IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 13, 2019, accepted April 25, 2019, date of publication May 22, 2019, date of current version June 10, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2918282

IVCN: Information-Centric Network Slicing Optimization Based on NFV in Fog-Enabled RAN

HAO JIN, HAIYA LU, YI JIN, AND CHENGLIN ZHAO

Key Laboratory of Universal Wireless Communications, Ministry of Education, Beijing University of Posts and Telecommunications, Beijing 100876, China

Corresponding author: Hao Jin (hjin@bupt.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61471062 and Grant 61431008, and in part by the State Major Science and Technology Special Projects under Grant 2017ZX03001014.

ABSTRACT Content service has become the most popular service that occupies plenty of caching and networking resources. In order to deliver content to end users with high QoE and low infrastructure cost, network function virtualization is deemed as a promising solution for operators to provide content service with geographically dispersed nodes in large scale, which paves the way towards the virtualized content network. In this paper, from the point of view of information-centric networking and service provision, an information-centric virtual content network (IVCN) slicing framework is proposed. A method on the VNF partition from the point of view of information-centric networking and service provision is put forward. The service function chaining and optimization of IVCN are investigated both on the data plane and on the control plane. The IVCN slicing is modeled as a use case in heterogeneous fog-enabled RAN. A method of performance optimization on the IVCN slicing is proposed based on SFC, which aims to optimize the mapping of virtual network functions and virtual content placement. A heuristic approach called IVCN-RANO is proposed to solve the NP-hard problem based on ant colony optimization algorithm. The performance of the IVCN-RANO is evaluated comparing with CEE-LRU, Prob-LRU, and popularity-based schemes. The simulation results reveal that IVCN-RANO outperforms on performances including hit rate, average weighted hops, and average content redundancy.

INDEX TERMS Fog computing, content network, network function virtualization, service function chain, heterogeneous radio access network.

I. INTRODUCTION

The motivation of the 5G mobile network is to support a much wider range of use case characteristics and corresponding access requirements [1]. Mobile operators are facing tremendous traffic growth due to rich content multimedia, cloud applications and vertical market services. Cloud computing is an effective technology to support varied market services because of its elasticity and scalability. However, the distance between the cloud and the end devices might be an issue for latency-sensitive applications. Service Level Agreements (SLA) may also impose processing at locations where the cloud provider does not have data centers. Fog computing is a novel paradigm to address such issues. It enables provisioning resources and services outside the cloud, at the edge of network, closer to end devices, or eventually, at locations stipulated by SLAs [2].

Content service is one of the important use cases in fog computing [2], [3]. Since the topology of mobile content network is deployed distributed around the world, content service is supported by cloud providers to provide flexible and cost-effective service, which motivates the virtualization of content network based on Network Function Virtualization (NFV). Based on virtualization of content network, contents are usually cached based on content popularity and geographical distribution of users, and they are delivered to end users with Quality of Experience (QoE) guarantees and with low infrastructure cost based on virtualization techniques and geographically dispersed content network nodes in large scale. Network slices are created for tailored virtualization of content network, and it enables

The associate editor coordinating the review of this manuscript and approving it for publication was Nan Zhao.

Mobile Network Operators (MNO) to open their physical network infrastructure platform to the concurrent instantiation of multiple logical self-contained networks [4]–[6]. Furthermore, through virtualized content network, not only physical resources but also contents can be shared [7], [8].

There are some research issues focusing on virtualization of mobile content network, which can be categorized into two classes, one is the issues focusing on the function virtualization of mobile content network and orchestration based on NFV. According to the MANO architecture, the service provision, virtual network function, virtual infrastructure and physical network related resources are managed and orchestrated based on SLA requirements from not only mobile users but also tenants and operators, which forms an ecological system for content delivery stakeholders, including mobile operators, content service providers, tenants and mobile users as well. The other mainly concentrates on resource allocation in mobile content network, which include communication, in-network caching and computing resources in fog-enabled radio access network.

Reference [4], [10]–[20] are research works focusing on function virtualization of content network and related orchestration based on NFV. One of the remarkable issues in virtualization of mobile content network is the virtualization of network function, which can be divided and composed from different point of view, and another is the orchestration and optimization based on the Virtual Network Function (VNF) from different stakeholders. In [4], a Content Delivery Network (CDN) as a Service (CDNaaS) platform is proposed which can create Virtual Machines (VMs) through the network of data centers and provide a customized slice of CDN to users. CDNaaS manages videos by means of caches, transcoders, and streamers hosted in different VMs. In [10], Caching-as-a-Service (CaaS) is proposed, including a caching virtualization framework on Cloud Radio Access Network (C-RAN) and the virtualization of Evolved Packet Core (EPC) based on VM. In [11], the CDN slice consists of four VNF types, namely virtual transcoders, virtual streamers, virtual caches, and a CDN-slice-specific Coordinator. In [12], a power management framework is proposed for NFV-based multimedia content delivery based on Central Processing Unit (CPU) frequency scaling. In [13], an architecture is presented for on-the-fly provisioning of CDN components by using NFV and microservice architecture principles. In [14], on the SONATA Service Platform, a mechanism was designed that allows developers to create and execute Service and Function Specific Managers, which are processes that define service or function specific orchestration behaviors.

Regarding VNF based optimization, collaboration between content delivery stakeholders and optimization based on SLA is investigated in [15]–[18]. In [15], a model for the collaboration between content delivery stakeholders is proposed. A high-level SLA is used for the negotiation of both computing resources and connectivity. In [16], the method of dynamic deployment and optimization of Virtual CDNs (vCDNs) is presented. The CDN collects historical data and uses these data to predict future bandwidth consumption, and generates SLAs with the Internet Service Provider (ISP) using traffic predictions. The ISP prices each SLA and embeds the vCDN as a Service Function Chain (SFC) into its network, and optimizes services with components to reduce costs based on dynamic connectivity demand. In [16], the vCDN solution is composed of two VNFs called the Virtual Streamer (vSTR) and the Virtual Media Gateway (VMG). In [17], a model is proposed for content delivery actors to collaborate over a virtualized infrastructure, which includes Streaming VNF, and Caching and Routing Orchestrator VNF. In [18], the fundamental tradeoff is addressed between deployment cost and service availability for on-demand content delivery service provision over a telecom operator's Network Functions Virtualization Infrastructure.

Regarding the optimization objectives based on VNF orchestration in mobile content network, in [4], the VMs placement problem is formulated aiming at minimizing the cost and maximizing QoE of streaming. In [10], system optimization for CaaS is also addressed to reduce inter-MNO traffic load and intra-MNO traffic load. In [11], the optimal placement of composing VNFs with adequate amount of virtual resources for each VNF is formulated by bargaining game to allocate an appropriate set of VNFs for each CDN slice to meet its performance requirements and minimize the cost in terms of allocated virtual resources. In [15], a linear programming formulation is presented for the VNF embedding to increase problem tractability with a small cost overhead. In [17], a game-theoretic analysis is presented for different ISP-CDN collaboration models and optimality conditions. In [18], a multi-objective optimization problem is formulated to assign computing resources to a set of virtual instances and place these virtual instances in a subset of available physical hosts. In [19], an online VNF Forwarding Graph (VNF-FG) placement in CDNs is addressed considering new VNF instantiations, migration, hosting and routing costs. The objective is to place the VNF-FGs such that total reconfiguration costs are minimized while Quality of Service (QoS) is satisfied. An Integer Linear Programming (ILP) formulation is evaluated in a small-scale scenario. In [20], the VNF components for video service are mixer, transcoder and compressor. To solve the fine-grained services as a chain of VNFs to be placed in CDN due to the specifics of the chains (e.g., one of their end-points is not known prior to the VNF placement), the objective is to find the optimal placement solution of VNFs with minimized cost and satisfied QoS with a large number of servers and end users.

