An SDN/NFV Based Framework for Management and Deployment of Service Based 5G Core Network

Lu Ma[1,*], Xiangming Wen[1,2], Luhan Wang[1,2], Zhaoming Lu[1,2], Raymond Knopp[3]

[1] Beijing Key Laboratory of Network System Architecture and Convergence, Beijing University of Posts and Telecommunications, Beijing 100876, China
[2] Beijing Advanced Innovation Center for Future Internet Technology, Beijing 100124, China
[3] Communication System Department, EURECOM, Biot, 06410, France
* The corresponding author, email: malu@bupt.edu.cn

Abstract: The traffic explosion and the rising of diverse requirements lead to many challenges for traditional mobile network architecture on flexibility, scalability, and deployability. To meet new requirements in the 5G era, service based architecture is introduced into mobile networks. The monolithic network elements (e.g., MME, PGW, etc.) are split into smaller network functions to provide customized services. However, the management and deployment of network functions in service based 5G core network are still big challenges. In this paper, we propose a novel management architecture for 5G service based core network based on NFV and SDN. Combined with SDN, NFV and edge computing, the proposed framework can provide distributed and on-demand deployment of network functions, service guaranteed network slicing, flexible orchestration of network functions and optimal workload allocation. Simulations are conducted to show that the proposed framework and algorithm are effective in terms of reducing network operating cost.

Keywords: service based architecture; 5G core network; SDN; NFV; workload allocation

I. INTRODUCTION

The mobile networks are experiencing unprecedented changes both in traffic and types of communication in recent years. Increasing data traffic is generated in mobile networks in which video traffic grows fast. According to the report of Cisco [1], mobile data traffic will grow at a compound annual growth rate (CAGR) of 47 percent from 2016 to 2021, reaching 49 exabytes per month by 2021. Meanwhile, M2M (machine to machine) connections are calculated to grow from 780 million in 2016 to 3.3 billion by 2021. Three typical scenarios in 5G era are considered as: enhanced Mobile Broadband (eMBB), Ultra-Reliable and Low Latency Communications (URLLC), and massive Internet of Things (mIoT) [2]. eMBB focuses on services characterized by high data rates, such as high definition (HD) videos, virtual reality (VR), augmented reality (AR), etc. URLLC provides latency-sensitive services (e.g. self-driving). mIoT consists of devices in a huge connection density (e.g. smart city, smart agriculture, etc.). Obviously, the characteristics and requirements of traffic in mobile networks are evolving to diversity.

As a result, the increasing traffic and diverse requirements bring unprecedented challenges for the mobile network [3]. First, the explosion of mobile traffic and development of new radio access technologies make the bottleneck of the mobile network shifted from radio interface towards the backhaul and core networks [4]. Second, although the connections between human to machine, machine to

Received: Apr. 15, 2018
Revised: Jun. 7, 2018
Editor: Gaogang Xie

machine require very low data rates and are not sensitive to latency, the number of connections is extremely huge. This makes the core network suffer from signaling storm [5]. Third, the requirements of low latency and extremely high reliability which are essential for some vertical industries (e.g. industrial automation, autonomous driving, etc.) bring challenges of reducing latency for mobile network. The latency from air interface in 5G system is considered less than 1 ms to meet the requirement of these vertical industries [6].

However, the traditional mobile core network is designed in the monolithic pattern, that is, the functions of network elements is tightly coupled in a monolith running on a dedicated hardware which increases both the capital and operational expense of network operators, meanwhile causing the problems of ossification [7]. Any changes or updates to the network are related to the replacing of network devices. Consequently, it is difficult for conventional mobile network to meet the challenges brought by the traffic explosion and diverse requirements. To address the challenges in mobile networks, new technologies are emerging in the development of 5G mobile network architecture [8]. Software defined network (SDN) [9] and network function virtualization (NFV) [10] are considered attractive solutions to decouple the control and data planes, and decouple network functions from dedicated hardware in the mobile network [11]-[14]. Researchers in [15]-[18] proposed some SDN based approaches for mobile network which divide entities of conventional EPC into control planes and data planes to promote the efficiency of transmission. An NFV based virtual network of EPC functions running on universal servers was proposed in [17], which decoupled the network functions from the dedicated hardware. Also, many works brought both SDN and NFV into the design of mobile network architecture [18]-[20].

The SDN based works have improved the data forwarding efficiency in mobile core network, and these works based on NFV aim at promoting the efficiency of deployment and management of EPC network. Although these existing works improve the resource utilization of mobile networks, the mobile networks are difficult to provide customized network functions for diversified requirements of users in 5G era, because the network functions are still coupled as a monolith (e.g. Mobility Management Entity, MME; Serving Gateway, SGW; Packet Data Network Gateway, PGW; etc.). For example, the MME is responsible for the functions of session management, mobility management, and access management, etc. Moreover, it may lead to a waste of system resources. For instance, we know that the capacity extension should be done when the load of a system is over 90 percent. At this time, a horizontal copy of the system should be deployed and load balance between them should be conducted. However, some modules without heavy load are also deployed twice in the systems.

Recently, the architectural style of microservice is attracting more and more attention to provide flexible and on-demand customized services for diverse applications. The service based architecture (SBA) is proposed for 5G core network [21]. Inspired by microservice architecture pattern in software industry [22], 5G SBA is redesigned. Different from the existing EPC based architectures, the conventional network elements in mobile core network are split and modularized into smaller and more lightweight network functions which expose APIs for communication. However, the management and deployment of network functions in service based 5G core network are still big challenges.

In this paper, we investigate the manage-

The SDN based works have improved the data forwarding efficiency in mobile core network, and these works based on NFV aim at promoting the efficiency of deployment and management of EPC network.

Table I. *The main differences between proposed framework and existing works.*

	Technology adoption	Customized network slicing	Flexibility	Scalability	Backhaul cost
Softcell [14]	SDN	✓	Low	High	High
MobileFlow [16]	SDN	-	Low	Low	High
Softepc [17]	NFV	-	Low	Low	High
Softair [19]	SDN+NFV	✓	Low	High	Low
Softnet [20]	SDN+NFV	-	Low	High	High
Proposed framework	SDN+NFV	✓	High	High	Low

ment and deployment framework for 5G service based core network with consideration of service guaranteed network slicing, service chain orchestration, and workload allocation. The main differences between the existing works and the proposed framework are shown as Table 1. The main contributions of this paper are summarized as follows.

