NFV and SFC: A Case Study of Optimization for Virtual Mobility Management

Hao Jin, Yi Jin, Haiya Lu, Chenglin Zhao, and Mugen Peng, *Senior Member, IEEE*

***Abstract*— To support the typical application scenarios defined in the fifth generation wireless network, such as the enhanced mobile broadband, massive machine type of communication, ultra-reliable and low-latency communication, virtual Mobility Management Entity (vMME) is a promising solution, which runs on universal servers and network functions virtualization instead of conventional hardware-dedicated mobility management entity. Among different vMME mapping solutions, the decomposing of MME into multiple components is a prospective approach to implementing distributed and virtualized mobility management. In this paper, the optimization of mobility management is addressed by using NFV and service function chain. A general signaling processing flow of vMME based on service function chain is analyzed. The performance of vMME is formulated considering the total signaling communication overhead cost, the total signaling communication overhead cost on backhauls as well as the migration overhead cost of state data under different function component placement of vMME. Since this optimization problem is NP-hard and the computation complexity is $O(n^k)$, a heuristic approach consisting of Min-TSCOC, Min-TSCOCB, and Min-MOCSD are presented, which aims to obtain the optimal solutions to the optimization problems by using genetic algorithm. The simulation results show that it is beneficial to decompose the function of mobility management, and the performance gains from vMME for different network function composition strictly depend on four mobility events.**

***Index Terms*— Network function virtualization, service function chain, mobility management.**

I. Introduction

With the rapid development of radio access network, more and more applications are deployed in mobile internet, which promotes the 5G research on typical service scenarios including Enhanced Mobile Broadband (eMBB), massive Machine Type of Communication (mMTC), Ultra Reliable and Low Latency Communication (URLLC) as well as vehicular network [1]–[3]. eMBB provides higher rate and access capabilities to support quality of experience of mobile users. The scenario for mMTC demands for optimized signaling control with high connection density, efficient access and management of large-scale, low-cost and low-power IoT devices. The scenario for URLLC leverages low latency and reliable information interaction capabilities to support real-time, sophisticated and secure collaboration among connected entities. The vehicular application scenario provides networking and communication for high-speed vehicles. The typical networking architecture of 5G radio access network and various service requirements brings a great challenge to the existing mobility management based on EPC.

The existing mobility management solution of cellular mobile network provides mobility management functions through functional modules including source base station, target base station, Mobility Management Entity (MME), Serving Gateway (S-GW), Packet Gateway (P-GW), HSS, and PCRF. The core functional modules of mobility management are deployed in the core network in a centralized manner including MME, S-GW, P-GW, HSS and PCRF. Therefore, the existing solution for mobility management brings about not only limitation of single point of failure, but also scalability and elasticity to meet various service requirements and radio access network architecture [4], [5]. In addition, in the existing mobility management solution, the control plane and the user plane are closely coupled. The control plane is responsible for maintaining the mobility context of the UE, while the user plane is for handling the routing of packets. The heterogeneity, dense deployment, and flattening characteristics of 5G mobile radio access networks directly affects the performance of control plane for mobility management.

To address the limitations of existing mobility management, two methodologies are provided, namely distributed mobility management [4], [5] and virtualized mobility management [6].

Distributed mobility management pushes the management of cellular mobility management toward the edge of the network to meet the various demands of services as well as the performance requirements of mobility management. A distributed system called DMME is presented to implement mobility management for LTE (Long Term Evolution), which confirms that distributed architectures are efficient to reliably support high-throughput, latency-sensitive control plane functions by prototype implementation and simulations [5].

In order to reduce the deployment cost of MME, virtual Mobility Management Entity (vMME) was proposed as an alternative solution based on Network Functions Virtualization (NFV) instead of conventional hardware-dedicated MME entity [6]. NFV-based mobility management provides more flexible, scalable and resilient mobility management service with on-demand orchestration of mobility management, which

Manuscript received April 30, 2018; revised August 3, 2018; accepted August 28, 2018. Date of publication September 27, 2018; date of current version November 28, 2018. This work was supported in part by the National Natural Science Foundation of China under Grants 61471062 and 61431008, and in part by the State Major Science and Technology Special Projects under Grant 2017ZX03001014. *(Corresponding author: Hao Jin.)*

The authors are with the Wireless Signal Processing and Network Laboratory, Key Laboratory of Universal Wireless Communications, Ministry of Education, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: hjin@bupt.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSAC.2018.2869967

enables distributed and semi-distributed solutions for mobility management.

Network function virtualization is essentially the relocation of network functions from standalone boxes based on dedicated hardware to software appliances running in the cloud environment or on general-purpose commodity servers. By using NFV, each conventional network function (NF) is now running on a virtual machine (VM) as a 1:1 mapping model or is decomposed into smaller components called Virtual Network Function Components (VNFC) running on multiple VMs as a 1:N mapping model [7]–[9], where 1 usually represents the decomposed network function, and N is the number of VNFCs after function decomposition. Programmable network connectivity provided by Software Defined Networking (SDN) brings about the combination of SDN and NFV to provide dynamic, flexible deployment and on-demand scaling of VNFs, which contributes a lot to the development of the mobile packet core towards 5G system [6], [10]. One typical case for NFV/SDN is virtualized Evolved Packet Core (vEPC), and the other is the NG core architecture standardized by 3GPP [11]–[14]. Since the NG core architecture is still in its early stage, most of the research issues on vMME are based on the vEPC based architecture, and few of them focus on the NG core architecture.

vEPC can be categorized into four ways, including a) Fully NFV-based EPC architecture, b) SDN/NFV-based EPC architecture with virtualized data/user plane, c) SDN/NFV-based EPC architecture with non-virtualized data/user plane, d) Fully SDN-based Mobile Packet Core (MPC) architecture [10]. In the above architectures, MME is still one of the function modules coupled in the NFV/SDN architecture instead of vMME.

In [6], EPC as a service is proposed, and four methods are described to deploy EPC in cloud environments which include 1:1 mapping, 1:N mapping, N:1 mapping and N:2 mapping depending on different VNF mapping approaches. vMME is also categorized into 1:1, 1:N, N:1 as well as N:2 mapping. In [15], the feasibility to design a flexible and adaptive mobile core network based on functional decomposition and network slicing is discussed, but how to decompose MME is still an open issue. In [16], the cost of VNF/VNFC deployment strategies is analyzed for a virtualized mobile network infrastructure providing Evolved Packet Core as a Service (EPCaaS) considering functional and administrative constraints. The cost is measured in terms of the utilization of data center infrastructure resources such as computation and networking. The vMME functionality is 1:N mapping, including Signaling Load Balancer (SLB) and Mobility Management Processor (MMP), where the MMP performs the processing task of MME. In [17], different VNF placement algorithms for carrier cloud are introduced aiming at minimizing path between UEs and PDN-GW VNFs of an underlying VNI to ensure acceptable QoE, as well as minimizing S-GW relocation frequency. Some factors are considered including the mobility features, service usage behavioral patterns of mobile users, and mobile operators' cost in terms of the total number of instantiated VNFs to build a Virtual Network Infrastructure (VNI). The algorithms are presented to deal with the optimal placement of VNFs on the user plane. In [18], a practical realization of cellular core called KLEIN is presented as a minimally disruptive design by using vEPC and SDN for service chaining in data centers, with a global resource management scheme to map traffic to different data center locations. The EPC function is mapped as several instances including MME (namely N:2 mapping).

In [19], a decentralized core network architecture optimized for the identified control events is proposed for Machine Type Communication which brings massive control signaling. The proposed control plane functions can be executed separately. The performance for each proposed control function is analyzed by using the main control events that generate signaling messages in the network. In [20], an implemental architecture of a vEPC is proposed to accommodate M2M services. Every vEPC is optimized by eliminating EPC components or replacing standardized interface protocols with internal application interworking. The validity of the proposed architecture is evaluated experimentally from the aspect of CPU resource consumption.

In [21] and [22], 1:N mapping of vMME is investigated, and vMME is decoupled into three functional entities, namely the Front-End (FE), the Worker (W) and the state DataBase (DB). The FE is the communication interface with other network entities and it is also responsible for balancing the load between Ws. W is only responsible for implementing vMME's functional logic and it does not store UE's session state information. The DB stores session state information of users. vMME is formulated as the three-tier queue model by using H2H, M2M service models and priority queues, the average response time of vMME in different scenarios are got [21]. The response time of the vMME is evaluated under different number of users assuming that the three functional entities of the vMME are running in the form of VMs without considering delay of intermediate nodes [22].

From the view point of Service Function Chain(SFC) and VNF embedding, a chained VNFs is defined as a chain-ordered set of service functions (SFs) that handles the traffic of the delivery (data plane), control, and monitoring (control plane) of a specific service/application [23]. Some typical network services such as IMS, load balancing and vCDN are modelled and analyzed from the perspective of SFC in current literatures. In [24], ScalIMS is presented as a management system for IMS which enables dynamic deployment and scaling of VNF service chains across multiple data centers on both control plane and data plane. In [25], two variants of Admission Control/FGE(Forwarding Graph Embedding) problems are investigated, which are formulated by Integer Linear Programming (ILP) and Mixed Integer Linear Programming (MILP). In [26], optimal placement of VNFs for network load balancing are formulated aiming at minimizing the distance from a smaller cluster of servers to the data center in a common cloud and network service approach, and three optimization methods are compared including Linear Programing (LP) model, Genetic Algorithm (GA) and Random Fit Placement Algorithm (RFPA) for the allocation and replication of VNFs.

When considering the 1:N mapping of vMME, and different VNF components of vMME are combined via different ways

to provide MME function and optimized performance with various service requirements, resource constraints and performance metrics, the process of the virtualization of MME can be regarded as a chained VNFs of vMME from the view point of SFC. From the above research issues, although there are some research issues on 1:N mapping of vMME, to the best knowledge of our research, there is no study concentrating on the optimization of vMME based on 1:N mapping and SFC. The focus of this paper is the optimization of vMME by using NFV and SFC for different service scenarios. The main contribution of this paper is as follows:

(1) A MANO-based vMME framework is proposed by using the 1:3 mapping, and a general signaling processing flow of vMME based on SFC is analyzed. (2) On the basis of the virtualized mobility management framework of 1:3 mapping, an optimization method of vMME is proposed, and the performances of vMME is formulated as optimization problems to minimize the total signaling communication overhead cost, the total signaling communication overhead cost on backhauls and the migration overhead cost of state data for different placement of VNF components of vMME. (3) In order to solve the NP-hard optimization problems, a heuristic approach is proposed including algorithms called Min-TSCOC, Min-TSCOCB and Min-MOCSD, and the performances of vMME are evaluated with different function placement solutions and various application scenario parameters.