Resource allocation in mobile content network is addressed in [8]–[10] and [21]–[28], and optimization on in-network caching and communication resource plays an important role in fog-enabled radio access network. A virtual caching management system is proposed including the global level managed by the infrastructure provider applying to all tenants or virtualized network operators and the tenant-specific level

managed by specific tenants in [9]. The information-centric wireless network virtualization architecture is addressed in [8], [21], and [22]. In [8], the key components are radio spectrum resource, wireless network infrastructure, virtual resources (including content-level slicing, network-level slicing, and flow level slicing), and information-centric wireless virtualization controller. The resource allocation and in-network caching strategy are formulated as an optimization problem considering virtualization. In [21], content is virtualized and shared by virtual service providers. Virtual resource allocation and caching strategies are formulated as a joint optimization problem considering virtualization and caching. In [22], the virtual resource allocation strategy is modeled as a joint optimization problem in heterogeneous networks including caching and computing. In [23] and [24], the joint resource allocation and content caching problem are formulated to utilize radio and content storage resources in the highly congested backhaul scenario, aiming at minimizing the maximum content request rejection rate of different mobile virtual network operators in different cells. In [25], in cache-enabled hybrid RAN, information-centric content-oriented slicing is modeled which includes slicing on content cache resources and communication resources. The optimization problem is formulated to minimize the average system cost to get the contents required by users. In [26], a framework is proposed which jointly considers networking, caching, and computing to support energy-efficient information retrieval in wireless networks. In [27], the optimization is to maximize gains from managing caches, utilizing network resources of the backhaul and spectrum. In [28], in small-cell networks, a mechanism is designed to deal with the competition for storage among multiple Service Providers (SP) based on multi-object auctions.

From above investigation, it reveals that virtual network function partition is indispensable to the orchestration and optimization of NFV based mobile content network. Composition and placement of VNF components make an impact on the resource allocation and performance of information-centric mobile network, and it affects the QoS based on SLA. Optimal partition of VNF component, function composition and networking orchestration on VNFs bring about not only gains on minimizing delay to get the content, but also gains on reducing the content forwarding bandwidth as well as content caching redundancy. On the basis of research works mentioned above, Liang *et al.* [8], Wang *et al.* [21], Zhou *et al.* [22], Tran and Le [23], [24], Jin *et al.* [25] only consider the resource slicing optimization in RAN without considering NFV and SFC from the respect of service provision in ICN. CDN slices based on SFC are investigated in detail in [4], [11]–[18], and [29], however, it is still a challenge on the research of orchestration and optimization from the respect of information-centric network and its service provision in fog-enabled heterogeneous RAN.

Based on the motivation for virtualization of information-centric mobile content network and its optimal orchestration, an Information-centric Virtual Content Network (IVCN) slicing framework is proposed in this paper. A method on the VNF partition from the point of view of information-centric networking and service provision is put forward. Based on the partition and composition of VNF proposed, not only content network and content service provision but also contents are virtualized and shared based on their copyright constraints in the IVCN slicing framework. IVCNs are generated and mapped to mobile content network slices optimally according to different optimization objects. The main contribution is summarized as follows:

1) An IVCN slicing framework is proposed to virtualize content service, which includes IVCN controllers, virtual Content Agents (vCA), virtual Caches (vCache), contents and virtual Links (vLinks). The key functional modules of Management and Orchestration (MANO) [30] are provided to orchestrate and manage the IVCN slices.
2) Based on the VNF partition from the point of view of information-centric networking and service provision, IVCN slicing is modeled as a use case in the scenario of heterogeneous fog-enabled RAN. A method of performance optimization on IVCN slicing is proposed based on SFC, which aims to optimize the mapping of virtual network functions and virtual content placement.
3) In order to obtain the optimized IVCN slice for content service, vCA mapping on the control plane and vCache mapping on the data plane are formulated to minimize the weighted hops for contents. A heuristic approach called IVCN-RANO (Information-centric Virtual Content Network Optimization in Heterogeneous Radio Access Network) is proposed to solve the NP-hard problem based on ant colony optimization algorithm.
4) The performance of IVCN-RANO is evaluated, and the simulation results indicate that it outperforms in the hit rate, the average weighted hops per user request on the data plane and that on the control plane, as well as in the average content redundancy compared with CEE-LRU, Prob-LRU and Popularity-Based schemes.

The rest of this paper is organized as follows. In section II, the framework and the procedure of the IVCN slicing is proposed. In section III, the system model is introduced. In section IV, the IVCN slicing is formulated as an optimization problem for virtual network function mapping and content placement aiming at minimizing the weighted hops for content in the heterogeneous fog-enabled RAN. In section V, performance evaluation of IVCN-RANO is provided and discussed. Section VI concludes the paper.

II. INFORMATION-CENTRIC VIRTUAL CONTENT NETWORK SLICING

In the information-centric content network, contents are usually cached in the servers on the forwarding path or pre-placed at the edge of the network. Network slicing is applied in order to provide varied content service with high hit rate for requests and low service response time. From the perspective of content providers, network slices are orchestrated by the

requirements of content providers, and content providers manage their own network slices to provide content service to users. However, for those contents with high popularity, different content providers provide duplicated contents, and many cache resources are occupied to cache those duplicated contents, which brings about not only caching redundant contents but also bandwidth and energy consumption, especially in the cases when network slices are independent of each other. Therefore, cooperative caching and content sharing optimization based on the virtualization of content network is of importance.

In this section, IVCN slicing framework based on MANO is proposed. The content service provision procedure based on IVCN slicing framework is presented, and then SFCs in IVCN slicing are described.

A. IVCN SLICING FRAMEWORK BASED ON MANO

The IVCN slicing framework based on MANO is illustrated in Fig.1, which is composed of three main layers including the content service layer, the content slice layer and the content-oriented resource layer. MANO-based entities are provided to orchestrate content service use cases into IVCNs, to orchestrate networking and virtual network function components for IVCNs, to map IVCNs into infrastructure resources, as well as to configure, monitor and optimize IVCN slices during their lifecycles [30], [31].

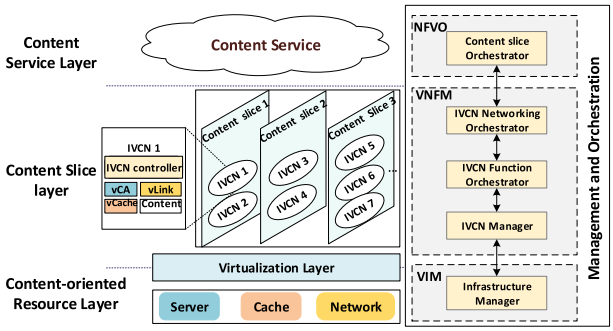

FIGURE 1. Framework of IVCN slice based on MANO.

Based on the concept of NFV, the main functions of an information-centric content network are partitioned as VNF components including VNF1 (including Content Naming, Content Identification and Path Mapping), VNF2 (Processing of Content Request, Content Inquiry, Cooperative Processing of Content Request), VNF3 (Content Cache Management, Content Placement and Path Optimization), as well as VNF4 (Data Forwarding on the Control/Data Plane). Content service is provided via composing and networking various VNF components by the entities of management and orchestration based on different optimization objectives and limited resources in the content network. In order to provide content service flexibly, IVCN is proposed, and the functional modules of an IVCN are defined and composed for content network by the above VNF components.

1) INFORMATION-CENTRIC VIRTUAL CONTENT NETWORK (IVCN)

An IVCN is defined as a virtual information-centric content network that consists of virtualized content-oriented network function modules including virtual Content Agents (vCA), virtual Caches (vCache), virtual Links (vLink), an IVCN controller, and virtual content items. The functional modules of an IVCN are described as follows:

- vCA: It is responsible to process content requests from users, inquiry the content from binded vCaches as well as from other vCAs, and respond to the requests of users. It receives control messages from the IVCN controller, including the optimal path to get the required content, control relationship between vCAs, etc. That is to say, it provides functions such as content identification and path mapping, processing of content request, content inquiry, content request cooperative processing, content cache management, content placement and path optimization, as well as data forwarding on the control plane. Different vCAs have different functions depending on various composition of VNF components of content network. An example of the VNF component composition for vCA is described in Fig.2.

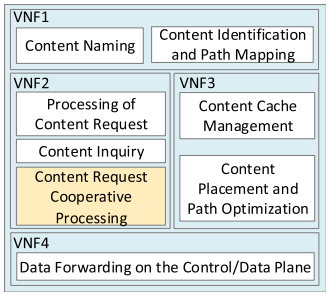

FIGURE 2. An example of VNF composition for a vCA with a selective function of Content Request Cooperative Processing in VNF2.