- We propose a novel management architecture for 5G service based core network based on NFV and SDN. In the proposed framework, the service management layer is responsible for the service management and orchestration, while the core and flow SDN controllers in the infrastructure management layer perform network function deployment, workload allocation, traffic scheduling, etc. Combined with SDN, NFV and edge computing, the proposed framework makes the development, deployment and management of 5G mobile network more flexible and efficient.
- The proposed framework can perform service guaranteed network slicing, flexible network function orchestration based on SDN and NFV technologies. And the end-to-end network slicing architecture is presented, where some network functions related to scenarios of eMBB and URLLC can be deployed on the edge servers.
- The optimal workload allocation for distributed 5G core network can be achieved in the proposed framework. We formulate the workload allocation as a total cost minimization problem with consideration of the bandwidth cost of backhaul network, the energy consumption of mobile cloud core and mobile edge core, revenue loss associated with backhaul delay. And an optimal workload allocation algorithm is proposed.
- Simulations are conducted to demonstrate that the proposed framework and workload allocation algorithm are effective in terms of reducing network operating cost.

The remainder of this paper is organized as follows. The introduction of SBA for 5G core network and the proposed management architecture for service based 5G core network are given in Section II. In Section III, we give some example network management applications and use cases that can be performed by leveraging the advantages of the proposed framework. Section IV is the evaluation of performance and analysis, and the conclusion and future work is in Section V.

II. The Management Architecture for 5G Service Based Core Network

In this section, we will first introduce the SBA for 5G core network, and then propose a novel architecture for management and deployment of 5G core network.

2.1 Service based architecture for 5G core network

The mobile communication network is experiencing explosive growth in data traffic and mobile device. In 5G communication system, the network operators will face great chances and challenges in order to accommodate the demands of diverse scenarios (e.g. eMBB, mIoT, URLLC, etc.). So that the 5G mobile network will be a service oriented network to enable more flexibility for the creation of new services and new applications. Actually, the conventional mobile core network is an integrate system, it consists of network elements (e.g. MME, SGW, PGW, etc.) which includes different function modules tightly coupled with dedicated hardware. The control plane and data plane are also coupled with hardware. This architecture cannot accommodates the diverse demands of users in 5G era because of the lacks in flexibility, scalability, and deployability.

Inspired by microservice architecture pattern in software industry, a service based architecture for 5G core network was proposed [21], as shown in figure 1. The 5G SBA is further designed according to microservice architecture pattern, the monolithic network elements are split into smaller network functions (NFs). Each NF may be responsible for a single task and can be deployed independently. Each NF exposes API that's consumed by other NFs or clients.

2.2 The proposed management architecture for service based 5G core network

In conventional cellular network, all mobile data traffic has to pass the core network to access services. However, in the 5G era, the global mobile data traffic is expected to grow at an extremely high speed because of the ultra-large content traffic. The explosion of mobile traffic and development of new radio access technologies make the bottleneck of the mobile network shifted from radio interface towards the backhaul and core networks.

On the other hand, 5G is expected to support ultra-reliable low-latency communications (URLLC) (i.e. mission-critical applications) which require uninterrupted and robust exchange of data [23]. For example, the autonomous driving and the smart grid controlling require radio latency of less than 1ms, and end-to-end latency of less than a few ms which cannot be guaranteed in current mobile network.

To solve the above challenges, we propose an NFV and SDN based architecture for management and deployment of 5G core network, as shown in figure 2. The 5G Service Portal is the entrance of NFs to provide different services for users. The Service Management Layer is responsible for orchestrating and configuring of NF modules based on the policy made by the Operation Support System (OSS) and Business Support System (BSS). The 5G Infrastructure Management Layer manages the infrastructure of 5G core, including flow scheduling, NF deployment, network slicing, etc. There are two kinds of SDN controllers in this layer: the Core SDN Controller in charge of NF management and coordination (e.g. service migration and deployment), the Flow SDN Controller in charge of efficient traffic dispatch in the backhaul network. Controlled by SDN controller, the 5G core user plane (UP) and some applications can be deployed on the edge servers as mobile edge core (MEC), while the control plane (CP) are deployed in the cloud date center as mobile cloud core (MCC). As a result, the requests related to mobile broadband and mission-critical applications could be served at edge of network to reduce the traffic anchoring to core network and eliminate backhaul delay.

Based on the proposed framework, we have accomplished the design and implementation of service based network slicing Orchestration and Management (O&M) architecture for 5G SBA, as shown in figure 3. OpenStack is utilized to realize the virtualization platform where the computing, storage resources of the NFVI are virtualized by means of a virtualized infrastructure manager (VIM). The NFV orchestration is implemented by HEAT module in OpenStack which can conduct deployment and lifecycle management of services based on HEAT templates. Ansible acts as the Virtualized Network Function Management (VNFM) which is an automation tool to configure, deploy, and orchestrate advanced tasks such as continuous deployment or zero downtime rolling updates. As for the management and orchestration of IP transport network resources, we make use of OpenDaylight (ODL) open source project. With ODL as the SDN controller, we can conduct the network resource management and structure multiple Virtual Tenant Network (VTN).

To implement the management and orchestration of network slicing, we introduce the concept of Network Operation System (NOS). The NOS mainly consists of common services, core orchestration and management function, drivers as well as the slice designer and specific UIs. The functions and implement of these

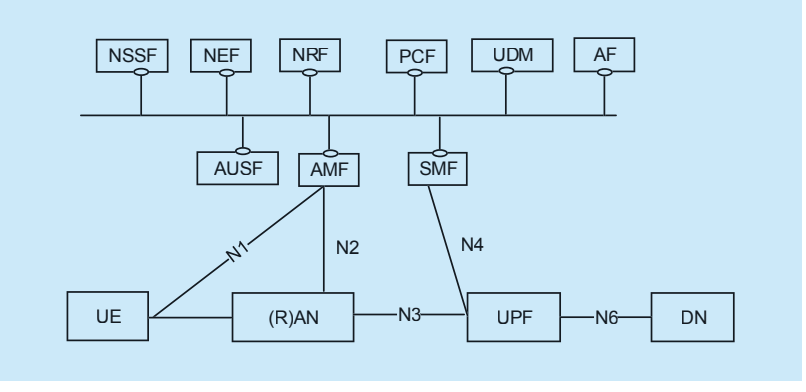

Fig. 1. *The service based reference architecture of 5G system.*

components are as follow:

1) The Common Services: provide the fundamental function for management of network slicing (e.g., tenant information management, service register and discovery, driver management, etc.):

- *Tenant Manager:* provides tenant information of management and authorization, authentication control.
- *System Manager:* provides system component management of slicing management system including register, update, and monitoring.
- *Catalogue:* provides catalogue management of the data related to the slice orchestration and management. The NGINX is utilized to implement this service
- *TOSCA:* tools provide common tools to interpret and construct TOSCA models. We make use of Apache ARIA TOSCA to implement the orchestrator of TOSCA model.
- *API GW (API Gateway):* is a significant component to support service oriented framework. It provides API management, register, publishing and coordination of network services. The open source WSO2 API Manager is utilized to implement the design, creation, publishing and management of APIs in each service.
- *Driver Manager:* provides the management of drivers.