The rest of this paper is organized as follows: system model for virtual mobility management framework based on MANO is presented in section II. In section III, the performances of vMME based on 1:3 mapping are modelled and optimized. In section IV, the performances of vMME are evaluated, and numerical simulation results are provided. Finally, conclusions and future research work are given in section V.

II. System Model

In this section, a MANO-based virtualized mobility management framework is proposed by using 1:3 mapping. The functions of the three virtualized function components are given. According to the decomposition of components, a general signaling processing flow of vMME based on service function chain is analyzed.

A. Scenario of vMME

Considering the scenario of a heterogeneous radio access network with several MBSs (Macro Base Station) and SBSs (Small Base Station) as shown in Fig.1, the SBSs provide hot spot coverage and the MBSs provide wide area coverage. Either wired link or wireless link is deployed between MBS and SBS for communication on control plane.

Based on the above scenario, mobility management function is deployed by vMME in the core network and radio access network respectively. That is to say, vMME is operated by using VNFCs in the data center, both MBSs and SBSs are equipped with universal servers. Since VNFCs are combined and located at different MBSs and SBSs to support vMME, the communication process on the control plane for mobility management composes service function chains, and some

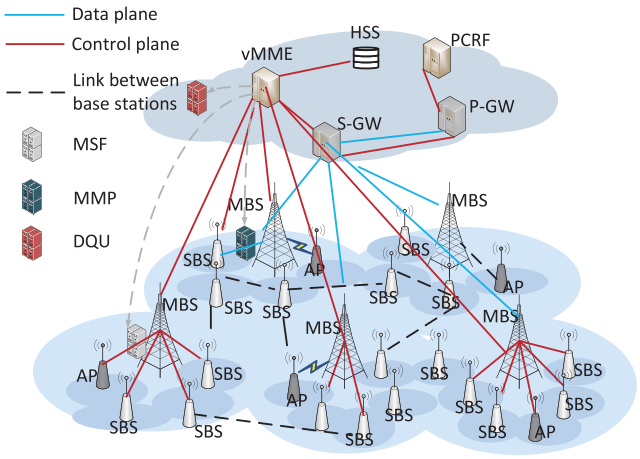

Fig. 1. Scenario of virtualized Mobility Management.

of the communication interaction among VNFCs need one-hop or multi-hop relay by MBS or SBS.

B. The Framework of vMME Based on MANO

The implementation of NFV is standardized by the architecture called MANO [9]. As a typical use case, mobility management can be virtualized as a service in the MANO based architecture [10], and vMME can be managed and optimized through MANO architecture. In the MANO architecture for vMME, NFVO is responsible for the orchestration of mobility management services, including vMME functional component partition, composition optimization as well as performance optimization. Mobility management slices are created and managed to meet different service requirements. The VNF Manager completes function composition optimization based on specified functional component partitioning and resource optimization in order to meet the performance requirements of mobility management in the virtualized environment. The Virtualized Infrastructure Manager (VIM) provide optimal mapping of virtual resources used by mobility management function modules to the physical resources including computation, storage and communication resources. Therefore, functional component partitioning and related optimization of vMME affects the performance of vMME, which is of paramount important to various service scenarios. However, decomposition of MME is a hard work to do.

Mobility management has been decomposed according to the control and data planes [4]. Generally, from the view point of function, the functions of mobility management contain handover management, data management and location management. Location management maintains the reachability of mobile users independent of the UE location or connected network. It has associated an identification database, containing bindings with UEs' identification and its current address. Data management is responsible for encapsulation of data packets through address translation, it does not provide any signaling and it only receives signaling from handover management. Handover management maintains sessions active

when a UE roams between networks, so it provides handover detection and negotiation, being responsible for the signaling that communicates with data management and location management after a handover. Another function of handover management is to maintain the mobility context and routes [27].

Based on the main functions of mobility management, the first step for decomposition of mobility management is to split the functions into control plane and data plane. All procedures related to the MME signaling are on the control plane depending on the interaction among entities related to the mobility management procedures. Taking the mobility management of LTE as an example, the most relevant control plane functions related with mobility includes, (1) Authenticating the UE as it accesses the system; (2) Managing UE state while the users are idle; (3) Supervising handovers between different base stations; (4) Establishing bearers as required by services for connectivity in a mobile context; (5) Generating billing information; (6) Implementing lawful interception policies, and overseeing a large number of features defined in 3GPP specifications [5], [28], [29]. The mobility management events frequently processed on control plane message procedures includes Attach, Idle, Wakeup and Handover [22], [30], and all of them require the interaction of MME with almost all of the entities, including UE and base stations(in the RAN), as well as entities of service gateways(S-GW and P-GW), and HSS(in the core network).

Concerning the above mentioned mobility events of MME, some functions requires interactions with database such as (1) and (2), and some of them can be implemented locally in RAN without database query such as (3), (4) for mMTC. Since the interactions involving database query leads to delay, which derivate the performance of MME, while distributed database brings about synchronization problem of state data of MME, the database query related interaction procedure and the related function placement becomes a tradeoff for the further decouple and design of vMME. Another problem is the hierarchical deployment of cellular radio access network, which aggravates the burden of vMME by the frequent interactions of mobility events on hotspot areas as well as the signaling load on the backhauls if some function components of vMME is located in the core network.

Considering the hierarchical deployment of cellular radio access network, the synchronization problem due to distributed databases, the different mobility features of user equipment as well as the signaling cost on communication links, the second step is decoupling MME into three components including a component which is located nearby RAN nodes, a component responsible for signaling processing and a component for database query. According to the principles of VNFC composition, the different composition patterns of vMME also provides flexible combination solutions of component replacement for various service scenarios. If the number of VNFCs is more than three, on the one hand, the methodology of decoupling MME becomes more complicated, on the other hand, the optimal replacement of decomposed components leads to not only more composition patterns, but also complexity of the synchronization of state data and frequent migration among VNFCs as well.

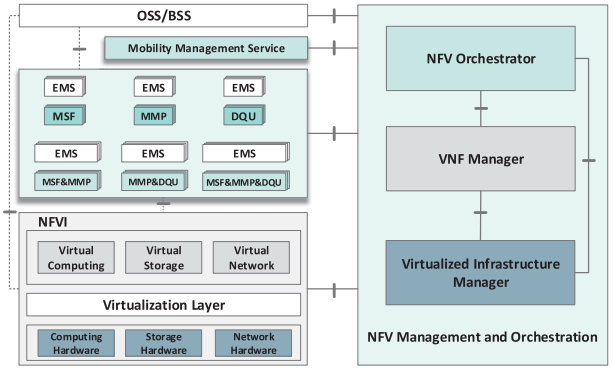

Fig. 2. MANO-based VMME framework with 1:3 mapping.

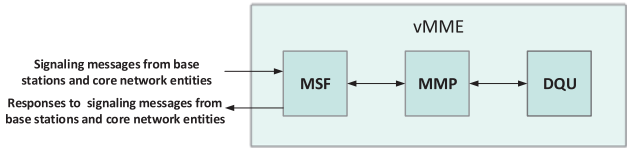

Fig. 3. Virtual Network Functional components of vMME.

Based on the above consideration of function decomposition and signaling interactions, we adopt the 1:3 mapping solution for vMME, and divide the vMME into three virtual components including Mobility Signaling Forwarding (MSF), Mobility Management Processor (MMP) and Data Query Update (DQU). For simplicity of analysis on signaling procedures, the external interfaces of vMME is MSF with BSs and core network (including service gateways such as P-GW and S-GW, HSS), while the internal interfaces are the interfaces of MSF with MMP, and MMP with DQU.

Fig.2 shows a MANO-based vMME framework with 1:3 mapping, and Fig.3 is the decomposition and interface of vMME for 1:3 mapping. The functions of the three VNFCs are described as follows:

(1) MSF: it is responsible for receiving all the mobility management signaling messages from core network and radio access network, and forwarding or replying some of the mobility management signaling messages. It actually acts as the communication interface between vMME and other network entities. Some of the mobility context information of mobile users is stored in MSF.

(2) MMP: It is in charge of implementing mobility management function logic, and mobility context information of mobile users is also stored in MMP. When MMP receives the signaling messages from other network entities forwarded from the MSFs, MMP proceeds the signaling messages according to its functional logic. If the data cannot be obtained in the MMP cache during the process, which are required for mobility management signaling process, then the MMP sends a query request to the DQU, and the required data is obtained for signaling process from the DQU's reply for the query request. When the MMP completes the processing, and the signaling message generated by the MMP needs to be sent to specified network entities, the MMP sends the signaling message to the corresponding MSF, which is used for forwarding the signaling message to that corresponding network entities. If the data in

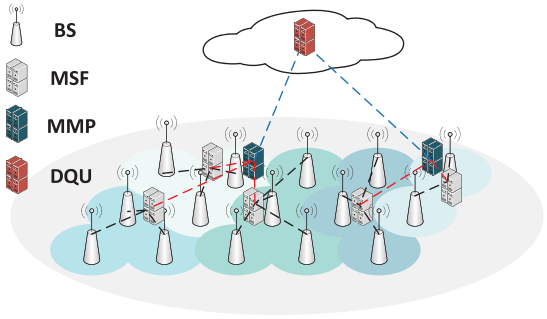

Fig. 4. An example of virtualized function composition and placement for vMME (MMPs and MSFs in RAN).

the database is needed to be updated according to the MMP processing, then the MMP sends an update database request to DQU, and the database update operation would be performed by DQU.

(3) DQU: It is responsible for querying and updating the relevant state information of mobile users, and it also stores some mobility context information for mobile users. When DQU receives the data query request from the MMP, it queries the related data and returns the query result to the MMP. When DQU receives the data updating request from the MMP, it completes the relevant data updating.

Fig.4 shows an example of deployment solutions, in which MMPs and MSFs are placed on several BSs, DQU is deployed in the core network, and MSF, MMP and DQU are connected via wireless or wired links.