- vCache: The functions of a vCache include binding with vCAs, caching virtual content items, recording virtual content item copyright information, delivering content, etc. Caching virtual content items is to cache the virtual content item information from content providers, including but not limited to content indexes, content features, and the mapping of virtual content items to content copies in physical content caches. Binding with vCA is to bind the vCache to the specific vCA. Recording virtual content item copyright information is to manage the copyright permission information of specific content to the specific IVCN. Delivering Content is to provide the way to users to get the requested content under the control of an IVCN controller and vCAs.
- vLink: It refers to the virtual link which indicates relationships between two vCAs, or between a vCA and a vCache. The relationships include control, cooperation and relaying. For example, for the relationship of two vCAs, control relationship means one vCA is controlled

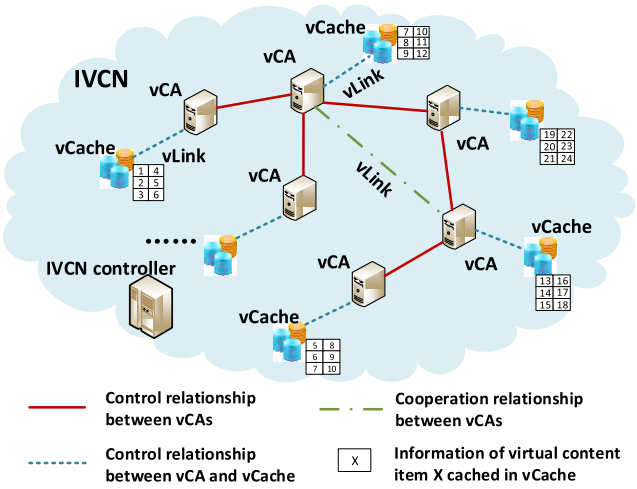

FIGURE 3. Control and cooperation relationship between two vCAs, vCA and vCache in an IVCN.

by the other vCA, and cooperation relationship indicates that the vCA meets some content requests with the help of the other vCA by forwarding the content requests to the vCA cooperating with it. Relaying relationship means that one vCA is used as the relay node to forward content requests to the other vCA on the control plane. In Fig.3, the control and cooperation relationship between two vCAs, vCA and vCache in an IVCN are illustrated as an example.

- IVCN controller: It is responsible to monitor and optimize the IVCN. It also manages the relationships between two vCAs, as well as between a vCA and a vCache.
- Virtual content item: It refers to the content item cached in an IVCN. The virtual content items include those content items owned by content providers and/or operators with copyright limitations, as well as content items uploaded by content publishers with or without copyright limitations. One content item cached in the vCache can be cached in the physical content caches with multiple copies according to caching management schemes.

Content slices can be generated flexibly by using one or several IVCNs according to the requirements of content service providers.

2) THE FUNCTIONS OF EACH LAYER IN THE IVCN SLICING FRAMEWORK

In the content service layer, content service is orchestrated based on various service requirements, and sliced into content slices composed of one IVCN or several IVCNs depending on service requirements and strategies, for example, different virtualization motivations including content categories, content provider preferences, cloud providers as well as geographical locations.

In the content slice layer, the VNF components are composed to generate an IVCN. When users send content requests, content service is provided to users by chaining various VNF components into SFCs both on the control plane and on the data plane.

In the content-oriented resource layer, the physical infrastructure resources are virtualized including common servers, caches and network resources. The functional elements and resources in an IVCN are mapped optimally into the physical infrastructure resources and a physical information-centric content network is obtained.

3) MANAGEMENT AND ORCHESTRATION OF IVCN SLICING FRAMEWORK

Based on the MANO architecture, six modules are presented to manage and orchestrate IVCN-based content slices including Content slice Orchestrator, IVCN Networking Orchestrator, IVCN Function Orchestrator, IVCN manager, IVCN controller and Infrastructure manager, which are explained as follows:

- Content slice Orchestrator: It is a use case of NFVO (NFV Orchestrator), which is used to transform the content service requirements into technical description. It also orchestrates a content slice with a profile with virtual resources for content service based on the performance evaluation results from the content slice layer and the content-oriented resource layer.
- IVCN Networking Orchestrator: It is a use case of VNFM (VNF Manager). It is in charge of orchestrating the networking of IVCNs for each content slice based on the performance evaluation reports from VIM (Virtualized Infrastructure Manager).
- IVCN Function Orchestrator: It is a use case of VNFM. It is designed to orchestrate the service function chains of IVCNs on the control plane and that on the data plane in each content slice based on the performance evaluation reports from VIM.
- IVCN manager: It is a use case of VNFM. It is used to monitor and manage multiple IVCNs, including lifecycle, function and performance based on the performance evaluation reports from VIM.
- IVCN controller: As a use case of VNFM, it is responsible to monitor and manage an IVCN via the performance evaluation reports from VIM.
- Infrastructure manager: As a use case of VIM, it is designed to monitor, evaluate and manage infrastructure resources as well as physical content networks, and map IVCNs into physical infrastructure.

The comparison between conventional network slices for content service and that of IVCN slices is shown in Fig.4. On the one hand, based on NFV, the VNF components can be composed to create an IVCN according to different optimization objectives, which brings advantages from VNF component optimal placement and composition, content request cooperative processing in order to increase hit rate of contents and reduce service response time. On the other hand, from the perspective of contents, barriers between content providers are broken, IVCN slices are orchestrated for all

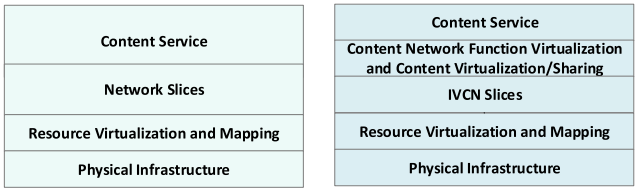

FIGURE 4. Comparison between conventional network slices and IVCN slices for content service provision.

contents, which are virtualized as virtual content items and shared considering copyright information limitation. Content placement in IVCN slices is optimized based on content copyright constraints; therefore, caching resources are used more effectively for caching those non-redundant contents. Optimization of an IVCN slice is also achieved by adjusting the connection relationship between IVCNs, changing the scale or network topology as well as content service type of IVCN.

B. CONTENT SERVICE PROVISION PROCEDURE BASED ON IVCN SLICING

Fig.5 shows an example on the process for content service provision based on IVCN slicing framework, including slice instantiation phase, slice scaling phase and content request phase. In the slice instantiation phase, a content slice is orchestrated and created for providing content service. In the slice scaling phase, MANO monitors and manages all IVCNs of content slices. When the slice owner requests for changes of IVCNs from the content service level, the IVCNs are adjusted. In the content request phase, when a content request is received by a vCA, it is processed and responded.

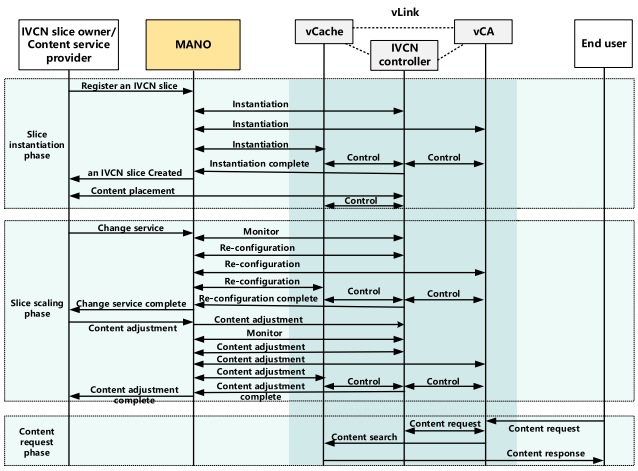

FIGURE 5. An example on the content service provision procedure based on IVCN slicing framework.

C. SERVICE FUNCTION CHAIN IN IVCN SLICING

As defined in [29], service function chaining is a network capability that provides support for application-driven-networking through the ordered interconnection of service functions. In the IVCN slicing framework, content service can be viewed as SFCs embedded in physical infrastructure, which are divided into two types, namely SFC on the control plane and SFC on the data plane. Fig.6 describes the SFC on the control plane and SFC on the data plane, respectively. The VNF component functions in vCAs differentiate depending on VNF component composition of vCAs.

With embedding vCAs and vCaches of an IVCN into the physical infrastructure network, the information-centric content service function chains are also mapped into the physical infrastructure as shown in Fig.6.