2) The core orchestration and management: is responsible for slice monitoring, configuration and slice-level orchestration and management.

- *Slice O&M:* provides the orchestration and management of end-to-end network slices including lifecycle and context management, monitoring, configuration. The Flask microframework is utilized to implement the slice O&M framework.
- *Plans Engine:* is responsible for workflow model design of network slice. The open source project Activiti is used to implement the workflow management and design engine.
- *E2E Resource O&M:* provides end-to-end infrastructure resource management of network slice, including the register, activation, management of computing and storage resources in data center (DC) and IP transport network resources.
- *NFM module*: maintains and manages the interface information and parameters of network function. It can perform configuration, lifecycle management, monitoring of each network function. A build-in monitor component is implemented for collection

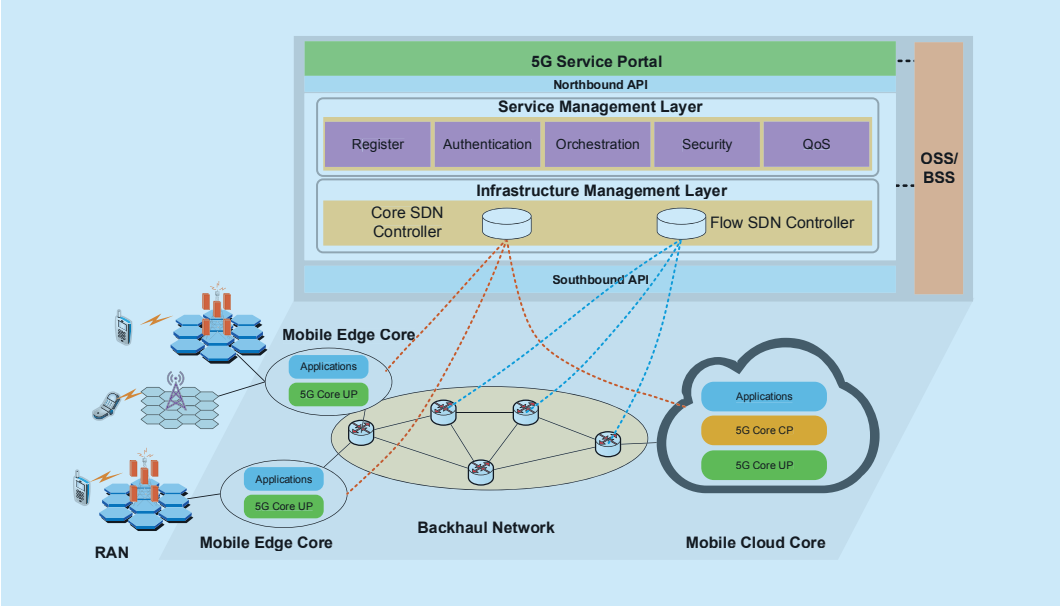

Fig. 2. *The management architecture for service based 5G core network based on SDN/NFV.*

and management of network function state information. The Build-in monitor is implemented by the open source Nagios which is monitoring system of running status information in the NOS.

3) The Driver layer: provides different kinds of drivers as adaptors between the NOS and SDN/NFV components.

4) The Slicing Designer: provides on-demand customization of network slice for users through generating meta data design files or scripts for TOSCA model.

The main novelties of the proposed management framework could be summarized as follows:

- In the proposed framework, all the services are managed and deployed on NFV based platform. It provides more flexible orchestration of NFs to get better system performance, satisfying diverse requirements of users.
- Two different SDN controllers are used in the infrastructure management layer. The Core SDN controller is responsible for deployment and management of NFs, and the Flow SDN controller handles efficient traffic dispatch of backhaul network. As a result, more flexible and on-demand deployment, monitoring and management of 5G core NFs can be realized. In the proposed framework, not all services will be implemented in the MEC. For instance, IMS service will continue to be implemented in MCC, while the ultra-real time services and video cache service will be handled by the MEC. Also, the challenges of the backhaul network can be addressed. On one hand, the flow SDN controller makes the flow transmission more efficient, on the other hand, some core network functions are distributed to the edge of the access networks, and a lot of ultra-broadband traffic are anchored the MEC, backhaul traffic will significantly decrease, thereby bringing down backhaul investment costs as well.
- In the proposed framework, 5G core network functions are deployed in a distributed way. As a complement to the MCC, the 5G MEC can process partial workloads locally on edge servers without transmitting them to the remote cloud data centers. As a consequence, the 5G core network can achieve lower end-to-end delay for URLLC traffic which need ultra-low latency of a few ms.
- In the proposed framework, the optimal strategy of minimizing the total cost of distributed 5G core network can be made in Core SDN controller by optimally deciding the number of active servers in cloud data centers, workload allocation between MEC and MCC.

III. Applications and Use Cases

In this section, we present three network management applications to demonstrate the advantages of the proposed framework based on SDN/NFV.

3.1 Service guaranteed network slicing

In the 5G era, there will be diverse range of service types which include three typical scenarios: eMBB, URLLC, mIoT. Hence, the operators have to serve a variety of devices with different characteristics and needs. For example, AR and VR have strict requirements on data rate and latency, while the mobile broadband service need high capacity and video cache. The demands in some vertical indus-

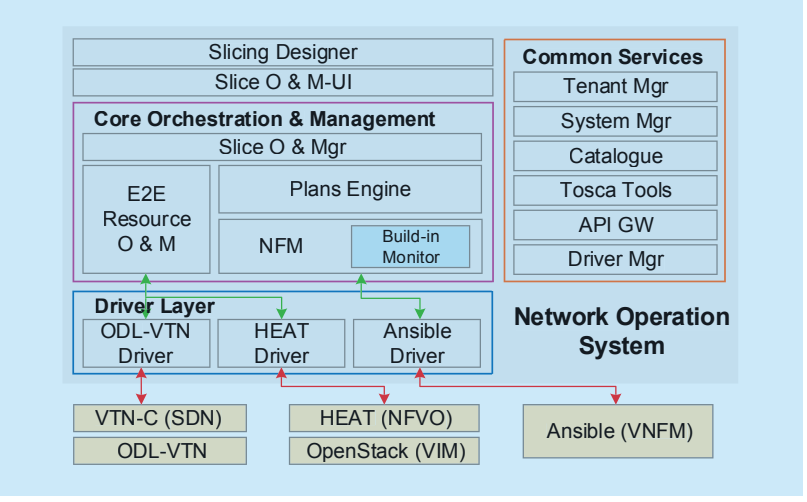

Fig. 3. *Implement of the service based network slicing O&M.*

tries are even more stringent because the terminals are normally machines with very low tolerance on performance degradation. As a result, the 5G network will be a service-centric network rather than the traditional user-centric network.

