth. For instance, in [11], the authors studied the optimization of computing offloading and wireless communication resource allocation in future 5G heterogeneous networks. The target is to minimize the total power consumption for information delivering and computation, constrained by the service delay. A comprehensive survey on joint optimization of communication and computing resources is provided in [9].

### Communication Load Aware Computing Resource Allocation

As illustrated above, the CU plays an important role in FogMNW, since it can be both a communication node and a fog computing node. Different from joint optimization, the authors in [12] focused on the computing efficiency of CUs in FogMNW, and tried to enhance the efficiency by exploiting the features of mobile communication load. CLA was proposed, demonstrating a systematic management of communication and computing resources. Note that CUs provide a resource pool, based on which VBSs are generated with virtualization techniques. Consider the communication protocol processing carried out in CUs using computing resources such as CPU and memories. Since one computing resource element (e.g., a group of CPUs) can handle protocol processing from several cells, one basic problem is how to map the protocol processing tasks from different cells to one computing resource element. *This computing resource allocation problem is closely related to the communication loads in the cells.* As shown in Fig. 5, if the protocol processing tasks from neighboring cells are grouped to be carried out on one computing resource element, to ensure that the resources are sufficient to process peak communication load at time $t_1$, only two cells can have their work processed by one element. Since the variations of load at neighboring cells are similar, it is highly possible that at time $t_2$, the load at both cells reduces and a large portion of computing resource is not used, resulting in low computing resource efficiency. The main idea of CLA is that *cells with large load discrepancies should be grouped together* such that at different times, the total computing resource needed are similar and overall efficiency can be improved.

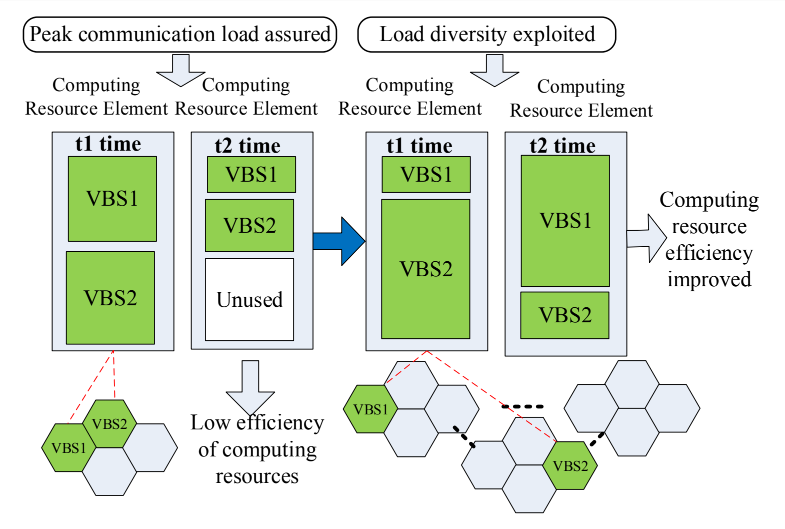
Peak communication load assured
Load diversity exploited
Computing Resource Element
Computing Resource Element
Computing Resource Element
Computing Resource Element
t1 time
t2 time
t1 time
t2 time
VBS1
VBS2
VBS1
VBS2
Unused
VBS1
VBS2
VBS1
VBS2
Computing resource efficiency improved
Low efficiency of computing resources
VBS2
VBS1
VBS1
VBS2

**Figure 5.** Illustration of CLA scheme.

The performance of CLA was verified via simulations. First, by running a practical LTE protocol software on a server with four CPUs (i.e., one computing resource element (CRE) with four processers), parameters needed for computing resource management were obtained, such as time for computing task migration. Then, based on the 24-hour network load statistics provided by China Mobile, the instant loads in N cells were generated via simulations, assuming a pure business or residential area. When the N cells are covered by a centralized mobile network, the average computing resource efficiency of CU is shown in Fig. 6 as a function of N. The comparing schemes, "peak/average load assured scheme (PLA/ALA)," would allocate computing resources in such a way that the resources needed for the peak/average load of each cell in one day should be ensured. It can be seen that *CLA outperforms both comparing schemes by exploiting the load diversity in the N cells.* PLA presents the lowest efficiency due to the large waste of computing resources when VBSs sharing the same CRE are all with low load. ALA improves the efficiency over PLA. The performance gap between CLA