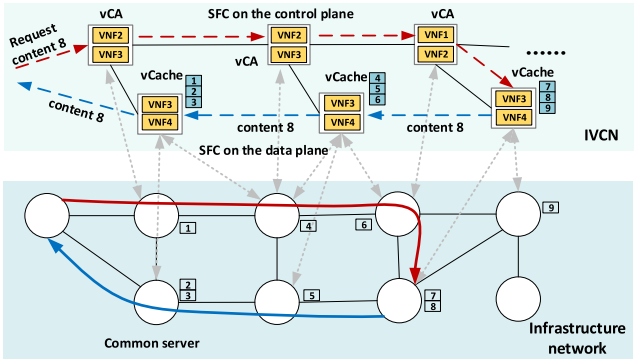

FIGURE 6. Information-centric content SFC mapping into the physical infrastructure.

Both of the SFCs on the control plane and that on the data plane can be optimized. On the control plane, the SFC is optimized according to the placement of VNF component instances in order to improve deployment efficiency and high hit rate, and reduce service response time for users. On the data plane, the SFC is optimized via the content optimal placement to reduce bandwidth cost, caching cost and service response time for the required content, by considering content copyright limitation information and content sharing. Since most of the content users are mobile users, it is a challenge to optimize the SFCs on the control plane and that on the data plane in the scenario of fog-enabled RAN.

III. MODELLING OF IVCN SLICING IN FOG-ENABLED RAN

In this section, the modelling of IVCN slicing is addressed in the scenario of heterogeneous fog-enabled RAN as a typical application case for fog computing. In this use case, the IVCN slice is composed of just one IVCN, we assume that all of the VNF components are available in every vCA. The modelling is introduced in five subsections including scenario, physical network model, content request model, user mobility model and SFC model for IVCN slicing. The key notations are listed in Table.1.

A. SCENARIO

A fog-enabled heterogeneous RAN is shown in Fig.7, which is divided into two levels including Macro Base Station (MBS) level and Small Base Station (SBS) level. In the MBS level, all the MBSs are fog nodes deployed with caches,

TABLE 1. Summary of key notations.

Notation	Description
S	Number of Macro Base Stations (MBSs)
$\mathbf{H}$	Connection matrix between the MBSs
$\mathbf{A}_l$	Set of Small Base Stations (SBSs) covered by MBS l
$\mathbf{B}^l$	Connection matrix between the SBSs covered by MBS l
d	The average connectivity degree of SBSs
M	Number of all the contents provided by all content providers
q_m^l	Probability of the content m in MBS l being requested
P_{ij}	Probability of users moving from MBS i to MBS j
Q_{ij}^l	Probability of users in MBS l moving from SBS i to SBS j
$\mathbf{\Pi}$	User steady-state distribution probability in the MBS level
$\mathbf{\Phi}^l$	user steady-state distribution probability in the SBS level of MBS l
N	Number of all users
MBS_vCA_k	Embedding decision for MBS vCA function in MBS k
$SBS\_vCA_k^l$	Embedding decision for SBS vCA function in SBS k in the coverage of MBS l
v	ID of MBS vCA in the MBS level
u_l	ID of SBS vCA in the coverage of MBS l
$c_{m,k}$	Placement decision for content m in the MBS k
$e_{m,j}^l$	Placement decision for content m in SBS j in the coverage of MBS l
L_{ij}^{MBS}	The shortest path from MBS i to MBS j
$L_{l,ij}^{SBS}$	The shortest path from SBS i to SBS j in the coverage of MBS l
x_k	The total cache space of MBS k
y_j^l	The total cache space of SBS j in the MBS l coverage
R_0	The regularized content size

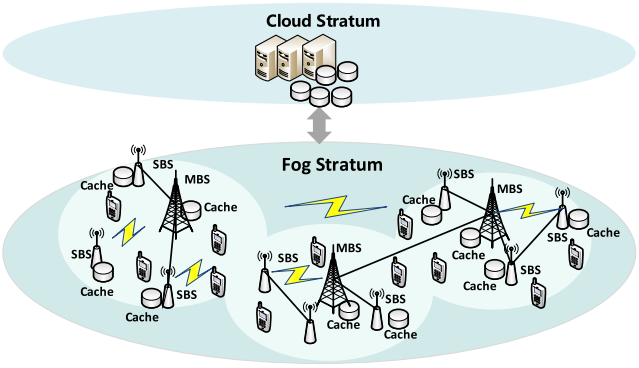

FIGURE 7. A fog-enabled heterogeneous RAN.

and each MBS communicates with other MBSs through wireless or wired links. In the SBS level, all the SBSs are also fog nodes deployed with caches, and each SBS communicates with other SBSs under the same MBS coverage through wireless or wired links. The proposed IVCN slicing framework is deployed to orchestrate and manage IVCN slices in the fog-enabled RAN.

B. PHYSICAL NETWORK MODEL

In order to investigate the impact of network scale to the IVCN-based mobile content network, the physical infrastructure network is divided into two levels. In the MBS level, the deployment of MBSs is shown in Fig.8(a), assuming that

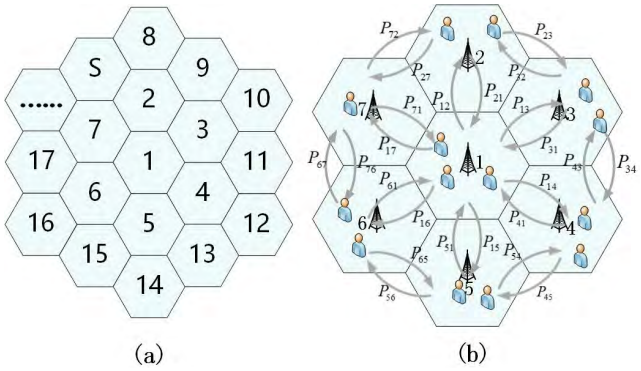

FIGURE 8. (a)Deployment of MBSs,(b)User mobility model in the MBS level.

the total number of MBSs is S. In the SBS level, the SBSs are deployed in the coverage of the MBS with different density and connection. Let the matrix $\mathbf{H}$ indicate the connection matrix between the MBSs, where the element of the matrix $H_{ij} = 1$ indicates connection between MBS i and MBS j, $H_{ij} = 0$ indicates no connection between MBS i and MBS j.

Assuming that all SBSs covered by MBS l can communicate with the MBS l. Let the set $\mathbf{A}_l$ indicate the set of SBSs covered by MBS l and the matrix $\mathbf{B}^l$ indicate the connection matrix between the SBSs covered by MBS l, where the element of the matrix $B_{ij}^l = 1$ indicates connection between SBS i and SBS j in $\mathbf{A}_l$, $B_{ij}^l = 0$ indicates no connection between SBS i and SBS j in $\mathbf{A}_l$.

C. CONTENT REQUEST MODEL

Content modelling is closely related to the content request, caching behavior and asymptotic performance of caching. Items (cacheline addresses for a cache) are ranked according to their popularity (occurrence frequency) and popularity is in a power-law relation with the rank, and it has since long been recognized as the most accurate way to represent sw-cache interactions and more generally computer programs [35]. Based on the investigation from the large scale datasets collected from ISPs, Zipf distribution has been regarded as the accurate model to formulate content distribution [36]. Recently, time series data can be processed based on big data and AI technology. The fluctuations of content popularity which depends on time progress is investigated, and some research issues are concentrating on the popularity prediction based on the analysis of time series data collected from content-oriented network [37]–[40]. According to the investigation of content popularity prediction, content distribution is proved to follow Zipf distribution averaged with time [37], [41], [42]. Based on the above investigation, content popularity is modeled as Zipf distribution in our paper.

In the above fog-enabled heterogeneous RAN, assume that the popularity of the content in different MBS is different, and the number of all the contents provided by the content provider is M. Let $rank_m^l$ indicate the popularity rank of content m and $\alpha_l \geq 0$ indicate the skewness of content

popularity in MBS l. According to Zipf law, the probability of the content m in MBS l requested by a user is given by $q_m^l = \frac{(rank_m^l)^{-\alpha_l}}{\sum_{i=1}^{M}(rank_i^l)^{-\alpha_l}}$. When $\alpha_l = 0$, all the contents in MBS l have the same popularity. The higher α_l is, the more probable the popular contents in MBS l are to be requested.