From the above different VNF composition case of vMME, it is a challenge to deploy VNFs of vMME taking into consideration different application scenarios and total CAPEX/OPEX of mobility management.

C. Service Function Chain Based Signaling Process

From the respect of optimal deployment of VNFs, the main objective is to achieve fast, scalable and dynamic composition and allocation of VNFs to execute a network service (NS). However, since a NS requires a set of VNFs, two problems are needed to be solved in order to achieve the efficiency of service coordination and management, namely optimal composition of VNFs for a determined NS, and efficiently allocation and scheduling the VNFs of a NS onto a substrate network [23].

vMME can be viewed as a network service. That is to say, the VNF components of vMME can be connected and chained together to execute the function of mobility management. The signaling process of vMME is essentially a process in which VNFCs call each other according to the signaling process flow, therefore, a service function chain is composed based on a signaling process of vMME. In other words, the optimization of vMME is not only related to VNF component optimal placement, but also constrained with the service function chains based on the signaling processes. The service function chain based on the signaling process is optimized with the placement and composition of VNF components of vMME without affecting the integrity of function of mobility management. Therefore, it is significant to investigate and optimize the service function chain based signaling process of vMME.

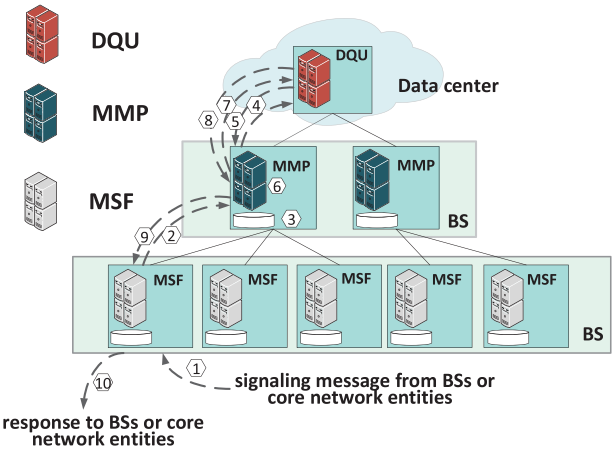

Fig. 5. The general signaling process flow based on 1:3 mapping of vMME.

In mobility management, the events which the vMME processes frequently are Attach, Idle, Wakeup, and Handover [22], [30]. These events are initiated either by UEs or by network, and they are triggered by UEs under specific conditions. Signaling messages are generated and interacted among UE, BSs, vMME as well as other network entities in order to complete the storage, deletion, and update of the relevant state information of UE for mobility management. In order to describe the general signaling process flow of vMME based on 1:3 mapping in detail, an example of vMME including 5 MSFs, 2 MMPs, and 1 DQU is used to support mobility management for UEs. The hierarchical control framework of DQU, MMP and MSF for vMME is illustrated in Fig.5. Five MSFs are placed on five different BSs according to geographical location and connection relationship among BSs, and each MSF manages at least one BS and processes all signaling requests of UEs attached to the BSs managed by the MSF. Two MMPs are placed on two different BSs. Each MMP manages at least one MSF and processes all signaling messages from the MSFs managed by the MMP. The DQU is placed in the data center and processes all database operation requests from the two MMPs. According to the above framework of vMME, the general signaling process flow based on 1:3 mapping of MSF, MMP and DQU is presented as follows:

(1) The MSF receives the signaling message coming from BSs, S-GW, P-GW and HSS as the communication interface between vMME and network entities related mobility management including BSs, S-GW, P-GW and HSS;

(2) The MSF checks the network entity information of the signaling message and forwards it to the corresponding MMP;

(3) The MMP receives the signaling message, and processes it according to its function logic. During the process, the MMP checks whether the data required for processing the signaling message is stored in its cache;

(4) If the data required for processing the signaling message cannot be found in its cache, then the MMP sends a data query request to the DQU to retrieve the required data from the database. During the database retrieval period, the MMP processes other signaling messages;

(5) The DQU processes the data query request, and send back the query result to the MMP;

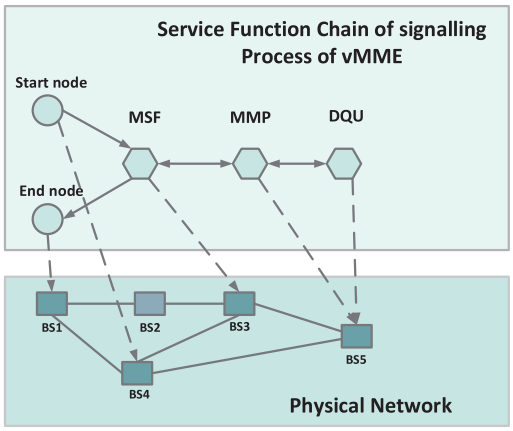

Fig. 6. Mapping from service function chain of vMME to physical network.

(6) When the MMP obtains the required data to process the signaling process, the signaling message is processed;

(7) After the MMP completes the processing of the signaling message, and if some mobility context information in the DQU is needed to be updated, the MMP sends a data update request to the DQU;

(8) When the DQU receives the data update request from the MMP, the DQU completes the data updating according to the data update request, and then sends a data update completion response to the MMP;

(9) As the MMP completes the processing of signaling message, it generates one or more response messages according to the function logic and sends them to the MSF;

(10) After receiving the response messages, the MSF forwards them to the corresponding network entities according to their network entity information in the signaling messages.

The mapping of the VNFC placement and service function chain corresponding to the signaling process to the physical network is described in Fig.6.

Fig.6 reveals that it is really sophisticated to efficiently deploy each VNF component of vMME on the optimal physical nodes by considering service scenario and requirements, the performance of vMME as well as other optimization objectives from mobile users and/or mobile network operators. In order to solve the above problems, we formulate them as optimization problems in the third section.

III. Problem Formulation

In this Section, based on the framework proposed in Section II, an optimization method of vMME is proposed, and the performances of vMME is formulated as optimization problems to minimize the total signaling communication overhead cost, the total signaling communication overhead cost on backhauls and the migration overhead cost of state data for different placement of VNF components of vMME.

The optimization of vMME can be considered from two ways, one is the function optimization, which takes into account the virtual function component partition and the optimal virtual function component composition under specified virtual function component partition result, the other is the performance optimization, which aims at the optimization of performance metrics oriented from MME, mobile users as well as operators on the basis of virtual function component partition and placement.

During the time when mobility management provides service for UEs, the occurrence of four mobility events leads to signaling interaction among communication links of MSF, MMP and DQU, as well as among BSs, S-GW, P-GW and HSS. Since communication resources are constrained, minimizing the total signaling communication overhead cost is a key performance metric to optimize the VNF component placement as well as minimize the delay of mobility management. Meanwhile, state data migration is also a burden to communication resources due to the mobility of UEs and distributed placement of MSFs and MMPs. Therefore, minimizing the total signaling communication overhead cost and the migration overhead cost of state data are chosen as the optimization objectives in the problem formulation.

From the viewpoint of function optimization, in vMME, virtual function component composition and optimization placement plays an important role in the optimization of vMME, and it also contributes to both the total signaling communication overhead cost and the migration overhead cost of state data.

The optimization of vMME is modeled in the scenario of Fig.1, where there are several MBSs, SBSs and UEs. The virtual function components of vMME include MSF, MMP and DQU, which can be deployed on MBSs, SBSs and in the core network. The virtual function components are connected by communication links. Therefore, the virtual function components deployed in the different locations cooperate to provide the function of mobility management.

Assuming there are N_{BS} Small Cells (SC) and there are connections between small cells. As presented in (1), the adjacency matrix $\mathbf{A}$ is used to represent the connection between SCs, where

$$A_{i,j} = \begin{cases} 1 & \text{connection between SC } i \text{ and } j, \\ 0 & \text{no connection between SC } i \text{ and } j. \end{cases} \tag{1}$$

Assuming that $\mathbf{\Omega}$ is the entities which have direct signaling interaction with vMME, including BS, S-GW, P-GW and HSS. Since SGW, PGW and HSS are located in the core network, for simplicity, let CN denotes S-GW, P-GW and HSS, then $\mathbf{\Omega}$ can be presented as in (2),

$$\mathbf{\Omega} = \{BS, CN\} \tag{2}$$

Let $\mathbf{\Phi}$ represent the frequently happened events, which include Attach, Idle, Handover and Wakeup as presented in (3), and λ_φ denotes the arrival rate of event φ, where $\varphi \in \mathbf{\Phi}$.

$$\mathbf{\Phi} = \{Attach, Idle, Handover, Wakeup\} \tag{3}$$

In the signaling process of event φ, there are $N^{\varphi}_{\omega,MSF}$ signaling interaction between entities ω and MSF, $N^{\varphi}_{MSF,MMP}$ signaling interaction between MSF and MMP, and $N^{\varphi}_{MMP,DQU}$ signaling interaction between MMP and DQU.

$U_{i,j}$ is a binary variable and $U_{i,j} = 1$ denotes that UE i is attached to BS j. $MSF_{k,j}$ is a binary variable and

$MSF_{k,j}=1$ denotes that MSF k is placed on BS j. $MMP_{l,j}$ is a binary variable and $MMP_{l,j}=1$ denotes that MMP l is placed on BS j.

Two cases are divided for signaling interaction between entities, namely signaling interaction between vMME and external entities and signaling interaction within vMME internal entities, respectively. Signaling interaction between vMME and external entities include interaction between BS and MSF, interaction between CN and MSF, as well as interaction between CN and DQU. Signaling interaction within vMME internal entities includes interaction between MSF and MMP and interaction between MMP and DQU.

Let $cost_{BS\text{-}MSF}$ be the signaling communication overhead cost between BS and MSF, which denotes the weighted interaction rate of signaling between all BSs and all MSFs, then $cost_{BS\text{-}MSF}$ can be obtained as (4),

$$\begin{aligned} IR_{j,k} &= \sum_{i=1}^{N_{UE}} \sum_{\varphi\in\Phi} U_{i,j}\lambda_{\varphi} N^{\varphi}_{BS,MSF} R_{j,k} \\ cost_{BS\text{-}MSF} &= \sum_{j=1}^{N_{BS}} \sum_{k=1}^{N_{MSF}} \sum_{q=1}^{N_{BS}} IR_{j,k} MSF_{k,q} H_{j,q} \end{aligned} \quad (4)$$

Where $R_{j,k}$ is a binary variable and $R_{j,k}=1$ denotes that BS j is controlled by MSF k. $H_{j,q}$ denotes the hops from BS j to BS q. N_{UE} denotes the number of UEs, N_{BS} is the number of BSs, N_{MSF} denotes the number of MSFs.