To achieve the service-centric network, it's not practical to deploy different kinds of dedicated networks for each service in current architecture. The service guaranteed network slicing can be implemented in the proposed framework to address this issue. The network slices are virtual and end-to-end networks which are each logically isolated including device, access, transmission and core network and dedicated for diversified requirements and characteristics of vertical industries.

Fig. 4 depicts the network slicing architecture based on the proposed framework. Microservice and NFV are the prerequisites to implement network slices. First, the NFs are installed onto VMs deployed on the virtualized commercial servers. Then, the virtualized network functions (VNFs) can be deployed in MEC or MCC depending on the types of services. As for the connectivity between VNFs in MEC, MCC, and traffic transmission in backhaul network, SDN plays an important role. SDN controller performs provisioning of routers in the backhaul network to create SDN tunnels (i.e. Virtual Private Networks, VPNs) for different slices and manage VNFs in the clouds. Consequently, operators can customize network slices in the way they want. For instance, the user plane functions, the cache, the V2X server can be placed in the mobile edge to accommodate the requirements of eMBB and URLLC slices, while the functions of mIoT and traditional voice will implemented in the mobile core cloud.

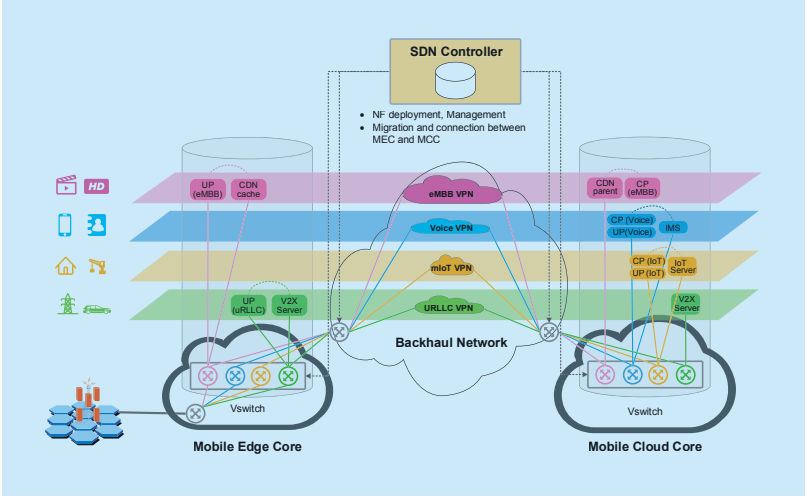

Fig. 4. *The service guaranteed network slicing architecture.*

For each network slice, isolation and dedicated resources such as resources within virtualized servers, network bandwidth and QoS are guaranteed. The proposed framework is capable to provide logical dedicated networks upon a common infrastructure and supports the operators to explore deeper business potentials by providing customized services.

3.2 Flexible orchestration of network functions

The proposed management framework can provide more flexible orchestration of NFs to get better system performance, satisfying diverse requirements of users. Through splitting network elements into more lightweight NFs and combining with SDN and NFV, the proposed management framework allows network operators to carry out more granular approaches for management of networks. Dynamic management of service chain is supported in the proposed framework. In 5G era, not only the traffic types are diverse, but also the resource requirements of each scenario are dynamically changing over time and diverse among different users. Hence, the mobile operators should have a collection of services which can be run on particular data flows. The number of services and the order in which they are applied depend on the traffic and users. In the proposed framework, a logically centralized controller is adopted and the network traffic can go through the required NFs to realize service chaining. Moreover, the service chains can be managed dynamically to accommodate the dynamic requirements of users. As shown in figure 5, a user activates two types of flows which are routed to different services. The red flow goes through App 1 and App 2, while the green flow is only handled by

App 3. The orchestration of service chains is dynamic and according to the traffic types and the policy made by the network operator. As a result, the proposed framework can provide dynamic network operation and maintenance (O&M).

Besides, the proposed management framework supports flexible and dynamic service migration by starting or terminating associated NVFs. The SDN controller will make a trade-off between the revenue and expense of the migration based on the current network status (e.g. latency, bandwidth, QoS, cost, etc.) and then determine whether migrate the service or not. The Migration Determining Factor (MDF) is defined to represent the relative cost of a service migration. The MDF of service node x is a function of the latency, the available load, bandwidth, service price and credit, etc. If y is defined as the previous service provider (i.e. the start point of migration), the MDF can be expressed as:

$$MDF(x) = \sum f(C_{lat}(x), C_{load}(x), C_{band}(x,y), C_{cre}(x), C_{pri}(x)) \quad (1)$$

where the $C_{lat}(x)$ is the latency of service transmission from providers to node x, the $C_{load}(x)$ means the available capacity of servers at node x, the $C_{band}(x,y)$ represents the max available bandwidth of the migration link from start point x to the end point y, the $C_{cre}(x)$ is the credit of service provider which is associated to feedbacks of users, the $C_{pri}(x)$ is the cost of renting server resources.

Consequently, we can get a matrix Q which consist of the determining factors of all service nodes. Through setting the weight coefficients for parameters and normalizing the matrix, we can get the relative cost of migration to each node. Then, the node has minimum cost will be selected as the service provider.

3.3 Optimal workload allocation in 5G core network

In the proposed framework, 5G core network functions can be deployed in a distributed way. As a complement to the MCC, the 5G MEC can process partial workloads locally on edge servers without transmitting them to the remote cloud data centers. As a consequence, 5G network can benefit a lot, in terms of supporting applications with stringent latency requirements, providing location-aware services, cutting down the cost of backhaul, etc. Since the high operational cost of cloud and the collaboration requirements between MCC and MEC of some services, it is important to manage the operational cost in a distributed 5G core network. In the proposed framework, the strategy of minimizing the total cost of distributed 5G core network can be made in Core SDN controller by optimally deciding the number of active servers in cloud data centers, workload allocation between MEC and MCC.

To investigate the problem, we assume that the 5G core network consists of N MECs and M cloud data centers in MCC to serve J types of requests. Edge servers are co-located with their corresponding base stations and collaborate with MCC to process the tasks from users. The problem modelling and formulation are given as flows.