D. USER MOBILITY MODEL

The mobility of users is modeled by Markov process according to [32]. The mobility of users is divided into two parts, namely mobility in the MBS level and mobility in the SBS level. In the MBS level, a user moves from one MBS to its neighbor MBSs shown in the Fig.8(b). In the SBS level, a user moves among the SBSs deployed in the MBS coverage, that is to say, the user moves from one SBS to its neighbor SBSs in the same MBS coverage.

In general, in the MBS level, let P_{ij} indicate the probability of users moving from MBS i to MBS j and Q_{ij}^l indicate the probability of users moving from SBS i to SBS j in MBS l. P_{ij} is the element of the transition probability matrix $\mathbf{P}$ and Q_{ij}^l is the element of the transition probability matrix $\mathbf{Q}^l$ for MBS l. According to the Markov process, the user steady-state distribution probability matrix in the MBS level and in the SBS level can be indicated as Π and Φ^l, respectively, where

$$\begin{cases} \mathbf{P}^T \Pi = \Pi \\ \sum_{i=1}^{S} \Pi_i = 1 \end{cases} \tag{1}$$

$$\begin{cases} (\mathbf{Q}^l)^T \Phi^l = \Phi^l \\ \sum_{i \in \mathbf{A}_l} \Phi_i^l = 1 \end{cases} \tag{2}$$

Based on the above mobility model, the steady-state probability of user attaching to the SBS i deployed in the coverage of MBS l is $\Pi_l \Phi_i^l$. Assuming the number of the users is N and users are uniformly distributed in each SBS, then the average number of users attaching to the SBS i deployed in the coverage of MBS l is $N \Pi_l \Phi_i^l$.

E. SFC MODEL FOR IVCN SLICING

The IVCN slicing in the fog-enabled heterogeneous RAN is shown in Fig.9. According to Fig.9, the procedure of service function chaining on the control plane and that on the data plane are given in section III-E.1 and III-E.2, respectively.

1) SFC MODEL ON THE CONTROL PLANE FOR IVCN SLICING

The provision and optimization of SFC can be formulated as an ILP problem by graph embedding with the given set of nodes and links. Where to place instances of VNFs on servers in a NFV-based infrastructure to accommodate the traffic for a given set of SFC requests depends on the operators/providers competing goals [29]. However, concerning the service function chaining of the IVCN slice in the framework, the SFC is not a graph with a given set of fog nodes and links, since the

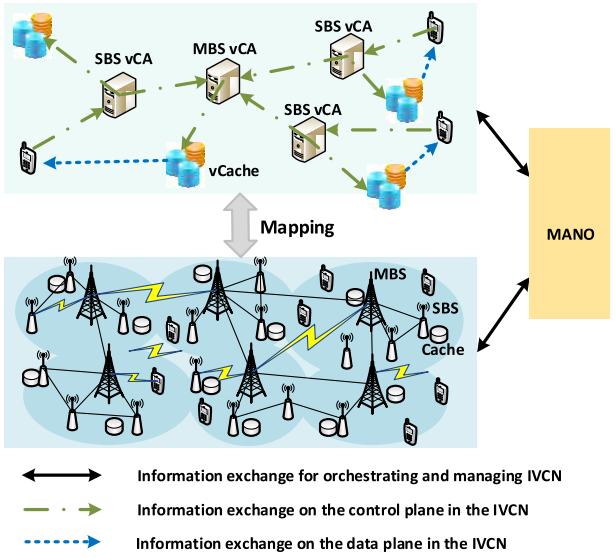

FIGURE 9. The SFCs of an IVCN slice in the heterogeneous RAN.

SFCs need to be optimized depending on the content inquiry schemes and content placement.

Assuming that in the fog-enabled heterogeneous RAN based on IVCN slicing, the MANO orchestrator, a vCA in the MBS level (called MBS vCA) and several vCAs in the SBS level (called SBS vCA) are included on the control plane. The MANO orchestrator exchanges control messages with the MBS vCA and SBS vCAs to orchestrate and manage the IVCN slice.

In the physical network, in the MBS level, the MBS vCA with vCache is deployed in a MBS. In the SBS level, some SBS vCAs with vCaches are deployed in the SBSs in the MBS coverage. The MBS vCA controls the SBS vCAs in the coverage of the MBS.

When a content request is received by a vCA, the content request is processed and a SFC is formed on the control plane, which is a sequential information-centric content network functional chain of vCAs including VNF components mentioned in section II. A simple example to process the content request on the control plane in the IVCN slice is described as follows:

Step 1: The user sends a content request to the attached SBS vCA. If the content request is met, the SBS vCA obtains the shortest data forwarding path from the vCache to the user requesting for the content in the SBS level in the IVCN slice, and responds to the user with the content information including the shortest data forwarding path; Else, go to **step 2**.

Step 2: If the content request cooperative processing on the SBS vCA is supported and embedded, go to **step 3**; Else, go to **step 6**.

Step 3: Based on the scheme of content request cooperative processing, the SBS vCA forwards the content request to the MBS vCA. If the content request can be met, the MBS vCA obtains the shortest forwarding path from the vCache to the user requesting for the content and responds to the SBS

vCA with the content information including the shortest data forwarding path; Else, go to **step 4**.

Step 4: If the content request cooperative processing on the MBS vCA is supported and embedded, the MBS vCA calls the virtual function named content request cooperative processing, and go to **step 5**; Else, the MBS vCA returns the inquiry failure message to the SBS vCA, and go to **step 6**.

Step 5: The MBS vCA forwards the content request to the data center in the core network. The content request of the user is met by the data center, and the shortest data forwarding path to get the content is returned to the user by the SBS vCA via the MBS vCA.

Step 6: The SBS vCA respond to the user with the inquiry failure message for the content request.

From the procedure illustrated above, the formed SFC is an optimal path to obtain the content, which is related to the cooperative processing of content request, the placement of content required by the user as well as the networking topology of the fog-enabled heterogeneous RAN.

2) SFC MODEL ON THE DATA PLANE FOR IVCN SLICING

On the data plane of IVCN slicing, vCaches allocated to each vCA (including the MBS vCA and the SBS vCAs) are deployed in the IVCN slice. The content items are cached in the vCaches. After the procedure of content inquiry on the control plane, the shortest data forwarding path to form the optimal SFC on the control plane is obtained either by a SBS vCA or the MBS vCA, and the content forwarding path from the vCache to the user is also got. That is to say, the SFC on the data plane is formed based on the shortest data forwarding path for the content.

As soon as the user requesting for the content receives the information including the shortest content forwarding path, the user gets the content with the shortest content forwarding path in the physical content network.

From the above content inquiry procedure, it is necessary to optimize the SFC of IVCN slicing on the control plane and data plane, which depends on the optimal placement and composition of VNF components, content placement and networking mapped to the physical content network. In section IV, optimization for SFC in the IVCN slice based on IVCN networking orchestration is formulated.

IV. PROBLEM FORMULATION

In order to formulate the SFC optimization for IVCN mapping to the physical network, some issues are needed to be considered: (1) vCA mapping, namely embedding MBS vCA and SBS vCAs into the physical network; (2) optimal virtual content item placement in the vCaches binded to the vCAs and mapping into the physical network. The vCA mapping problem is related to the optimization of SFC on the control plane, while the content placement in the vCaches is related to the optimization on both control plane and data plane.

In this section, the SFC optimization problem called IVCN-RANO in the IVCN slicing is formulated to optimize the vCA mapping and content placement on the control plane and data plane, which are resided in the IVCN Networking Orchestrator. The VNF component placement and related optimization resided in the IVCN Function Orchestrator are left for research in the future.