Assuming $cost_{MSF\text{-}CN}$ is the signaling communication overhead cost between MSF and CN, which denotes the weighted interaction rate of signaling between all MSFs and CN, similarly, $cost_{MSF\text{-}CN}$ is got as in (5),

$$\begin{aligned} IR_{k,CN} &= \sum_{i=1}^{N_{UE}} \sum_{\varphi\in\Phi} \sum_{j=1}^{N_{BS}} U_{i,j}\lambda_{\varphi} N^{\varphi}_{CN,MSF} R_{j,k} \\ cost_{MSF\text{-}CN} &= \sum_{k=1}^{N_{MSF}} \sum_{q=1}^{N_{BS}} IR_{k,CN} MSF_{k,q} H_{q,CN} \end{aligned} \quad (5)$$

Where $R_{j,k}$ is a binary variable and $R_{j,k}=1$ denotes that BS j is controlled by MSF k. $H_{q,CN}$ denotes the hops from BS q to CN.

The signaling communication overhead cost between DQU and CN is represented as $cost_{DQU\text{-}CN}$, which denotes the weighted interaction rate of signaling between all DQUs and CN, then $cost_{DQU\text{-}CN}$ is obtained as in (6),

$$\begin{aligned} IR_{l}^{DQU,CN} &= \sum_{i=1}^{N_{UE}} \sum_{\varphi\in\Phi} \sum_{j=1}^{N_{BS}} \sum_{k=1}^{N_{MSF}} U_{i,j}\lambda_{\varphi} N^{\varphi}_{DQU,CN} R_{j,k} S_{k,l} \\ cost_{DQU\text{-}CN} &= \sum_{l=1}^{N_{MMP}} \sum_{m=1}^{N_{DQU}} IR_{l}^{DQU,CN} T_{l,m} H_{m,CN} \end{aligned} \quad (6)$$

Where $R_{j,k}$ is a binary variable and $R_{j,k}=1$ denotes that BS j is controlled by MSF k. $S_{k,l}$ is a binary variable and $S_{k,l}=1$ denotes that MSF k is controlled by MMP l. $T_{l,m}$ is a binary variable and $T_{l,m}=1$ denotes that MMP l is controlled by DQU m. $H_{m,CN}$ denotes the hops from DQU m to CN. N_{MMP} denotes the number of MMPs, N_{DQU} denotes the number of DQUs.

Let $cost_{MSF\text{-}MMP}$ be the signaling overhead cost between MSF and MMP, which denotes the weighted interaction rate of signaling between all MSFs and all MMPs, then $cost_{MSF\text{-}MMP}$ can be obtained in (7),

$$\begin{aligned} &IR_{k,l} \\ &= \sum_{i=1}^{N_{UE}} \sum_{\varphi\in\Phi} \sum_{j=1}^{N_{BS}} U_{i,j}\lambda_{\varphi} N^{\varphi}_{CN,MSF} R_{j,k} S_{k,l} \\ &cost_{MSF\text{-}MMP} \\ &= \sum_{q=1}^{N_{BS}} \sum_{r=1}^{N_{BS}} \sum_{k=1}^{N_{MSF}} \sum_{l=1}^{N_{MMP}} IR_{k,l} MSF_{k,q} MMP_{l,r} H_{q,r} \end{aligned} \quad (7)$$

Where $R_{j,k}$ is a binary variable and $R_{j,k}=1$ denotes that BS j is controlled by MSF k. $S_{k,l}$ is a binary variable and $S_{k,l}=1$ denotes that MSF k is controlled by MMP l. $H_{q,r}$ denotes the hops from BS q to BS r.

Assuming that $cost_{MMP\text{-}DQU}$ is the signaling overhead cost between MMP and DQU, which is the weighted interaction rate of signaling between all MMPs and all DQUs, then $cost_{MMP\text{-}DQU}$ is got in (8) as,

$$\begin{aligned} &IR_{l,DQU} \\ &\quad = \sum_{i=1}^{N_{UE}} \sum_{\varphi\in\Phi} \sum_{j=1}^{N_{BS}} \sum_{k=1}^{N_{MSF}} U_{i,j}\lambda_{\varphi} N^{\varphi}_{MMP,DQU} R_{j,k} S_{k,l} \\ &cost_{MMP\text{-}DQU} \\ &\quad = \sum_{r=1}^{N_{BS}} \sum_{l=1}^{N_{MMP}} IR_{l,DQU} MMP_{l,r} H_{r,DQU} \end{aligned} \quad (8)$$

Where $R_{j,k}$ is a binary variable and $R_{j,k}=1$ denotes that BS j is controlled by MSF k. $S_{k,l}$ is a binary variable and $S_{k,l}=1$ denotes that MSF k is controlled by MMP l. $H_{r,DQU}$ is the hops from BS r to DQU.

The total signaling communication overhead cost of all UEs is the sum of signaling communication overhead cost between vMME and external entities, and signaling communication overhead cost within vMME internal entities, which can be presented in (9) as,

$$\begin{aligned} total\text{-}cost = {} & cost_{BS\text{-}MSF} + cost_{MSF\text{-}CN} + cost_{DQU\text{-}CN} \\ & + cost_{MSF\text{-}MMP} + cost_{MMP\text{-}DQU} \end{aligned} \quad (9)$$

Considering the constraints of C1 to C8, the total signaling communication overhead cost should be minimized, which can be formulated as an optimization problem as follows [31]:

$$\begin{aligned} \min \ & total\text{-}cost \\ \text{s.t.} \ & \sum_{q=1}^{N_{BS}} MSF_{k,q} = 1 && (C1) \\ & \sum_{k=1}^{N_{MSF}} R_{j,k} = 1 && (C2) \\ & \sum_{j=1}^{N_{BS}} R_{j,k} \geq 1 && (C3) \\ & \sum_{r=1}^{N_{BS}} MMP_{l,r} = 1 && (C4) \end{aligned}$$

$$\sum_{l=1}^{N_{MMP}} S_{k,l} = 1 \qquad (C5)$$
$$\sum_{k=1}^{N_{MSF}} S_{k,l} \geq 1 \qquad (C6)$$
$$\sum_{m=1}^{N_{DQU}} T_{l,m} = 1 \qquad (C7)$$
$$\sum_{l=1}^{N_{MMP}} T_{l,m} \geq 1 \qquad (C8) \qquad (10)$$

C1 to C3 are constraints for MSF, where C1 is the constraint which each MSF can only be placed on one BS. C2 denotes that each BS can only be controlled by one MSF, while C3 means that the number of BSs controlled by each MSF is not less than one.

C4 to C8 are constraints related to MMP, where C4 denotes that each MMP can only be placed on one BS, C5 denotes that each MSF can only be controlled by one MMP, C6 denotes that the number of MSFs controlled by each MMP is not less than one, C7 denotes that each MMP can only be controlled by one DQU, C8 denotes that the number of MMPs controlled by each DQU is not less than one.

In order to reduce the signaling communication overhead cost on backhauls, we denote the total signaling communication overhead cost of all UEs on backhauls as $backhaul\text{-}cost$, then it is given as (11),

$$backhaul\text{-}cost = cost_{MSF\text{-}CN} + cost_{DQU\text{-}CN} + cost_{MMP\text{-}DQU} \qquad (11)$$

Since the backhaul resource is also constrained, the total signaling communication overhead cost on backhauls is formulated as an optimization problem as follows:

$$\min \ backhaul\text{-}cost$$
$$\text{s.t. } C1, C2, C3, C4, C5, C6, C7, C8 \qquad (12)$$

Assuming that $mig\text{-}cost_{MSF\text{-}MSF}$ is the migration overhead cost of state data between MSF and MSF, which is the weighted migration rate of state data between MSF and MSF, then $mig\text{-}cost_{MSF\text{-}MSF}$ is got in (13) as,

$$mig\text{-}cost_{MSF\text{-}MSF} = \sum_{j=1}^{N_{BS}} \sum_{i=1}^{N_{UE}} U_{i,j} \lambda_{HO} p_{MSF} A_{MSF} H_{MSF} \qquad (13)$$

Where p_{MSF} is the probability of a UE switching from one MSF to another when a switch event occurs, and A_{MSF} denotes the amount of state data migration between MSFs when the above situation occurs. H_{MSF} is the average hops between MSFs under the current deployment strategy.

Let $mig\text{-}cost_{MMP\text{-}MMP}$ be the migration overhead cost of state data between MMP and MMP, which is the weighted migration rate of state data between MMP and MMP, then $mig\text{-}cost_{MMP\text{-}MMP}$ can be obtained as (14),

$$mig\text{-}cost_{MMP\text{-}MMP} = \sum_{j=1}^{N_{BS}} \sum_{i=1}^{N_{UE}} U_{i,j} \lambda_{HO} p_{MMP} A_{MMP} H_{MMP} \qquad (14)$$

Where p_{MMP} is the probability of a UE switching from one MMP to another when a switch event occurs, and A_{MMP} denotes the amount of state data migration between MMPs when the above situation occurs. H_{MMP} is the average hops between MMPs under the current deployment strategy.

Since most of the migration overhead cost of state data comes from the data migrated between MSFs and the data migrated between MMPs due to the mobility of UEs, we denote the migration overhead cost of state data of all UEs as $mig\text{-}cost$, then it is given as (15),

$$mig\text{-}cost = mig\text{-}cost_{MSF\text{-}MSF} + mig\text{-}cost_{MMP\text{-}MMP} \qquad (15)$$

The migration overhead cost of state data is formulated as an optimization problem as follows:

$$\min \ mig\text{-}cost$$
$$\text{s.t. } C1, C2, C3, C4, C5, C6, C7, C8 \qquad (16)$$

The optimization problems in (10), (12) and (16) are general integer linear programming problems, which is a NP hard problem [31]. It is difficult to find an optimized solution in an efficient way when the network size becomes large.

In order to get the optimized solutions of (10), (12) and (16), a heuristic approach is proposed including Min-TSCOC(Minimization on Total Signaling Communication Overhead Cost), Min-TSCOCB(Minimization on Total Signaling Communication Overhead Cost on Backhauls) and Min-MOCSD(Minimization on Migration Overhead Cost of State Data), respectively. The three algorithms are listed.