1) Energy cost of MEC: We assume that there is a linear relationship between energy consumption and workload. Let $x_{i,j}$ denote the request rate of application j allocated to MEC i (in requests/second). Hence, we can express the energy cost of MECs with duration T by:

$$C^{edge} = \sum_{i=1}^{N}\sum_{j=1}^{J} h_i a_i x_{i,j} T. \quad (2)$$

where $a_i > 0$ is the predetermined parameter which is associated with the deployment of MEC, h_i is the electricity price associated with electric region that edge server is located

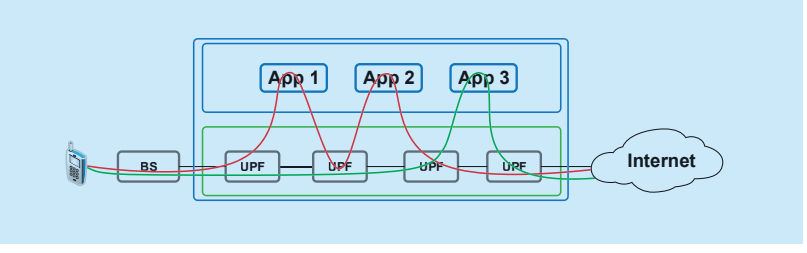

Fig. 5. *Dynamic service chain management in the proposed framework.*

(in dollars/MWh).

2) Processing delay of MEC: M/M/1 model is adopted to evaluate the processing delay of application j at MEC i [24]. Then, we have:

$$D_{i,j}^{edge} = \frac{1}{v_{i,j} - x_{i,j}}. \quad (3)$$

where $v_{i,j}$ is the service rate of application j at the MEC i (in requests/second).

3) Energy cost of MCC: Let $y_{i,j,m}$ be the request rate of application j allocated from MEC i to data center m (in requests/second), $\mu_{j,m}$ be the service rate of application j in data center m. In addition, we denote the idle power and peak power of servers for application j in data center m by $p_{j,m}^{idle}$ and $p_{j,m}^{peak}$ respectively. Then the energy cost of MCC could be estimated by [25]:

$$C^{cloud} = \sum_{j=1}^{J}\sum_{m=1}^{M} f_m \left(c_{j,m} p_{j,m}^{idle} + (p_{j,m}^{peak} - p_{j,m}^{idle}) \frac{\sum_i y_{i,j,m}}{\mu_{j,m}} \right) T \quad (4)$$

where $c_{j,m}$ is the number of active servers for application j in data center m, f_m is the electricity price factor associated with electric region that data center m is located (in dollars/MWh).

4) Processing delay of MCC: M/M/n model is adopted to evaluate the processing delay of application j at data center m. Then, we have:

$$D_{j,m}^{cloud} = \frac{1}{c_{j,m}\mu_{j,m} - \sum_i y_{i,j,m}} + \frac{1}{\mu_{j,m}}. \quad (5)$$

5) Bandwidth cost of backhaul network: The bandwidth cost of backhaul can be expressed as:

$$C^{bandwidth} = \sum_{i=1}^{N}\sum_{j=1}^{J}\sum_{m=1}^{M} y_{i,j,m} s_j P_m. \quad (6)$$

where s_j is the request size of application j (in Mb/request), P_m is the bandwidth price of backhaul to data center m (in dollars/Mbps).

6) Revenue loss related to backhaul delay: Because the backhaul delay usually leads to revenue loss, we deem it as a kind of cost. Denote the delay of backhaul network associated with MEC i and data center m by $d_{i,m}$ (in ms). Then, the revenue loss related to backhaul delay can be obtained by:

$$C^{delay} = \sum_{i=1}^{N}\sum_{j=1}^{J}\sum_{m=1}^{M} \varepsilon_j d_{i,m} y_{i,j,m} T. \quad (7)$$

where ε_j is the penalty factor of application j (in dollar/ms).

Consequently, the optimal workload allocation can be formulated as total system cost minimization problem:

$$\begin{aligned} &\min \ C^{edge} + C^{bandwidth} + C^{delay} + C^{cloud}. \quad (8)\\ &s.t. \ C1: \ D_{i,j}^{edge} \le t_j^{\max}, \ \forall\, i, j\\ &C2: \ x_{i,j} \ge 0, \ \forall\, i, j\\ &C3: \ l_{i,j} = x_{i,j} + \sum_{m=1}^{M} y_{i,j,m}, \ \forall\, i, j\\ &C4: \ \sum_{i=1}^{N}\sum_{j=1}^{J} y_{i,j,m} s_j \le B_m^{\max}, \ \forall\, m \quad (9)\\ &C5: \ y_{i,j,m} \ge 0, \ \forall\, i, j\\ &C6: \ D_{j,m}^{cloud} \le t_j^{\max}, \ \forall\, m\\ &C7: \ 0 \le c_{j,m} \le C_{j,m}, \ c_{j,m} \in \mathbf{N}^+, \ \forall\, j, m \end{aligned}$$

where $t_j^{\max}$ is the maximum tolerant delay of application j, $l_{i,j}$ denotes the request rate of application j arrived at MEC i, and $C_{j,m}$ is the total number of servers for application j in data center m. The decision variables are $x_{i,j}$, $y_{i,j,m}$, and $c_{j,m}$.

Because $c_{j,m}$ is an integer variable, the problem is a mixed nonlinear integer programming (MNIP) which is difficult to solve. Hence, we first study the objective function when $c_{j,m}$ is regarded as non-integer variable. In this case, we give the following proposition:

Proposition 1: If each $c_{j,m}$ is regarded as non-integer variable, total system cost minimization problem is a convex optimization problem.

Proof: Since $\sum_{m=1}^{M} y_{i,j,m} = l_{i,j} - x_{i,j}, \ \forall\, i, j$ and for simplicity, we use $\{z_1, z_2, ..., z_N, z_{N+1}, ..., z_{N+M}\}$ to substitute $\{x_1, x_2, ..., x_N, c_1, ..., c_{M-1}, c_M\}$ for each request j, the objective function f can be transformed into a function of variables $\{z_1, z_2, ..., z_N, z_{N+1}, ..., z_{N+M}\}$. Hence the hessian

matrix of f is $\mathbf{H}=\begin{bmatrix}\mathbf{P} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}_{(N+M)\times(N+M)}$, where $\mathbf{P}=\begin{bmatrix}\frac{\partial^2 f}{\partial z_1^{\,2}} & 0 & 0\\ 0 & \ldots & 0\\ 0 & 0 & \frac{\partial^2 f}{\partial z_N^{\,2}}\end{bmatrix}_{N\times N}$. Because the hessian matrix $\mathbf{H}$ is a symmetric and semi-positive definite matrix, and constraints in (9) are linear, the system cost minimization problem in (8) is a convex optimization problem [26].