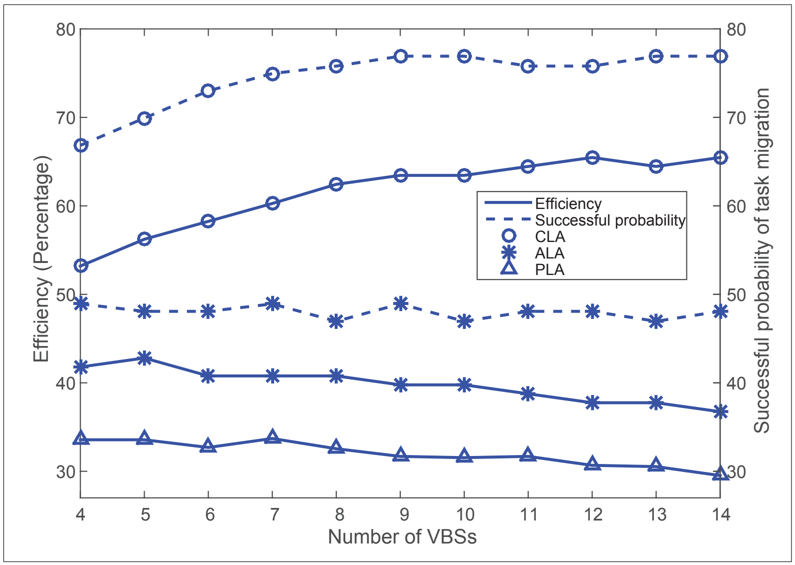

**Figure 6.** Performance of CLA as a function of the number of VBSs.

and PLA/ALA increases with N, since a larger statistical multiplexing gain can be achieved in resource pooling of CUs. *The proposed scheme can improve the efficiency over PLA by more than 100 percent* when there are 14 cells. Moreover, the successful probability of task migration is also shown. When the load of a VBS increases and the processor does not have sufficient computing resources to handle it, the task of the VBS needs to be migrated to other processors. If no other processor can handle it, the task migration fails. Note that PLA does not need the task migration since it is designed for the peak load. It can be seen that the task migration is more likely to be successfully implemented using CLA. Using ALA, VBSs associated to adjacent cells can be mapped to a certain CRE. When the load of all these VBSs reaches a peak simultaneously, the task migration is more likely to fail. Fortunately, this situation rarely occurs in CLA, which tries to map the VBSs with complementary loads to a certain CRE.

## Conclusions and Open Issues

The introduction of fog computing in mobile communication systems makes COM$^2$P possible. Various research has been carried out on FogMNW, investigating network architecture, system capacity and resource management. It has been shown that enabled by fog computing, COM$^2$P has the potential to improve the communication capacity of FogMNW, and a systematic management of communication and computing resources can enhance system performance significantly. However, there are still many open issues to be solved.

**Architectures for Future Smart Mobile Networks:** The introduction of fog computing makes the future mobile network not only a heterogeneous communication and hierarchical computing one, but also has the potential to be an intelligent one. For example, the network may learn to optimize its performance by itself, by collaboratively exploiting the communication and computing capabilities of all nodes. To achieve this, the proposed network architecture may need to be modified according to the requirements of learning algorithms. Moreover, to further support the sustainable development of mobile networks, more scalability, elasticity, transparency and openness are needed in future mobile network architectures.

**New Metrics to Measure the Capacity of FogMNW:** Although the framework proposed in [1] can be used to illustrate some practical examples, it could only provide the capacity gain and there still lacks a unified theoretical framework to obtain the capacity of FogMNW with both communication and computing resources. Since the Shannon capacity cannot evaluate the impact of computing resources on system performance, it may be insufficient to describe the capability of FogMNW. New metrics need to be defined, which should be theoretically tractable and could guide the design of COM$^2$P schemes. Moreover, considering COM$^2$P schemes, although coded multicast has clearly demonstrated the advantage of COM$^2$P, simple content access is considered to facilitate the theoretical analysis, that is, a user can always get the content as long as it is stored at the BS. However, in practical mobile networks, the content may not be able to reach the user due to the poor quality of wireless channels. Hence, *it is necessary to develop novel COM$^2$P schemes for FogMNW considering both content access and wireless channel access*.

**Systematic Management of Communication and Computing Resources:** Existing works on joint communication and computing resource management are mainly motivated by computing offloading. However, in practice, computing tasks cannot be split arbitrarily. How to perform optimization with these practical constraints is still an open problem. Moreover, to collaboratively manage communication and computing resources, the features of mobile services should be exploited. For example, the time-reversal feature of file delivery services was exploited by coded multiplex, such that files can be stored in a MS before the MS needs the service. Hence, developing service-aware cross-layer information exchanging schemes should also be a future focus.

## References

[1] H. Liu, Z. Chen, and L. Qian, "The Three Primary Colors of Mobile Systems," *IEEE Commun. Mag.*, vol. 54, no. 9, Sept. 2016, pp. 15–21.