The function GAboOF(Genetic Algorithm Based on Objective Function) mentioned in the three algorithms is omitted due to limited space.

Algorithm 1 Min-TSCOC

1: **Input:**maxHops_MSF2BS,maxHops_MMP2MSF, network_topology
2: objective←MSF
3: obj_func←1
4: constraints←{C1,C2,C3,maxHops_MSF2BS}
5: deployment_strategy_MSF←GAboOF(objective, obj_func,constraints)
6: objective←MMP
7: obj_func←$total\text{-}cost$
8: constraints←{C4,C5,C6,C7,C8,maxHops_MMP2MSF}
9: deployment_strategy_MMP←GAboOF(objective, obj_func,constraints)
10: **Output:**deployment_strategy_MSF, deployment_strategy_MMP

IV. Performance Evaluation

In this section, the performance of vMME is evaluated under the virtual function component partition including MSF, MMP and DQU. Considering the heterogeneous radio access network with several MBSs and SBSs, according to the reasonable composition for the three virtual function components deployed in radio access network and data center, six feasible

Algorithm 2 Min-TSCOCB

Input:maxHops_MSF2BS,maxHops_MMP2MSF, network_topology
objective←MSF
obj_func←1
constraints←{C1,C2,C3,maxHops_MSF2BS}
deployment_strategy_MSF←GAboOF(objective, obj_func,constraints)
objective←MMP
obj_func←$backhaul\text{-}cost$
constraints←{C4,C5,C6,C7,C8,maxHops_MMP2MSF}
deployment_strategy_MMP←GAboOF(objective, obj_func,constraints)
Output:deployment_strategy_MSF, deployment_strategy_MMP

Algorithm 3 Min-MOCSD

Input:maxHops_MSF2BS,maxHops_MMP2MSF, network_topology
objective←MSF
obj_func←1
constraints←{C1,C2,C3,maxHops_MSF2BS}
deployment_strategy_MSF←GAboOF(objective, obj_func,constraints)
objective←MMP
obj_func←$mig\text{-}cost$
constraints←{C4,C5,C6,C7,C8,maxHops_MMP2MSF}
deployment_strategy_MMP←GAboOF(objective, obj_func,constraints)
Output:deployment_strategy_MSF, deployment_strategy_MMP

cases of vMME function composition placement are presented, which are illustrated as follows:

(1) MSFMMPDQU: MSF, MMP and DQU are placed separately, where MSF and MMP are placed in RAN and DQU is placed in the core network.

(2) MMP&DQUCN: MSF is placed in RAN, and MMP and DQU are merged and placed in the core network.

(3) MMP&DQUAN: MSF is placed in RAN, MMP and DQU are merged and placed in RAN, but MSF is placed separately from MMP&DQU.

(4) MSF&MMPAN: MSF and MMP are merged and placed in RAN, and DQU is placed independently in the core network.

(5) MSF&MMP&DQUCN: MSF, MMP and DQU are merged and placed in the core network.

(6) MSF&MMP&DQUAN: MSF, MMP and DQU are merged and placed in RAN.

Since the optimal solution of vMME depends on the service scenarios, and the mobility feature of UEs is different under different service scenarios, one of the feasible ways to describe the mobility of UEs is the difference of four mobility management events measured from realistic networks. We select four typical application scenarios including vehicular network, shared bicycle, IoT(smart home) and IoT(smart meter) as the application cases, and give different parameter settings according to their typical mobility features extracted and calculated from [22].

The performance metrics in the simulation are defined as follows:

(1) Total signaling communication overhead cost per UE is the average of the sum of signaling communication overhead cost between vMME and external entities and signaling communication overhead cost within vMME internal entities for every UE;

(2) Total signaling communication overhead cost on backhauls per UE is the average of the signaling communication overhead cost on backhauls for every UE;

(3) Migration overhead cost of state data per UE is the average of the sum of migration overhead cost of state data between MSFs and migration overhead cost of state data between MMPs for every UE.

The optimization approach of vMME is evaluated including algorithms of Min-TSCOC, Min-TSCOCB and Min-MOCSD.

The optimization of vMME is evaluated by MATLAB with Monte Carlo method. The radio access network including SBS and MBS is generated by BA model [32] considering different degrees of nodes of SBS and MBS. Based on the network generated, the placement of MSF located at MBS and SBS are got according to the distribution of UEs and the constrained hops from UE to the MSF (The constrained hops is 2 in the simulation). The placement of MMP is determined by choosing the nodes within one hop with the MSFs. Therefore, the instances of MSF, MMP and DQU varies with the network topology including the number of the BSs, the degrees of each node, the parameters of UE distribution, the constrained hops between UE and MSF, the constrained hops between MSF and MMP as well as the placement of DQU.

With the initialized number of the instances of MSF, MMP and DQU, the optimal placement and number of MSF, MMP and DQU are optimized with different object functions including minimizing the total signaling communication overhead cost, the total signaling communication overhead cost on backhauls as well as the migration overhead cost of state data among MSF, MMP and DQU under different cases of vMME function composition placement and application scenarios, respectively.

A. Simulation Parameters

All of the parameters in the simulation include the networking parameters of heterogeneous radio access network in the simulation, the number of signaling interactions between different entities, the arrival rate of events for mobility management, the arrival rate of various events in different service scenarios, the size of state data migration that occurs between entities, as well as the probability of state data migration that occurs between VNF component entities.

In order to evaluate the performance of vMME in a scalable network, and at the same time considering the computational complexity of the algorithms, the number of nodes (including SBS and MBS) is 50 in the simulation. The number of UEs in

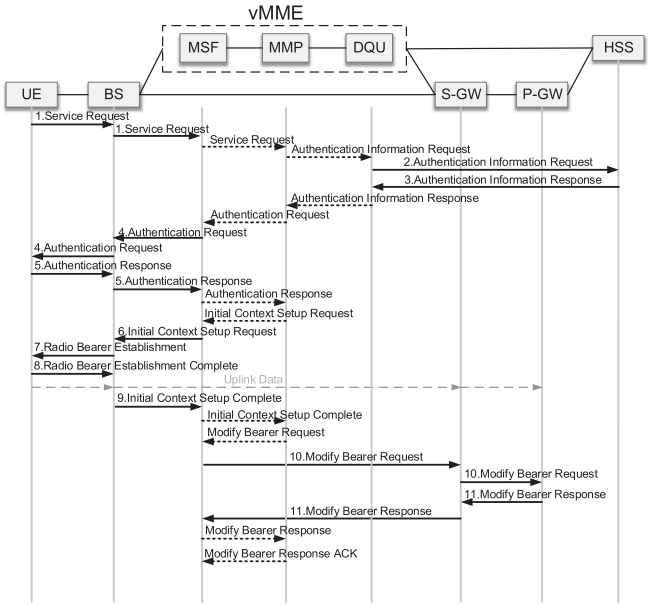

Fig. 7. The flow chart of signaling interaction of Wakeup.

TABLE I
THE NUMBER OF SIGNALING INTERACTIONS AMONG ENTITIES

Name of events	Attach	Idle	Wakeup	Handover
MSF-BS	5	3	5	2
MSF-CN	2	2	2	2
MSF-MMP	6	2	8	2
MMP-DQU	4	0	2	0
DQU-CN	4	0	2	0

each base station follows a uniform distribution of 5 to 20 [33]. In order to differentiate the weight of link type (link type of RAN and that of core network), the parameter of each hop to the core network is 5, and each hop in RAN is 1.

According to the flow chart of signaling interaction between vMME and other entities in the core network and radio access network, the number of signaling interactions of virtual function components entities are analyzed by using the concept of service function chain. Taking the event of Wakeup as a use case, Fig.7 shows the flow chart of signaling interaction of Wakeup according to [28] and [29].

One of the optimal results of the number of signaling interactions between entities of MSF, MMP and DQU is given in Table I.

The arrival rate of four kinds of mobility management event follows Poisson distribution denoted as P(λ) (procedures/s), which are extracted from [22] as: the arrival rate of Idle is P(4.5×10^{-3}), the arrival rate of Wakeup is P(4.5×10^{-3}), the arrival rate of Handover is P(1.2×10^{-3}), the arrival rate of Attach is set as P(1.2×10^{-5}).

Four application scenarios of vMME are considered, which are vehicular network, shared bicycle, IoT (smart home) and IoT(smart meter). The parameters are calculated assuming that the average intervals of the four events of Vehicular network are 30 min for Attach, 1 min for Idle, 1 min for Wakeup and 0.6 min for Handover; the intervals of the four events of Shared bicycle are 10 min for Attach, 10 min for Idle, 10 min for Wakeup and 3.6 min for Handover; the intervals of the four events of IoT (Smart home) are 5 min for Attach, 1 min for Idle, 1 min for Wakeup and no Handover events; and the intervals of the four events of IoT (Smart meter) are 14 days for Attach, 14 days for Idle, 14 days for Wakeup and no Handover events. The diameter of the small cell is 600 m, the speed of the Shared bicycle is 10 km/h and the speed of the Vehicular network is 60 km/h. The arrival rate of mobility management events in these application scenarios are calculated and listed in TABLE II.

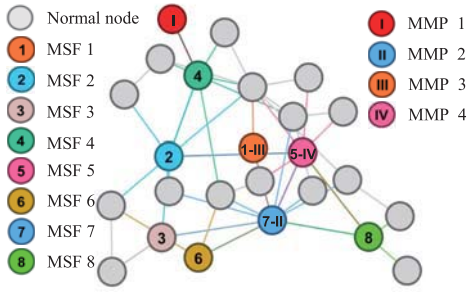

Fig. 8. The MSF, MMP and DQU initially located in the network.

When the UE switches from one MSF to another MSF due to its mobility, the migration of state data of UE is necessary, for example, the state data of the UE is migrated from one MSF to another MSF and the state data of the UE is migrated from one MMP to another MMP. According to [5], the state data migration of the UE is usually only a few KB. Considering the diversity of future services as well as the overhead cost due to virtualization of functional components [34], the state data migration of the UE would be more than a few KB, therefore, the size of state data migration between different VNFCs are set as follows: N_{MSF}=20 KB, N_{MMP}=40 KB, $N_{MSF\&MMP}$=60 KB, $N_{MMP\&DQU}$=60 KB and $N_{MSF\&MMP\&DQU}$=80 KB.