The IVCN slice is composed of three main elements including virtual content agent (denoted as **vCA**), virtual cache (denoted as **vCache**) and virtual link (denoted as **vL**). Let the set $\mathbf{MBS\_vCA} = \{MBS\_vCA_1, MBS\_vCA_2, ..., MBS\_vCA_S\}$ indicate the MBS vCA function embedding binary decision set in MBS level, where $MBS\_vCA_k \in \{0, 1\}$ indicates whether the MBS vCA function is embedded in MBS k in the MBS level. Let the set $\mathbf{SBS\_vCA}^l = \{SBS\_vCA^l_1, SBS\_vCA^l_2, ..., SBS\_vCA^l_{|\mathbf{A}_l|}\}$ indicate the SBS vCA function embedding binary decision set in SBS level in the MBS l coverage, where $SBS\_vCA^l_k \in \{0, 1\}$ indicates whether the SBS vCA function is embedded in SBS k in the coverage of MBS l. So $\mathbf{vCA} = \{\mathbf{MBS\_vCA}, \mathbf{SBS\_vCA}^1, \mathbf{SBS\_vCA}^2, ..., \mathbf{SBS\_vCA}^S\}$.$\{v = k|MBS\_vCA_k = 1\}$ indicates the ID of MBS vCA in the MBS level while $\{u_l = k|SBS\_vCA^l_k = 1\}$ indicates the ID of SBS vCA in the coverage of MBS l.

Let $c_{m,k}$ indicate the binary decision variable for content m placement in the MBS k, as the element of placement decision matrix $\mathbf{C}$ in MBS level. $e^l_{m,j}$ indicates the binary decision variable for content m placement in SBS j in the coverage of MBS l, as the element of placement decision matrix $\mathbf{E}^l$ in SBS level of MBS l. So the set $\mathbf{vCache} = \{\mathbf{C}, \mathbf{E}^1, \mathbf{E}^2, ..., \mathbf{E}^S\}$ indicates the content placement in vCache for the MBS vCA and SBS vCAs.

In addition, let L^{MBS}_{ij} indicate the shortest path from MBS i to MBS j as the element of available physical link resource set $\mathbf{MBS\_L}$ in MBS level. $|L^{MBS}_{ij}|$ indicates the shortest hops from MBS i to MBS j. $L^{SBS}_{l,ij}$ denotes the shortest path from SBS i to SBS j in MBS l coverage as the element of available physical link resource set $\mathbf{SBS\_L}^l$ in SBS level in MBS l coverage. So $\mathbf{vL} = \{\mathbf{MBS\_L}, \mathbf{SBS\_L}^1, \mathbf{SBS\_L}^2, ..., \mathbf{SBS\_L}^S\}$ indicates the available physical link resource set for mapping the virtual links of IVCN.

Assuming that the optimization algorithms are executed every time interval T_{op}, three algorithms are designed including IVCN-RANO-J-C, IVCN-RANO-J-DC and IVCN-RANO-S-D.

A. IVCN-RANO-J-C

In IVCN-RANO-J-C, vCA and vCache mapping with content placement are jointly optimized, which means it aims to minimize the content request forwarding weighted hops in the IVCN slice with the cache space constraints only on the control plane.

When all the users in the coverage of MBS l request content m which is met in the SBS vCA in the coverage of MBS l,

the weighted hops on the control plane in the IVCN slice mapping into the physical network is as (3):

$$SFC\_Hop^{C}_{SBS}(m,l) = \sum_{i\in\mathbf{A}_l} w_1 N \Pi_l \Phi^l_i q^l_m (|L^{SBS}_{l,iu_l}| + \sum_{j\in\mathbf{A}_l} e^l_{m,j}|L^{SBS}_{l,ju_l}|) \quad (3)$$

where $|L^{SBS}_{l,iu_l}| = \sum_{k\in\mathbf{A}_l} SBS\_vCA^l_k |L^{SBS}_{l,ik}|$, w_1 is the weight of the physical link between fog nodes in the SBS level.

When all the users in the coverage of MBS l request content m which is met in the MBS vCA, the weighted hops on the control plane mapping into the physical network can be obtained as (4):

$$SFC\_Hop^{C}_{MBS}(m,l) = \sum_{i\in\mathbf{A}_l} N \Pi_l \Phi^l_i q^l_m (w_3 |L^{MBS}_{lv}| + w_2 + w_3 \sum_{k=1}^{S} c_{m,k}|L^{MBS}_{kv}|)(\prod_{j\in\mathbf{A}_l}(1-e^l_{m,j})) \quad (4)$$

where $|L^{MBS}_{lv}| = \sum_{k=1}^{S} MBS\_vCA_k |L^{MBS}_{lk}|$, w_2 is the weight of the physical link between the fog node in the SBS level and the fog node in the MBS level, w_3 is the weight of the physical link between fog nodes in the MBS level.

When all the users in the coverage of MBS l request content m which is met in the core network, the weighted hops on the control plane mapping into the physical network is (5):

$$SFC\_Hop^{C}_{core}(m,l) = \sum_{i\in\mathbf{A}_l} w_4 N \Pi_l \Phi^l_i q^l_m \cdot (\prod_{j\in\mathbf{A}_l}(1-e^l_{m,j}))(\prod_{k=1}^{S}(1-c_{m,k})) \quad (5)$$

where w_4 is the weight of the physical link between the fog node in the MBS level and the core network.

Define $NC^{SBS-SBS}_{l,ij}$ as the number of content transmission from SBS j to SBS i in SBS level of MBS l coverage in steady state, $NC^{MBS-SBS}_{l,i}$ as the number of content transmission from MBS l to SBS i in MBS l coverage in steady state, $NC^{MBS-MBS}_{lk}$ as the number of content transmission from MBS k to MBS l in steady state, $NC^{core-MBS}_{l}$ as the number of content transmission from the core network to MBS l in steady state. Then, (6), (7), (8) and (9) are obtained as,

$$NC^{SBS-SBS}_{l,ij} = \sum_{m=1}^{M} N \Pi_l \Phi^l_i q^l_m e^l_{m,j} B^l_{ij} \quad (6)$$

$$NC^{MBS-SBS}_{l,i} = \sum_{m=1}^{M} N \Pi_l \Phi^l_i q^l_m (\prod_{j\in\mathbf{A}_l}(1-e^l_{m,j})) \quad (7)$$

$$NC^{MBS-MBS}_{lk} = \sum_{m=1}^{M} N \Pi_l q^l_m (\prod_{j\in\mathbf{A}_l}(1-e^l_{m,j}))c_{m,k}H_{lk} \quad (8)$$

$$NC^{core-MBS}_{l} = \sum_{m=1}^{M} N \Pi_l q^l_m (\prod_{j\in\mathbf{A}_l}(1-e^l_{m,j})) \cdot (\prod_{k=1}^{S}(1-c_{m,k})) \quad (9)$$

To describe communication link resources among SBSs, MBSs and the core network, let $V^{SBS-SBS}_{l,ij}$ denote the total traffic volume of regularized data allocated to the content transmission between SBS i and SBS j in MBS l coverage during T_{op}. $V^{MBS-SBS}_{l,i}$ denotes the total traffic volume of regularized data allocated to the content transmission between MBS l and SBS i in MBS l coverage during T_{op}. $V^{MBS-MBS}_{lk}$ denotes the total traffic volume of regularized data allocated to the content transmission between MBS l and MBS k during T_{op}. $V^{core-MBS}_{l}$ denotes the total traffic volume of regularized data allocated to the content transmission between MBS l and the core network during T_{op}.

The vCA and vCache mapping with content placement problem for IVCN-RANO-J-C can be formulated as follows:

$$min \sum_{m=1}^{M}\sum_{l=1}^{S}(SFC\_Hop^{C}_{SBS}(m,l) + SFC\_Hop^{C}_{MBS}(m,l) + SFC\_Hop^{C}_{core}(m,l)) \quad (10)$$

$$s.t. \sum_{k\in\mathbf{A}_l} SBS\_vCA^l_k = 1, SBS\_vCA^l_k \in \{0,1\}, \forall l \quad (C1)$$

$$\sum_{k=1}^{S} MBS\_vCA_k = 1, MBS\_vCA_k \in \{0,1\} \quad (C2)$$

$$\sum_{m=1}^{M} c_{m,k}R_0 \le x_k, c_{m,k} \in \{0,1\}, \forall k \quad (C3)$$

$$\sum_{m=1}^{M} e^l_{m,j}R_0 \le y^l_j, e^l_{m,j} \in \{0,1\}, \forall l, \forall j \in \mathbf{A}_l \quad (C4)$$

$$R_0 NC^{SBS-SBS}_{l,ij} \le V^{SBS-SBS}_{l,ij}, \forall i,j \in \mathbf{A}_l, \forall l \quad (C5)$$