Therefore, the interior-point algorithm can be used to solve the convex optimization problem, which has obvious advantages in convergence and computation speed.

Based on the above modelling and formulation, we develop a workload allocation algorithm to minimize the system cost, which can be implemented in the proposed framework. The details of the workload allocation algorithm are described as Algorithm 1.

In the proposed algorithm, the SDN controllers collect the system parameters from servers in MECs and cloud data center in MCC. And then it solves the cost minimization problem regarding $c_{j,m}$ as a non-integer variable, because the problem is a mixed nonlinear integer programming (MNIP) which is difficult to solve. After the decision variables are calculated, $c_{j,m}^{(1)}$ is rounded-up to get the optimal integer variable $c_{j,m}^{*}$. Finally, the SDN controllers make the optimal workload allocation strategy based on the result and send it to MECs and MCC.

IV. Evaluation and Analysis

In this section, we evaluate the optimal workload allocation and provide simulation results to show the advantages of proposed framework, compared with a baseline algorithm. The baseline means minimizing the consumption of core network without collaborating with edge network when processing the workload from users. The simulations are performed with MATLAB on a single Intel Core i5-2410U 2.39GHz PC with 8G RAM. The main system parameters are given as follows, $N=10$, $J=2$, $M=3$, $T=1$ hour, $a_i \sim U(1,2)$, $C_{j,m}=[100,100,80;\ 80,90,100]$, $t_j^{\max}=[1,\ 0.8]$ s, $l_{i,j} \sim U(2.4,3)$ requests/s, $v_{i,j} \sim U(3,4)$ requests/s, $d_{i,m} \sim U(0.05,1)$ ms, $P_m \sim U(1000,1500)$ dollars/Mbps, $\mu_{j,m}=[2,1.8,1.5;\ 1.75,1.6,1.9]$ requests/s, $B_m^{\max}=[10,9,8]*10^4$ Mbps, $p_{j,m}^{peak}=2p_{j,m}^{idle}=[220,190,200;\ 200,180,240]$ Watts, $\varepsilon_j=[2,3]*10^{-4}$ dollars/ms/request, $s_1=2$ Mb, $s_2=1$ Mb, $f_m=[80,70,90]$ dollars/MWh.

Fig. 6 shows the total system cost when the energy price h_i of edge servers increase from 10 to 70. As can be seen from the figure, the proposed framework can reduce system cost when the energy price of edge servers is less. And the total cost of proposed framework is equal with the baseline when the energy price of edge servers is too high, because all the workload will be allocated to MCC.

Fig. 7 illustrates the workload allocated to MECs as the increase of h_i, and the impacts of backhaul delay. We can see that more workload is allocated to MECs when the cost of edge servers is less. As the increase of edge server

Algorithm 1. Workload allocation algorithm.

Input: $l_{i,j}$, $v_{i,j}$, $t_j^{\max}$, ε_j, s_j, f_m, $d_{i,m}$, $C_{j,m}$, p_j^{idle}, p_j^{peak}, P_m, $B_m^{\max}$, $\mu_{j,m}$

Output: $x_{i,j}$, $y_{i,j,m}$, $c_{j,m}$

1: Initiation: Choose initial point $x_{i,j}^{(0)}$, $y_{i,j,m}^{(0)}$, $c_{j,m}^{(0)}$ in feasible domain.
2: **for** Each time period T **do**
3: The SDN controllers collects status messages from MECs and MCC to get the input parameters
4: Solve (8) regarding $c_{j,m}$ as a non-integer variable, and the solution is $x_{i,j}^{(1)}$, $y_{i,j,m}^{(1)}$, $c_{j,m}^{(1)}$
5: $x_{i,j}^{*}=x_{i,j}^{(1)}$, $y_{i,j,m}^{*}=y_{i,j,m}^{(1)}$, $c_{j,m}^{*}=\lceil c_{j,m}^{(1)} \rceil$
6: **return** $x_{i,j}^{*}$, $y_{i,j,m}^{*}$, $c_{j,m}^{*}$
7: The SDN controllers make the optimal workload allocation strategy based on the result and send it to MECs and MCC
8: **end for**

cost, the workload processing at MEC reduces to zero. It can be seen from the figure, with the same value of h_i, more workload is allocated to MECs when the backhaul network has higher revue loss associated with delay. The major reason is that the MECs are more proximate to users, reducing the backhaul delay.

Fig. 8 shows the workload allocated to MCC when h_i increases from 10 to 70, and the impacts of the bandwidth cost of backhaul network. It can be seen from the figure, as the increase of edge server cost, more workload is processed at MCC. Also, with the same value of h_i, less workload is allocated to MCC when the bandwidth price of backhaul network is higher, and processing workload at MECs reduces the system cost.

Fig. 9 shows the system cost when the request size increases from 1 to 10 Mb/request. As can be seen from figure 8, the system cost of proposed framework is lower than the baselines, even when it has hybrid backhaul which