The probability of state data migration between different VNFCs is given, namely, p_{MSF}=0.2, p_{MMP}=0.1, $p_{MSF\&MMP}$=0.2, $p_{MMP\&DQU}$=0.1 and $p_{MSF\&MMP\&DQU}$=0.2.

B. Performance Evaluation

In this subsection, the impact of UE and networking of radio access network to the location and number of instances of MSF, MMP and DQU are presented in *1)*. The impact of the arrival rate of four mobility management events to the total signaling communication overhead cost and the total signaling communication overhead cost on backhauls are investigated in *2)* and *3)*. The impact of vMME function composition to the total signaling communication overhead cost, the total signaling communication overhead cost on backhauls and the migration overhead cost of state data are presented in *4)* and *5)*.

1) Network Generated Based on Complex Network and Placement of Instances: The network topology of radio access network is generated based on complex network so as to

TABLE II
THE ARRIVAL RATE OF FOUR MOBILITY MANAGEMENT EVENTS UNDER DIFFERENT SERVICE SCENARIOS

Service scenarios	Attach	Idle	Wakeup	Handover
Vehicular network	P(5.6×10^{-4})	P(16.7×10^{-3})	P(16.7×10^{-3})	P(27.8×10^{-3})
Shared bicycle	P(1.7×10^{-3})	P(1.7×10^{-3})	P(1.7×10^{-3})	P(4.6×10^{-3})
IoT(Smart home)	P(3.3×10^{-3})	P(16.7×10^{-3})	P(16.7×10^{-3})	0
IoT(Smart meter)	P(8.3×10^{-7})	P(8.3×10^{-7})	P(8.3×10^{-7})	0

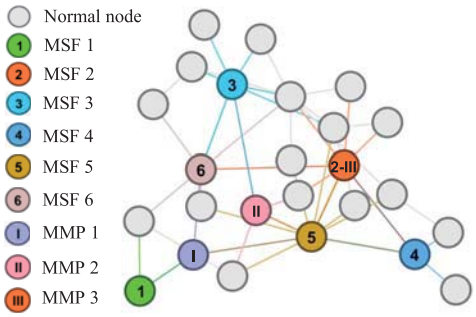

Fig. 9. The optimal placement result of MSF, MMP after optimization.

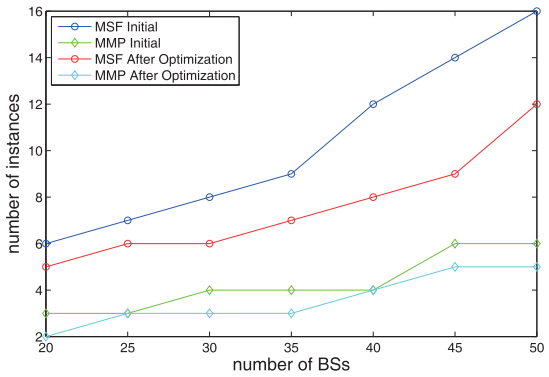

Fig. 10. The impact of the network size to the instances of MSF and MMP.

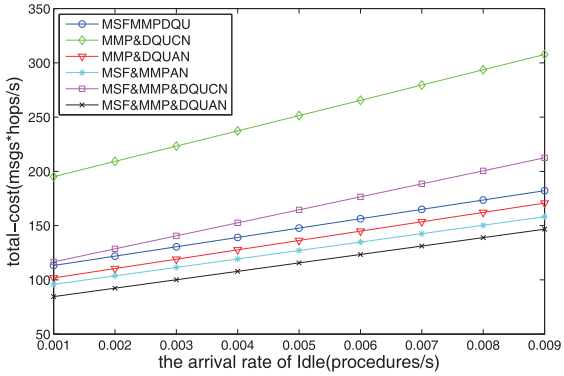

Fig. 11. The impact of the arrival rate of Idle event to *total-cost*.

evaluate the performances of vMME, in which the connection relation of BSs follows BA model.

In the simulation (section *2)*, *3)*, *4)*, *5)* and C), the number of nodes in the network is 50, and the average degree value is 2. With the generated radio access network, the number of VNFCs of vMME including MSF and MMP are initiated as 16 and 6, respectively. With the initiated number of the instances of VMME components, the optimization process is executed, and the number of instances of MSF and MMP are 12 and 5 after optimization algorithm of Min-TSCOC.

In order to evaluate the impact of network size to the number of instances of MSF and MMP, Fig.8 is the simulation result of the initiated placement of MSF and MMP when the number of nodes is 25, and Fig.9 is the optimal placement result of MSF, MMP and DQU after optimization. In Fig.10, the impact of network size (number of nodes) to the instances of MSF and MMP is given when the average degree value is 2.

2) Impact of the Arrival Rate of Four Events to Total-Cost: Fig.11, Fig.12, Fig.13 and Fig.14 show the impact of the arrival rate of four mobility management events on the total signaling communication overhead cost when the arrival rate of one event changes while the arrival rate of other events takes the default value. The total signaling communication overhead cost is the sum of the signaling communication overhead cost of the radio access network and the signaling communication overhead cost of the core network.

The total signaling communication overhead cost increases with the increasing of the arrival rate of the event in each figure. The *total-cost* of MMP&DQUCN is much larger than that of other cases since there is both more frequent signaling interaction between MSF and MMP and more hops between RAN and the core network. The *total-cost* of MSF&MMP&DQUAN is the minimal among all the cases since MSF, MMP and DQU are merged and placed in RAN and there is no signaling communication overhead cost among the three vMME function component entities. The difference of the *total-cost* of the remaining four cases is not obvious, but the *total-cost* of MSF&MMP&DQUCN (the existing mobility management solution) is larger, indicating that the 1:3 mapping method can reduce the *total-cost*. MSF&MMP&DQUAN seems to have the least *total-cost*, however, more state data migration between VNFC instances is needed, which is a tradeoff between the *total-cost* and the cost for state data migration among VNFCs.

3) Impact of the Arrival Rate of Four Events to Backhaul-Cost: The impact of the arrival rate of four mobility management events to the total signaling communication overhead cost on backhauls are given in Fig.15, Fig.16, Fig.17 and Fig.18. The total signaling communication overhead cost on backhauls increases with the increasing of the arrival rate of the events in each figure.

The curve of MSFMMPDQU and that of MSF&MMPAN coincide in the four figures. The reason is that since MSF, MMP and DQU are deployed in RAN, the signaling

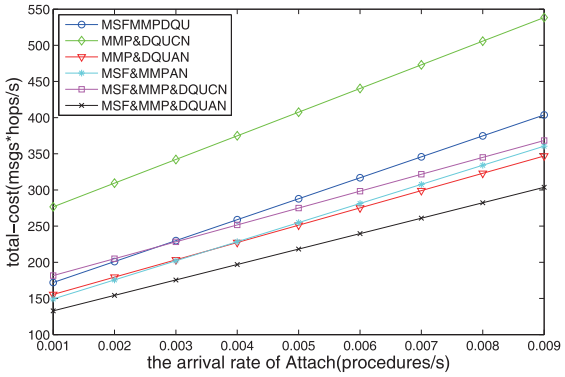

Fig. 12. The impact of the arrival rate of Attach event to *total-cost*.

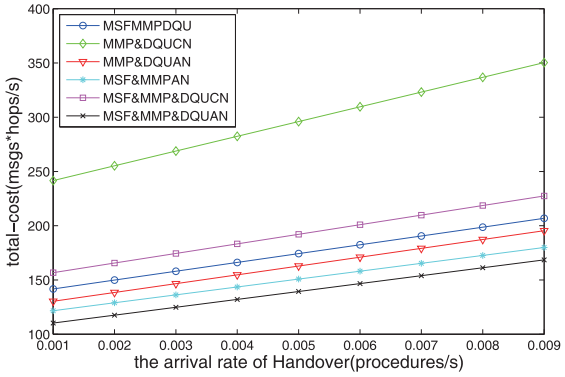

Fig. 13. The impact of the arrival rate of Handover event to *total-cost*.

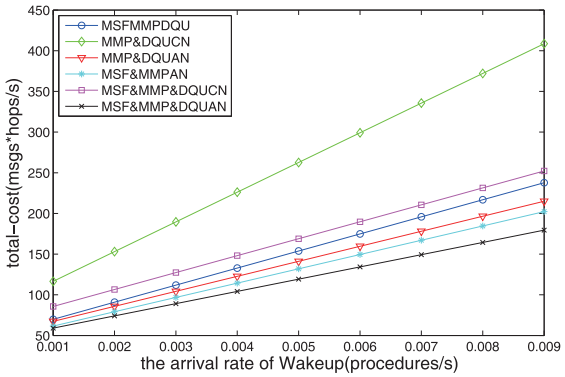

Fig. 14. The impact of the arrival rate of Wakeup event to *total-cost*.

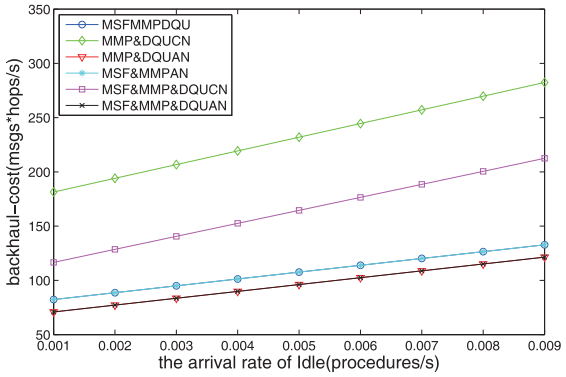

Fig. 15. The impact of the arrival rate of Idle event to *backhaul-cost*.

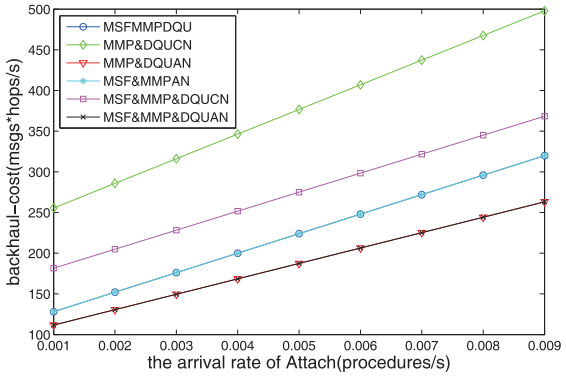

Fig. 16. The impact of the arrival rate of Attach event to *backhaul-cost*.