$$R_0 NC^{MBS-SBS}_{l,i} \le V^{MBS-SBS}_{l,i}, \forall i \in \mathbf{A}_l, \forall l \quad (C6)$$

$$R_0 NC^{MBS-MBS}_{lk} \le V^{MBS-MBS}_{lk}, \forall l,k \quad (C7)$$

$$R_0 NC^{core-MBS}_{l} \le V^{core-MBS}_{l}, \forall l \quad (C8)$$

Constraints (C1) means that there is only one SBS vCA in each MBS coverage and constraint (C2) means that there is only one MBS vCA in the MBS level. Constraints (C3) is to limit the cache space in the MBS level for the IVCN slice and constraints (C4) is the limit of cache space in the SBS level for the IVCN slice. Constraint (C5) means that the content transmission traffic from SBS j to SBS i in MBS l coverage cannot beyond the limitation $V^{SBS-SBS}_{l,ij}$. Constraint (C6) means that the content transmission traffic from MBS l to SBS i in MBS l coverage cannot beyond the limitation $V^{MBS-SBS}_{l,i}$. Constraint (C7) means that the content transmission traffic from MBS k to MBS l cannot beyond the limitation $V^{MBS-MBS}_{lk}$. Constraint (C8) means that the content

transmission traffic from the core network to MBS l cannot beyond the limitation $V_l^{core-MBS}$.

Considering the copyright permission of contents, for those contents which are not permitted to share among caches, they can also be modeled and added as the content placement constraints in the IVCN-RANO algorithms.

B. IVCN-RANO-J-DC

In IVCN-RANO-J-DC, the vCA and vCache mapping with content placement are jointly optimized. It aims to minimize the content request forwarding weighted hops with the cache space constraints both on the data plane and on the control plane.

When all the users in the coverage of MBS l request content m which is met in the SBS vCA in the coverage of MBS l, the weighted hops on the data plane mapping into the physical network is (11), and the weighted hops both on the data plane and control plane which are mapped into the physical network is (12):

$$SFC\_Hop^{D}_{SBS}(m,l) = \sum_{i\in\mathbf{A}_l} w_1 N \Pi_l \Phi_i^l q_m^l \Big(\sum_{j\in\mathbf{A}_l} e_{m,j}^l |L_{l,ij}^{SBS}|\Big) \tag{11}$$

$$SFC\_Hop^{DC}_{SBS}(m,l) = SFC\_Hop^{D}_{SBS}(m,l) + SFC\_Hop^{C}_{SBS}(m,l) \tag{12}$$

When all the users in the coverage of MBS l request content m which is met in the MBS vCA, the weighted hops on the data plane mapping into the physical network is (13), and the weighted hops both on the data plane and control plane which are mapped into the physical network can be got in (14):

$$SFC\_Hop^{D}_{MBS}(m,l) = \sum_{i\in\mathbf{A}_l} N \Pi_l \Phi_i^l q_m^l \prod_{j\in\mathbf{A}_l}(1-e_{m,j}^l) \cdot \Big(\sum_{k=1}^{S} c_{m,k}(w_3|L_{lk}^{MBS}| + w_2)\Big) \tag{13}$$

$$SFC\_Hop^{DC}_{MBS}(m,l) = SFC\_Hop^{D}_{MBS}(m,l) + SFC\_Hop^{C}_{MBS}(m,l) \tag{14}$$

When all the users in the coverage of MBS l request content m which is met in the core network, the weighted hops on the data plane mapping into the physical network can be obtained as (15), and the weighted hops both on the data plane and control plane mapping into the physical network is (16):

$$SFC\_Hop^{D}_{core}(m,l) = \sum_{i\in\mathbf{A}_l} (w_4 + w_2) N \Pi_l \Phi_i^l q_m^l \cdot \Big(\prod_{j\in\mathbf{A}_l}(1-e_{m,j}^l)\Big)\Big(\prod_{k=1}^{S}(1-c_{m,k})\Big) \tag{15}$$

$$SFC\_Hop^{DC}_{core}(m,l) = SFC\_Hop^{D}_{core}(m,l) + SFC\_Hop^{C}_{core}(m,l) \tag{16}$$

The vCA and vCache mapping with content placement problem for IVCN-RANO-J-DC can be formulated as follows:

$$\begin{aligned} min\ & \sum_{m=1}^{M}\sum_{l=1}^{S}(SFC\_Hop^{DC}_{SBS}(m,l) \\ & + SFC\_Hop^{DC}_{MBS}(m,l) + SFC\_Hop^{DC}_{core}(m,l)) \\ s.t.\ & (C1)(C2)(C3)(C4)(C5)(C6)(C7)(C8) \end{aligned} \tag{17}$$

C. IVCN-RANO-S-D

The vCA and vCache mapping with content placement can also be optimized separately. In IVCN-RANO-S-D, firstly, the vCA mapping is optimized including the MBS vCA and SBS vCAs, then the vCache mapping with content placement is optimized.

1) VIRTUAL CONTENT AGENT MAPPING

In IVCN-RANO-S-D, the request forwarding path from all users to SBS vCAs and the MBS vCA in the IVCN slice should be the shortest. In order to optimize the content forwarding hops on the control plane, vCAs are selected and embedded in the MBS level and SBS level first, and then vCaches with content placement are optimally placed in the SBS and MBS level.

The request forwarding hops of SFC on the control plane from user to the SBS vCA and MBS vCA in the IVCN slice mapping into the physical network is written as (18)

$$SFC\_CP\_Hops = \sum_{l=1}^{S}\sum_{i\in\mathbf{A}_l}\Big(\sum_{k\in\mathbf{A}_l} SBS\_vCA_k^l |L_{l,ik}^{SBS}| + \sum_{k=1}^{S} MBS\_vCA_k |L_{lk}^{MBS}|\Big)\cdot N\Pi_l\Phi_i^l \tag{18}$$

The vCA mapping problem for IVCN-RANO-S-D can be formulated as follows:

$$\begin{aligned} & min\ SFC\_CP\_Hops \\ & s.t.\ (C1)(C2) \end{aligned} \tag{19}$$

2) VIRTUAL CACHE MAPPING WITH CONTENT PLACEMENT

Based on the vCA mapping result in section IV-C.1, the vCache mapping with content placement is formulated only on the data plane to minimize the content forwarding weighted hops in the IVCN slice with the cache space constraint. The weighted hops on the data plane in the IVCN slice mapping into the physical network is already modeled in section IV-B. So the vCache mapping with content placement problem for IVCN-RANO-S-D is obtained as follows,

$$\begin{aligned} min\ & \sum_{m=1}^{M}\sum_{l=1}^{S}(SFC\_Hop^{D}_{SBS}(m,l) \\ & + SFC\_Hop^{D}_{MBS}(m,l) + SFC\_Hop^{D}_{core}(m,l)) \\ s.t.\ & (C3)(C4)(C5)(C6)(C7)(C8) \end{aligned} \tag{20}$$

D. THE SOLUTION OF IVCN-RANO

The optimization problem of the IVCN-RANO algorithms in (10), (17) and (20) are general integer linear programming problems, which is a NP hard problem [43]. The solution space of IVCN-RANO-J-C and IVCN-RANO-J-DC is $2^{(M+1)*(S+\sum_{l=1}^{S}|\mathbf{A}_l|)}$, while the solution space of IVCN-RANO-S-D is $2^{M*(S+\sum_{l=1}^{S}|\mathbf{A}_l|)}$. It is difficult to find an optimized solution in an efficient way, especially when S, M, and $|\mathbf{A}_l|$ are large. In order to solve the NP hard problems in IVCN-RANO, the distributed heuristic solutions of IVCN-RANO based on ACO (Ant Colony Optimization algorithm) are provided. ACO algorithm is used to optimize vCA mapping and content placement in vCache both in the SBS level and in the MBS level, since it is widely applied in solving NP-hard problem as a traditional heuristic algorithm [44].

1) IVCN-RANO-J

The IVCN-RANO-J algorithm is proposed to solve the optimization problem of IVCN-RANO-J-C and IVCN-RANO-J-DC. In IVCN-RANO-J, vCA and vCache mapping with content placement are jointly optimized. The process of IVCN-RANO-J is described in Algorithm 1.