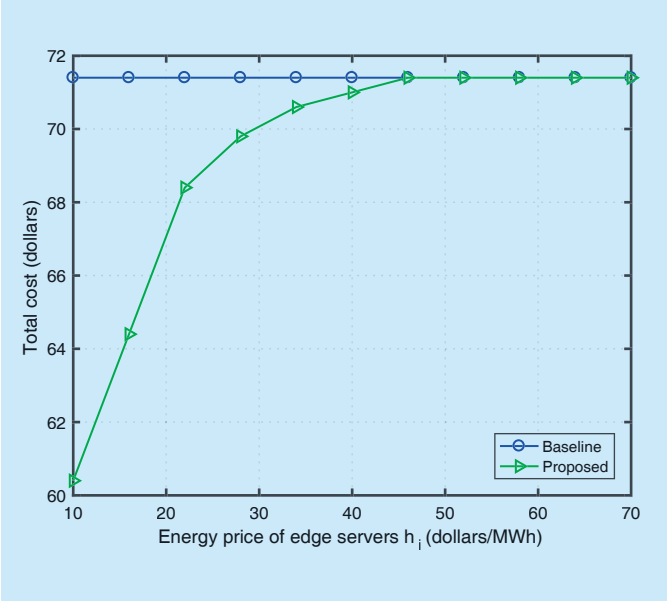

Fig. 6. *Total system cost.*

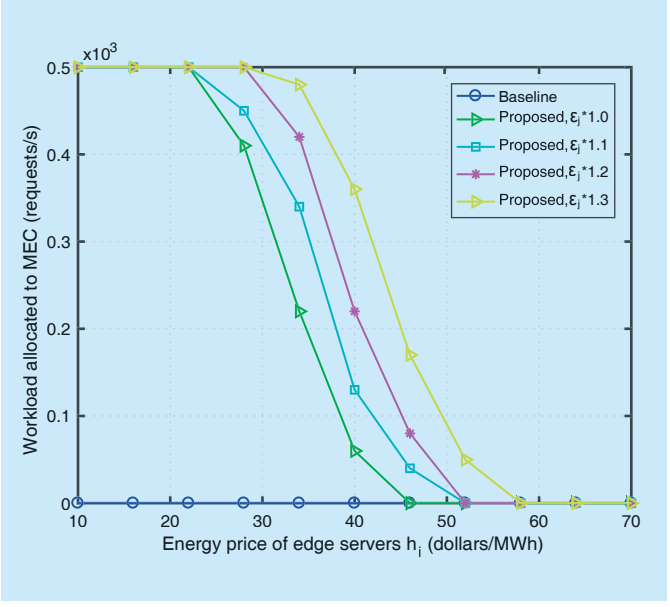

Fig. 7. *Workload allocated to MEC.*

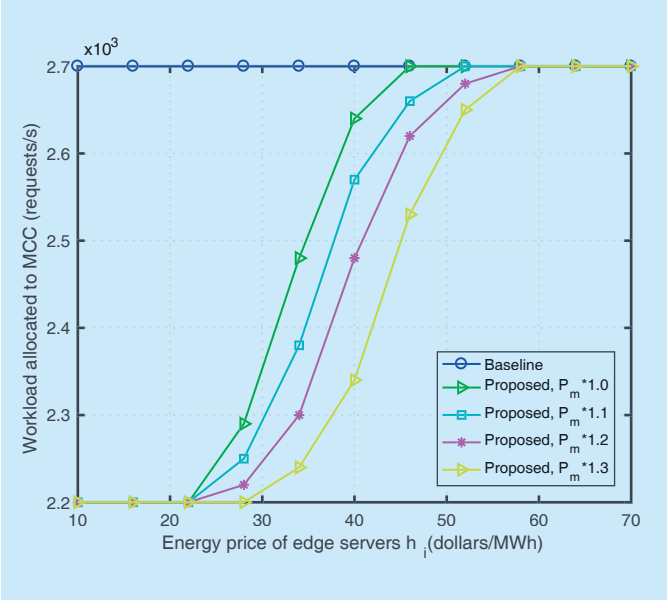

Fig. 8. *Workload allocated to MCC.*

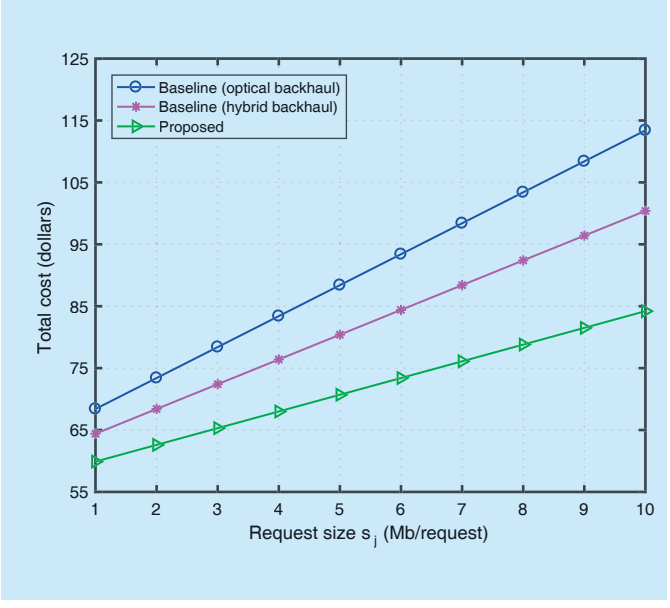

Fig. 9. *The impact of request size.*

is cheaper than optical backhaul. It also can be seen that the performance gaps between the proposed method and baselines are widening with the increase of request size. This is because larger request size brings more backhaul cost, and the MECs can reduce backhaul cost due to the distributed architecture.

V. Conclusion

In this paper, the SBA for 5G core network is introduced, which is considered as a promising architecture to improve the flexibility, scalability, and deployability of traditional core network and to provide customized services for diverse requirements. Then, we propose a novel management architecture for 5G core network SBA and give three use cases. Also, an optimal workload allocation algorithm is developed. Combined with SDN, NFV and edge computing, the proposed framework can provide distributed and on-demand deployment of network functions, service guaranteed network slicing, flexible orchestration of network functions and optimal workload allocation. The simulation results show that the proposed framework and algorithm are effective in terms of reducing network operating cost.

Acknowledgements

This work is supported by China Ministry of Education-CMCC Research Fund Project No. MCM20160104, National Science and Technology Major Project No. No. 2018ZX03001016, Beijing Municipal Science and technology Commission Research Fund Project No. Z171100005217001 and Fundamental Research Funds for Central Universities NO. 2018RC06. Besides, we would like to thank EURECOM and OpenAirInterface Alliance for their support and help.

References

[1] Cisco, "Cisco visual networking index: Global mobile data traffic forecast update, 2016-2021," *Cisco White Paper*, 2017.

[2] 3GPP, "5g: A tutorial overview of standards, trials, challenges, deployment and practice," *Technical Report TR38.913*, 2016.

[3] M. Shafi, A. F. Molisch, P. J. Smith, T. Haustein, P. Zhu, P. D. Silva, F. Tufvesson, A. Benjebbour, and G. Wunder, "5g: A tutorial overview of standards, trials, challenges, deployment, and practice," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 6, 2017, pp. 1201–1221.

[4] Y. I. Choi, J. H. Kim, and N. I. Park, "Revolutionary direction for 5g mobile core network architecture," *IEEE 2016 International Conference on Information and Communication Technology Convergence (ICTC)*, 2016, pp. 992–996.

[5] H. Baba, M. Matsumoto, and K. Noritake, "Lightweight virtualized evolved packet core architecture for future mobile communication," *2015 IEEE Wireless Communications and Networking Conference (WCNC)*, 2015, pp. 1811–1816.

[6] B. Holfeld, D. Wieruch, T. Wirth, L. Thiele, S. A. Ashraf, J. Huschke, I. Aktas, and J. Ansari, "Wireless communication for factory automation: an opportunity for lte and 5g systems," *IEEE Communications Magazine*, vol. 54, no. 6, 2016, pp. 36–43.

[7] Y. Li and M. Chen, "Software-defined network function virtualization: A survey," *IEEE Access*, vol. 3, 2015, pp. 2542–2553.