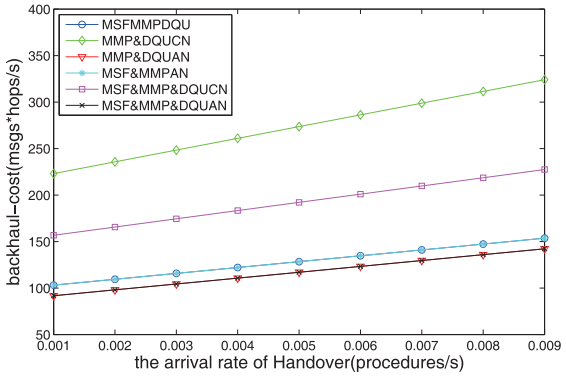

Fig. 17. The impact of the arrival rate of Handover event to *backhaul-cost*.

interaction between the MSF and the MMP only generates signaling communication overhead cost in RAN, and has nearly no effect on traffic load to backhauls. The curve of MMP&DQUAN and that of MSF&MMP&DQUAN also coincide because the signaling interaction between the MSF and the MMP has no impact to backhauls. The *backhaul-cost* of MMP&DQUCN is much larger than that of other cases since there is more frequent signaling interaction between MSF and MMP as well as more hops between RAN and the core network. The *backhaul-cost* of MSF&MMP&DQUAN and that of MMP&DQUAN are smaller because there is less signaling interaction between the DQU and the core network.Compared with existing mobility management schemes(namely MSF&MMP&DQUCN), the pattern of MSFMMPDQU, MSF&MMPAN, MMP&DQUAN and MSF&MMP&DQUAN can significantly reduce the *backhaul-cost*. The *backhaul-cost* of MSF&MMP&DQUAN reduces to almost half of the *backhaul-cost* compared with MSF&MMP&DQUCN with the increase of four mobility events.

4) Impact of vMME Function Composition to Total-Cost and Backhaul-Cost: The impact of vMME function composition to *total-cost* per UE in different application scenarios is evaluated in Fig.19. The *total-cost* for vehicular application scenario is the largest, while the *total-cost* of IoT (smart meter) is the smallest, since the arrival rate of four mobility management events for vehicular application scenario is much greater than that of the arrival rate of four events for IoT (smart meter). The *total-cost* of MMP&DQUCN is the largest and the *total-cost* of MSF&MMP&DQUAN is

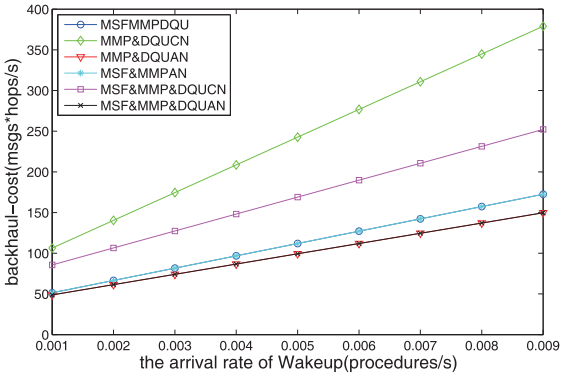

Fig. 18. The impact of the arrival rate of Wakeup event to *backhaul-cost*.

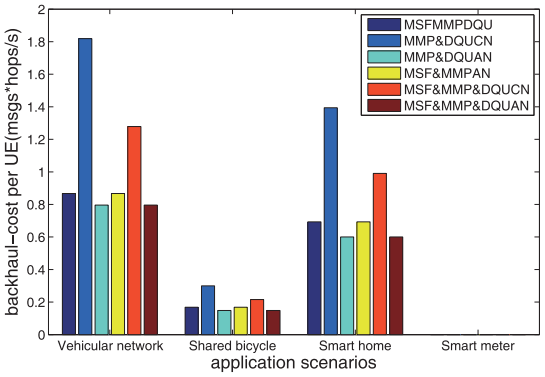

Fig. 20. The impact of vMME function composition to *backhaul-cost* per UE.

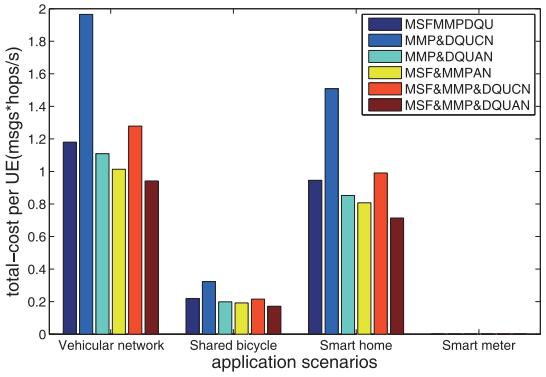

Fig. 19. The impact of vMME function composition to *total-cost* per UE.

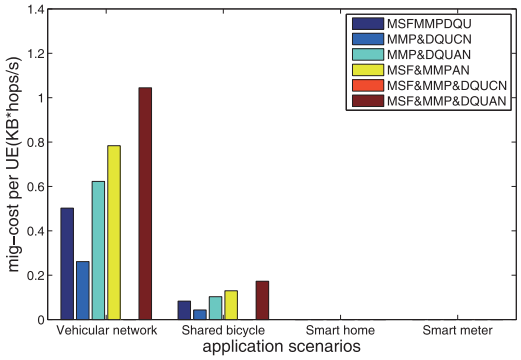

Fig. 21. The impact of vMME function composition to *mig-cost* per UE.

the smallest. From the perspective of total signaling communication overhead cost, MSF&MMP&DQUAN has better performance than that of other cases in different application scenarios.

The simulation results of the impact of vMME function composition to *backhaul-cost* in four application scenarios are given in Fig.20. Since the arrival rate of four mobility management events for vehicular application scenario is much greater than that of IoT(smart meter), the *backhaul-cost* for vehicular application scenario is the largest. The *backhaul-cost* for IoT(smart meter) is the smallest. The *backhaul-cost* of MMP&DQUCN is the largest and the *backhaul-cost* of MSF&MMP&DQUAN is the smallest. From the view point of the total signaling communication overhead cost on backhauls, MSF&MMP&DQUAN outperforms than that of other cases in different application scenarios.

From Fig.19 and Fig.20, the *total-cost* and the *backhaul-cost* of Smart Home is higher compared to those restrictive mobility application scenarios such as Shared Bicycle, which would be taken into consideration for vMME procedure optimization due to its frequent mobility management events.

5) Impact of vMME Function Composition to Mig-Cost: Fig.21 reveals the effect of vMME function composition on *mig-cost*. In scenarios of smart home and smart meter, most of the mobile equipment has nearly no mobility events of handover, therefore they do not generate migration overhead cost of state data. From the simulation results, the *mig-cost* of MSF&MMP&DQUAN is the largest since the switching probability of UE in the case of MSF&MMP&DQUAN is large and the amount of state data that needs to be migrated are large. the *mig-cost* of MSF&MMP&DQUCN is the smallest because MSF&MMP&DQU is placed in the core network and UEs need not switch among MSF&MMP&DQU although when they are roaming in a wider area, it is actually the existing solution of mobility management deployed in the core network.

From the simulation results, it reveals that the composition of MMP&DQUCN is a candidate solution because of its low migration cost compared with the composition of MMP&DQUAN and MSF&MMPAN.

C. Complexity of the Optimization Approach

From the formulation of the optimization problems, the computational complexity of traversing all feasible solutions is $O(n^k)$, and the computation complexity depends on the number of base stations and the number of MSFs, MMPs and DQUs. That is to say, for an example, when the number of base stations is set to 50 and the number of MSFs, MMPs and DQUs are set to 12, 5 and 1 respectively, the computation times to traversing all feasible solutions is 10^{91}. Obviously, it becomes a bottleneck to get the solutions to the optimization problems when considering scalability of network. By using the proposed heuristic approach, the computational complexity is reduced to $O(n^2)$.

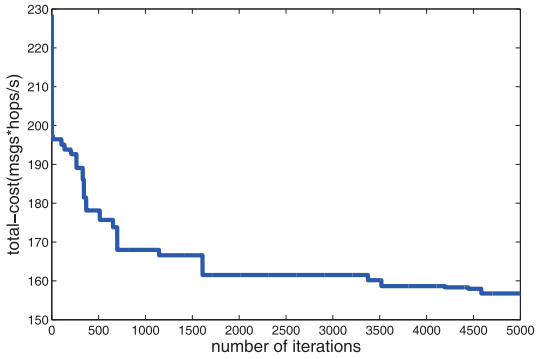

Fig. 22. The computation complexity of the algorithm of Min-TSCOC.

Taking the algorithm of Min-TSCOC as a use case, Fig.22 shows the impact of the number of iterations to the total signaling communication overhead cost, where the arrival rate of four events takes the default value, the number of base stations is set to 50 and the constrained hops between UE and MSF is set to 2. It indicates that the heuristic algorithm has converged when the number of iterations reaches 5000.

V. Conclusion

In this paper, the optimization of MME is investigated by using network function virtualization and service function chain for 1:3 function mapping of vMME. A general signaling processing flow of vMME based on service function chain is analyzed. The performance of vMME is formulated as optimization problems to minimize the total signaling communication overhead cost, the total signaling communication overhead cost on backhauls and the migration overhead cost of state data for different function component placement of vMME. To solve the NP-hard problem, a heuristic approach is proposed including algorithms of Min-TSCOC, Min-TSCOCB and Min-MOCSD. The approach is evaluated by simulation under various composition of vMME and service scenarios. The simulation results reveals that MSF&MMP&DQUAN is an alternative solution for reducing $total\text{-}cost$ and $backhaul\text{-}cost$. From the aspect of migration cost, the composition of MMP&DQUCN is an alternative solution because of its low migration cost compared with the composition of MMP&DQUAN and MSF&MMPAN. The Smart Home leads to large $total\text{-}cost$ and $backhaul\text{-}cost$ because of its frequent mobility management events, which becomes a challenge for the optimal design of mobility management architecture and procedures.