[2] F. Bonomi *et al.*, "Fog Computing and its Role in the Internet of Things," *Proc. ACM 1st Ed. MCC Workshop Mobile Cloud Comput.*, Helsinki, Finland, Aug. 2012, pp. 13–16.

[3] ETSI, "Mobile-edge computing–Introductory technical white paper," White Paper, Sophia Antipolis, France, Sept. 2014. https://portal.etsi.org/portals/0/tbpages/mec/docs/mobileedge computing - introductory technical white paper v1%2018- 09-14.pdf, accessed May 15, 2016.

[4] N. Chen *et al.*, "Fog as a Service Technology," *IEEE Commun. Mag.*, vol. 56, no. 11, Nov. 2018, pp. 95–101.

[5] E. K. Markakis *et al.*, "Computing, Caching, and Communication at the Edge: The Cornerstone for Building a Versatile 5G Ecosystem," *IEEE Commun. Mag.*, vol. 55, no. 11, Nov. 2017, pp. 152–57.

[6] 3GPP TS 23.501, "System architecture for the 5G system (v1.6.0)," Technical Specification, Valbonne, France, Dec. 2017. https://portal.3gpp.org/desktopmodules/Specifications/Specification Details.aspx?specificationId=3144, accessed Dec. 28, 2017.

[7] IEEE Std. 1934TM-2018, "IEEE standard for adoption of OpenFog Reference Architecture for fog computing," IEEE, New York, Aug. 2018, pp. 1–176.

[8] M. Maddah-Ali and U. Niesen, "Fundamental Limits of Caching," *IEEE Trans. Inf. Theory*, vol. 60, no. 5, May 2014, pp. 2856–67.

[9] Y. Mao *et al.*, "A Survey on Mobile Edge Computing: The Communication Perspective," *IEEE Commun. Survey & Tutorials*, vol. 19, no. 4, Aug. 2017, pp. 2322–58.

[10] S. Kim, "5G Network Communication, Caching, and Computing Algorithms Based on the Two-Tier Game Model," *ETRI J.*, vol. 40, no. 1, Feb. 2018, pp. 61–71.

[11] K. Zhang *et al.*, "Energy-Efficient Offloading for Mobile Edge Computing in 5G Heterogeneous Networks," *IEEE Access*, vol. 4, Oct. 2016, pp. 5896–5907.

[12] G. Zhai *et al.*, "Load Diversity Based Optimal Processing Resource Allocation for Super Base Stations in Centralized Radio Access Networks," *Science China* (Information Sciences), vol. 57, no. 4, 2017, pp. 1–12.

[13] L. Liu *et al.*, "Load Aware Joint CoMP Clustering and Inter-Cell Resource Scheduling in Heterogeneous Ultra Dense Cellular Networks," *IEEE Trans. Veh. Technol.*, vol. 67, no. 3, Nov. 2017, pp. 2741–55.

[14] V. Garcia, Y. Zhou, and J. Shi, "Coordinated Multipoint Transmission in Dense Cellular Networks with User-Centric Adaptive Clustering," *IEEE Trans. Wireless Commun.*, vol. 13, no. 8, Apr. 2014, pp. 4297–4308.

[15] China Mobile Research Institute, "White paper of next generation fronthaul interface (v1.0)," Technical Report, Beijing, China, Oct. 2015. http://www.docin.com/p-1687649630.html, accessed Oct. 21, 2016.

## Biographies

Yiqing Zhou [S'03,M'05,SM'10] (zhouyiqing@ict.ac.cn) received the B.S. and M.S. degrees from Southeast University, China, and the Ph.D. degree from The University of Hong Kong, Hong Kong. She is currently a professor at ICT/CAS. She has published over 150 papers in the areas of wireless mobile communications. She received the 2014 Top 15 Editor Award from the IEEE TVT.

Lin Tian [M'07,SM'17] (tianlindd@ict.ac.cn) received the B.S. and M.S. degrees from Beihang University, and the Ph.D. degree from ICT/CAS, China, in 2002, 2005 and 2012, respectively. She is currently a professor at ICT/CAS. She has published more than 50 research papers and three book/book chapters in the area of mobile communication and computing.

Ling Liu [M'18] (liuling@ict.ac.cn) received the B.S. degree from Nanchang University in 2012 and the Ph.D. degree from the University of Chinese Academy of Sciences in 2018. She is currently an assistant professor at ICT/CAS. Her research focuses on interference and resource management in ultra-dense networks, and the convergence of communication, computing and caching.

Yanli Qi (qiyanli@ict.ac.cn) received the B.S. and M.S. degrees from Jiangxi University of Science and Technology, China, in 2014 and 2017, respectively. She is currently a Ph.D. candidate at ICT/CAS, University of Chinese Academy of Sciences. Her research focuses on mobile edge computing, convergence of communication, computation and cache, admission control and resource management.