[8] V. G. Nguyen, A. Brunstrom, K. J. Grinnemo, and J. Taheri, "Sdn/nfvbased mobile packet core network architectures: A survey," *IEEE Communications Surveys Tutorials*, vol. 19, no. 3, 2017, pp. 1567–1602.

[9] T. Chen, M. Matinmikko, X. Chen, X. Zhou, and P. Ahokangas, "Software defined mobile networks: concept, survey, and research directions," *IEEE Communications Magazine*, vol. 53, no. 11, 2015, pp. 126–133.

[10] B. Han, V. Gopalakrishnan, L. Ji, and S. Lee, "Network function virtualization: Challenges and opportunities for innovations," *IEEE Communications Magazine*, vol. 53, no. 2, 2015, pp. 90–97.

[11] R. Riggio, A. Bradai, D. Harutyunyan, T. Rasheed, and T. Ahmed, "Scheduling wireless virtual networks functions," *IEEE Transactions on Network and Service Management*, vol. 13, no. 2, 2016, pp. 240–252.

[12] Yuhuai Peng, Xiaoxue Gong, Lei Guo and Dezhi Kong, "A survivability routing mechanism in SDN enabled wireless mesh networks: Design and evaluation," *China Communications*, vol. 13, no. 7, 2016, pp. 32-38.

[13] B. Han, V. Gopalakrishnan, L. Ji, and S. Lee, "Network function virtualization: Challenges and opportunities for innovations," *IEEE Communications Magazine*, vol. 53, no. 2, 2015, pp. 90–97.

[14] L. Tao, W. Xiangming, L. Zhaoming, Z. Xing, L. Yangchun and Z. Biao, "SWN: An SDN based framework for carrier grade Wi-Fi networks," *China Communications*, vol. 13, no. 3, 2016, pp. 12-26.

[15] L. V. J. R. Jin Xin, Li Erran Li, "Softcell: Scalable and flexible cellular core network architecture,"

Proceedings of the ninth ACM conference on Emerging networking experiments and technologies, 2013, pp. 163–174.

[16] V. Yazici, U. C. Kozat, and M. O. Sunay, "A new control plane for 5g network architecture with a case study on unified handoff, mobility, and routing management," *IEEE Communications Magazine*, vol. 52, no. 11, 2014, pp. 76–85.

[17] K. Pentikousis, Y. Wang, and W. Hu, "Mobileflow: Toward softwaredefined mobile networks," *IEEE Communications Magazine*, vol. 51, no. 7, 2013, pp. 44–53.

[18] B. Zhu and H. Jiang and L. Wu and S. Yi and H. Wang, "A dynamic bandwidth allocation solution based on user behavior and software defined EPS," *China Communications,* vol. 13, no. 9, 2016, pp. 80–90.

[19] F. Z. Yousaf, J. Lessmann, P. Loureiro, and S. Schmid, "Softepc: Dynamic instantiation of mobile core network entities for efficient resource utilization," *2013 IEEE International Conference on Communications (ICC)*, *IEEE,* 2013, pp. 3602–3606.

[20] W. P. Akyildiz I F, Lin S C, "Wireless software-defined networks (wsdns) and network function virtualization (nfv) for 5g cellular systems: An overview and qualitative evaluation," *Elsevier Computer Networks*, vol. 93, 2015, pp. 66–79.

[21] L. S. C. Akyildiz I F, Wang P, "Softair: A software defined networking architecture for 5g wireless systems," *Elsevier Computer Networks,* vol. 85, 2015, pp. 1–18.

[22] H. Wang, S. Chen, H. Xu, M. Ai, and Y. Shi, "Softnet: A software defined decentralized mobile network architecture toward 5g," *IEEE Network*, vol. 29, no. 2, 2015, pp. 16–22.

[23] 3GPP, "3gpp ts 23.501, system architecture for the 5g system (release 15)," *3rd Generation Partnership Project: Sophia-Antipolis, France* 2017.

[24] M. Fowler, "Microservices guide," *[Online]. Available: http://martinfowler.com/microservices/*.

[25] P. Schulz, M. Matthe, H. Klessig, M. Simsek, G. Fettweis, J. Ansari, S. A. Ashraf, B. Almeroth, J. Voigt, I. Riedel, A. Puschmann, A. Mitschele-Thiel, M. Muller, T. Elste, and M. Windisch, "Latency critical iot applications in 5g: Perspective on the design of radio interface and network architecture," *IEEE Communications Magazine*, vol. 55, no. 2, 2017, pp. 70–78.

[26] R. Deng, R. Lu, C. Lai, T. H. Luan, and H. Liang, "Optimal workload allocation in fog-cloud computing toward balanced delay and power consumption," *IEEE Internet of Things Journal*, vol. 3, no. 6, 2016, pp. 1171–1181.

[27] F. Jalali, K. Hinton, R. Ayre, T. Alpcan, and R. S. Tucker, "Fog computing may help to save energy in cloud computing," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 5, 2016, pp. 1728–1739.

[28] Boyd S, Vandenberghe L. Convex optimization[M]. *Cambridge university press*, 2004.

Biographies

Lu Ma, received the B.S. degree in Communication Engineering from North China Electric Power University. He is currently a Ph.D candidate in Information and Communication Engineering at Beijing University of Posts and Telecommunications (BUPT), China. His current research interests include 5G mobile communication network, network virtualization, mobile core network.

Luhan Wang, received the Ph.D in Beijing University of Posts and Telecommunications (BUPT) in 2017. He joined the School of Information and Communication Engineering in BUPT as an assistant professor in 2017. His current research interests include network architecture, network function virtualization and soft-defined networks.

Zhaoming Lu, received the Ph.D in Beijing University of Posts and Telecommunications (BUPT) in 2012. He joined the School of Information and Communication Engineering in BUPT in 2012. His research includes Open Wireless Networks, QoE management in wireless networks, and so on.

Xiangming Wen, received the B.E., M.S., and Ph.D. degrees from Beijing University of Posts and Telecommunications (BUPT), all in electrical engineering. He is currently the Vice President of BUPT, where he is also the Director of the Beijing Key Laboratory of Network System Architecture and Convergence. His current research is focused on broadband mobile communication theory, multimedia communications, and so on.

Raymond Knopp, received the B.Eng. (Hons.) and the M.Eng. degrees from McGill University, Montreal, QC, Canada, in 1992 and 1993, respectively, and the Ph.D. degree from the Swiss Federal Institute of Technology (EPFL), Lausanne, in 1997. He is currently a Professor of EURECOM. He is the General Secretary of the OpenAirInterface open-source software alliance.