Regarding future work, several challenges lie ahead. The proposed optimization approach is concentrated on the vEPC based architecture, while the NG architecture separates the current EPC functions into more fine-granular modular network functions which can be composed to build a control plane service tailored to a network slice [11], [12]. Therefore, the first challenge is to analyze the mobility management procedures based on the NG architecture. The second is considering the delay of intermediate nodes in the modelling. In addition, from the view point of virtualized instances of vMME, although we have considered the impact of UEs and networking to the number of instances in the simulation, it is still a challenge to formulate the auto-scaling of instances of vMME from the perspective of computation resource in the future.

References

[1] E. Dahlman *et al.*, "5G wireless access: Requirements and realization," *IEEE Commun. Mag.*, vol. 52, no. 12, pp. 42–47, Dec. 2014.

[2] H.-J. Einsiedler, A. Gavras, P. Sellstedt, R. Aguiar, R. Trivisonno, and D. Lavaux, "System design for 5G converged networks," in *Proc. IEEE Eur. Conf. Netw. Commun. (EuCNC)*, Paris, France, Jun. 2015, pp. 391–396.

[3] B. M. Masini, A. Bazzi, and E. Natalizio, "Radio access for future 5G vehicular networks," in *Proc. IEEE 86th Veh. Technol. Conf. (VTC-Fall)*, Toronto, ON, Canada, Sep. 2017, pp. 1–7.

[4] *Distributed Mobility Management Deployment Scenario and Architecture*, document draft-liu-dmm-deployment-scenario-02, IETF, 2015.

[5] X. An, F. Pianese, I. Widjaja, and U. G. Acer, "DMME: Virtualizing LTE mobility management," in *Proc. IEEE 36th Conf. Local Comput. Netw.*, Bonn, Germany, Oct. 2011, pp. 528–536.

[6] T. Taleb *et al.*, "EASE: EPC as a service to ease mobile core network deployment over cloud," *IEEE Netw. Mag.*, vol. 29, no. 2, pp. 78–88, Mar. 2015.

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

[8] R. Mijumbi, J. Serrat, J.-L. Gorricho, N. Bouten, F. De Turck, and R. Boutaba, "Network function virtualization: State-of-the-art and research challenges," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 1, pp. 236–262, 1st Quart., 2016.

[9] *Network Functions Virtualisation (NFV); Management and Orchestration; Report on Policy Management in MANO; Release 3*, document ETSI GR NFV-IFA 023 V3.1.1, 2017.

[10] V.-G. Nguyen, A. Brunstrom, K.-J. Grinnemo, and J. Taheri, "SDN/NFV-based mobile packet core network architectures: A survey," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 3, pp. 1567–1602, 3rd Quart., 2017.

[11] I. Afolabi, T. Taleb, K. Samdanis, A. Ksentini, and H. Flinck, "Network slicing and softwarization: A survey on principles, enabling technologies, and solutions," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 2429–2453, 3rd Quart., 2018, doi: 10.1109/COMST.2018.2815638.

[12] K. Mahmood, T. Mahmoodi, R. Trivisonno, A. Gavras, D. Trossen, and M. Liebsch, "On the integration of verticals through 5G control plane," in *Proc. IEEE Eur. Conf. Netw. Commun. (EuCNC)*, Oulu, Finland, Jun. 2017, pp. 1–5.

[13] *Technical Specification Group Services and System Aspects; System Architecture for the 5G System; Stage 2 (Release 15)*, 3GPP TS Standard 23.501, 2018.

[14] *Technical Specification Group Services and System Aspects; Procedures for the 5G System; Stage 2 (Release 15)*, 3GPP TS Standard 23.502, 2018.

[15] M. R. Sama, X. An, Q. Wei, and S. Beker, "Reshaping the mobile core network via function decomposition and network slicing for the 5G era," in *Proc. IEEE Wireless Commun. Netw. Conf.*, Doha, Qatar, Apr. 2016, pp. 90–96.

[16] F. Z. Yousaf, P. Loureiro, F. Zdarsky, T. Taleb, and M. Liebsch, "Cost analysis of initial deployment strategies for virtualized mobile core network functions," *IEEE Commun. Mag.*, vol. 53, no. 12, pp. 60–66, Dec. 2015.

[17] T. Taleb, M. Bagaa, and A. Ksentini, "User mobility-aware virtual network function placement for virtual 5G network infrastructure," in *Proc. IEEE Int. Conf. Commun. (ICC)*, London, U.K., Jun. 2015, pp. 3879–3884.

[18] Z. A. Qazi, P. K. Penumarthi, V. Sekar, V. Gopalakrishnan, K. Joshi, and S. R. Das, "KLEIN: A minimally disruptive design for an elastic cellular core," in *Proc. Symp. SDN Res. (SOSR)*, Santa Clara, CA, USA, Mar. 2016, pp. 1–12.

[19] A. Roozbeh, "Distributed cloud and de-centralized control plane: A proposal for scalable control plane for 5G," in *Proc. IEEE/ACM. 8th Int. Conf. Utility Cloud Comput. (UCC)*, Limassol, Cyprus, Dec. 2015, pp. 348–353.

[20] H. Baba, M. Matsumoto, and K. Noritake, "Lightweight virtualized evolved packet core architecture for future mobile communication," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, New Orleans, LA, USA, Mar. 2015, pp. 1811–1816.

[21] P. Andres-Maldonado, P. Ameigeiras, J. Prados-Garzon, J. J. Ramos-Munoz, and J. M. Lopez-Soler, "MME support for M2M communications using network function virtualization," in *Proc. IEEE 12th Adv. Int. Conf. Telecommun. (AICT)*, Valencia, Spain, May 2016, pp. 106–111.
[22] J. Prados-Garzon, P. Ameigeiras, J. J. Ramos-Munoz, P. Andres-Maldonado, and J. M. Lopez-Soler, "Analytical modeling for virtualized network functions," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, Paris, France, May 2017, pp. 979–985.
[23] A. M. Medhat, T. Taleb, A. Elmangoush, G. A. Carella, S. Covaci, and T. Magedanz, "Service function chaining in next generation networks: State of the art and research challenges," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 216–223, Feb. 2017.
[24] J. Duan, C. Wu, F. Le, A. X. Liu, and Y. Peng, "Dynamic scaling of virtualized, distributed service chains: A case study of IMS," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 11, pp. 2501–2511, Nov. 2017.
[25] M. A. T. Nejad, S. Parsaeefard, M. A. Maddah-Ali, T. Mahmoodi, and B. H. Khalaj, "vSPACE: VNF simultaneous placement, admission control and embedding," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 3, pp. 542–557, Mar. 2018.
[26] F. Carpio, S. Dhahri, and A. Jukan, "VNF placement with replication for Load balancing in NFV networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Paris, France, May 2017, pp. 1–6.
[27] T. Condeixa and S. Sargento, "Studying the integration of distributed and dynamic schemes in the mobility management," *Comput. Netw.*, vol. 60, no. 2, pp. 46–59, Feb. 2014.
[28] *Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall Description; Stage 2 (Release 15)*, 3GPP TS Standard 36.300, 2018.
[29] *Technical Specification Group Services and System Aspects; General Packet Radio Service (GPRS) Enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) Access (Release 15)*, 3GPP TS Standard 23.401, 2018.
[30] H. Lindholm, L. Osmani, H. Flinck, S. Tarkoma, and A. Rao, "State space analysis to refactor the mobile core," in *Proc. ACM 5th Workshop Things Cellular, Oper., Appl. Challenges*, London, U.K., Aug. 2015, pp. 31–36.
[31] C. Baolin, *Theory and Algorithms for Optimization*, 2nd ed. Beijing, China: Tsinghua Univ. Press, 2005.
[32] M. E. J. Newman, "The structure and function of complex networks," *SIAM Rev.*, vol. 45, no. 2, pp. 167–256, Jun. 2003.
[33] N. Golrezaei, K. Shanmugam, A. G. Dimakis, A. F. Molisch, and G. Caire, "FemtoCaching: Wireless video content delivery through distributed caching helpers," in *Proc. IEEE INFOCOM*, Orlando, FL, USA, Mar. 2012, pp. 1107–1115.
[34] P. Yu, X. Ma, J. Cao, and J. Lu, "Application mobility in pervasive computing: A survey," *Pervasive Mobile Comput.*, vol. 9, no. 1, pp. 2–17, Jul. 2013.

Hao Jin received the Ph.D. degree from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 1996. She is currently an Associate Professor with BUPT. Her research interests include future network architecture, optimization of mobile wireless communication, mobile edge computing, and data mining.

Yi Jin received the B.Eng. degree from Northeastern University at Qinhuangdao in 2016. He is currently pursuing the master's degree with the Key Laboratory of Universal Wireless Communications, Ministry of Education, Beijing University of Posts and Telecommunications, Beijing, China. His current research interests include mobility management in mobile network, mobile edge computing, and NFV.

Haiya Lu received the B.Eng. degree from the Nanjing University of Posts and Telecommunications in 2016. He is currently pursuing the master's degree with the Key Laboratory of Universal Wireless Communications, Ministry of Education, Beijing University of Posts and Telecommunications, Beijing, China. His current research interests include mobile caching, optimization on mobile edge computing, and NFV.

Chenglin Zhao received the bachelor's degree in radio technology from Tianjin University in 1986, and the master's degree in circuits and systems and the Ph.D. degree in communication and information system from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 1993 and 1997, respectively. He is currently a Professor with BUPT. His current research interests include emerging technologies of short-range wireless communication, cognitive radios, 60 GHz millimeter-wave communications, and Internet of Things.

Mugen Peng (M'05–SM'11) received the Ph.D. degree in communication and information systems from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2005. He has been a Full Professor with BUPT since 2012. In 2014, he was an Academic Visiting Fellow with Princeton University, USA. He has authored and co-authored over 390 papers. His main research areas include wireless communication theory, radio signal processing, cooperative communication, self-organization networking, cloud communication, and Internet of Things. He received the First Grade Award of Technological Invention Award three times in China. He was a recipient of the 2014 IEEE ComSoc AP Outstanding Young Researcher Award, WCNC 2015, the Best Paper Award in JCN 2016, and the 2018 Heinrich Hertz Prize Paper Award. He is currently or has been on the Editorial/Associate Editorial Board of the *IEEE Communications Magazine*, the IEEE ACCESS, the IEEE INTERNET OF THINGS JOURNAL, and *IET Communications*.