Algorithm 1 IVCN-RANO-J

Description: This algorithm is to solve the optimization problem IVCN-RANO-J-C and IVCN-RANO-J-DC.
Input:S,$\mathbf{H}$,$\mathbf{B}^l$,q_m^l,Π,Φ^l,$\mathbf{A}_l$,$l \in [1, S]$,$m \in [1, M]$, $opt \in \{C, DC\}$
Initialize the set **vCA**, **vCache** and **vL**;
Set the optimization value $target = +\infty$;
DO
  If $opt == C$ //choose IVCN-RANO-J-C
    For $l \leftarrow 1 : S$ //traverse all S MBSs
      Using ACO to solve (10) to obtain the matrix $\mathbf{E}^l$ and the set $\mathbf{SBS\_vCA}^l$ of SBS level in MBS l;
    End For;
    Using ACO to solve (10) to obtain the matrix **C** and the set **MBS_vCA** of MBS level;
    Compute the new optimization (10) value *target*;
  Else $opt == DC$ //choose IVCN-RANO-J-DC
    For $l \leftarrow 1 : S$ //traverse all S MBSs
      Using ACO to solve (17) to obtain the matrix $\mathbf{E}^l$ and the set $\mathbf{SBS\_vCA}^l$ of SBS level in MBS l;
    End For;
    Using ACO to solve (17) to obtain the matrix **C** and the set **MBS_vCA** of MBS level;
    Compute the new optimization (17) value *target*;
  End If;
Until *target* converges;
Output:the set **vCA**, **vCache**

2) IVCN-RANO-S

The IVCN-RANO-S algorithm is presented to solve the optimization problem of IVCN-RANO-S-D. In IVCN-RANO-S, the vCA and vCache mapping with content placement are optimized separately. The process of IVCN-RANO-S is given in Algorithm 2.

Algorithm 2 IVCN-RANO-S

Description: This algorithm is to solve the optimization problem IVCN-RANO-S-D.
Input:S,$\mathbf{H}$,$\mathbf{B}^l$,q_m^l,Π,Φ^l,$\mathbf{A}_l$,$l \in [1, S]$,$m \in [1, M]$
Initialize the set **vCA**, **vCache** and **vL**;
For $l \leftarrow 1 : S$ //traverse all S MBSs
  Using Dijkstra to solve (19) to obtain the set $\mathbf{SBS\_vCA}^l$ of SBS level in MBS l;
End For;
Using Dijkstra to solve (19) to obtain the set **MBS_vCA** of MBS level;
Set the optimization value $target = +\infty$;
DO
  For $l \leftarrow 1 : S$ //traverse all S MBSs
    Using ACO to solve (20) to obtain the matrix $\mathbf{E}^l$ of SBS level in MBS l;
  End For;
  Using ACO to solve (20) to obtain the matrix **C** of MBS level;
  Compute the new optimization (20) value *target*;
Until *target* converges;
Output:the set **vCA**, **vCache**

V. PERFORMANCE EVALUATION

In information-centric network, the QoS and SLA requirements can be evaluated by two categories, one is user-centric QoS requirements, which can be evaluated by the average hit rate and average response time for requested contents to ensure QoS of content users, the other is network-centric metrics, which includes the cost to provide the content service, and is often related to the caching resource and communication resource for Opex besides Capex. Therefore, average hit rate, average hops to get the contents and caching redundancy are the metrics which are often used for performance evaluation in information-centric network [34].

In this section, the performance of IVCN-RANO optimization algorithm is evaluated by MATLAB with Monte Carlo method. The performance metrics include hit rate, average weighted hops per request on the data plane and control plane as well as average content redundancy, which are described in section V-A.

The parameters are selected including the number of the MBS S, the Zipf parameter α, and the number M of the served contents. The physical network in the SBS level is generated by ER model [33]. The parameter in ER model is the average network connectivity degree which is labelled as d. The detail parameter setting is listed in section V-B.

Regarding the selection of baseline schemes, to the best of our knowledge, there is no paper concentrating on the information-centric resource allocation from the point of

view of information-centric networking and service provision based on NFV yet. According to IVCN slicing optimized by IVCN-RANO, embedding of an IVCN includes optimal vCA mapping, vCache mapping and placement of virtual content items into physical caches. The vCA and its binded vCache are placed at the same physical node to save the communication bandwidth cost between the vCA and the binded vCache, unless additional configuration requirement for vCA and vCache is needed. That is to say, optimal embedding of an IVCN is the optimal embedding of vCA, vCache and optimal content placement depending on various optimization objectives. However, content placement is only the content insertion, content eviction is performed in ICN due to limited caching capacity. Therefore, mapping of vCA and vCache, content management including content placement and content eviction should be considered jointly. In the ICN without IVCN mapping, every node is responsible for request inquiry and content management, and the request inquiry procedure is flat compared to that in IVCN. The typical content management schemes in ICN mainly include CEE, Prob and Popularity-Based schemes, and content eviction schemes include LRU, LFU, etc [34]. Based on the above comparison, in order to evaluate content request cooperative processing and optimal caching performance of IVCN, the baseline algorithms are selected according to the following two important factors: (1) For node function mapping, every IVCN components are mapped into every node except the VNF component of content request cooperative processing; (2) For content management schemes, some typical caching management schemes in ICN are selected including CEE-LRU, Prob-LRU and Popularity-Based schemes, then, the baselines are composed with node mapping and content management schemes, which are named and described as follows,

(a) CEE-LRU: every IVCN components are mapped into every node except the content request cooperative processing; the content insertion policy is CEE (Cache Everything Everywhere) and the content eviction policy is LRU (Least Recently Used).

(b) Prob-LRU: every IVCN components are mapped into every node except the content request cooperative processing; the content is cached in the base stations on the content forwarding path with the probability 0.3.

(c) Popularity-Based scheme: every IVCN components are mapped into every node except the content request cooperative processing; the most popularity contents ranked from 1 to y_j^l are cached in the SBS j in the coverage of the MBS l, while the contents ranked after y_j^l are cached in the MBS l.

A. PERFORMANCE METRICS USED IN THE PERFORMANCE EVALUATION

Since the IVCN slice is embedded in the physical infrastructure, the method to evaluate the performance of the IVCN slice can be considered from two aspects, namely from the aspect of the performance of physical content network and from the viewpoint of virtual content network orchestration. Thus, the performance metrics are divided into two parts, the first part are the performance metrics for the physical content network mapped by the IVCN slice, which include hit rate, service response time, and system cost to meet one content request based on the cache and bandwidth resources (for example, the optimal hops to get the content); while the second part are the performance metrics used to evaluate gains from VNF component composition and networking orchestration as well as content sharing, which are reflected in the performance metrics on the cost to get the content after orchestration, as well as reduction of average content redundancy.

In order to evaluate the performances of three algorithms proposed in section IV, three performance metrics are provided including the hit rate of user requests, the average weighted hops per request on the data plane and that on the control plane, and the average content redundancy. The hit rate of user requests is used to evaluate the performance of the embedded IVCN slice in the physical network. The average weighted hops per request on the data plane and that on the control plane aim to assess the communication resources allocated to the contents and the service response time for user requests of the embedded IVCN slice in the physical network. The average content redundancy aims to measure the average content redundancy of the embedded IVCN slice in the physical network benefited from content networking orchestration and content sharing. The definition of the three metrics are illustrated as follows.

(1) The hit rate of user requests

The hit rate usually refers to the ratio of the content requests met by the system among total content requests required by users in the system. In the IVCN-based heterogeneous fog-enabled RAN, it includes the hit rate in the SBS level, the hit rate in the MBS level and the total hit rate of the embedded IVCN slice in the physical network, which are defined in (21), (22) and (23) as,

$$HT_{SBS} = \frac{hit\_num\_SBS}{total\_req\_num} \tag{21}$$

$$HT_{MBS} = \frac{hit\_num\_MBS}{req\_num\_to\_MBS} \tag{22}$$

$$HT_{slice} = \frac{hit\_num\_SBS + hit\_num\_MBS}{total\_req\_num} \tag{23}$$

where $total\_req\_num$ is denoted as the total number of the user requests sent to vCAs in the IVCN slice embedded in the physical network, hit_num_SBS indicates the number of the requests which are hit in the SBS vCAs, hit_num_MBS indicates the number of the requests which are hit in the MBS vCA, and $req\_num\_to\_MBS$ is the number of requests which are sent to MBS vCAs.

(2) The average weighted hops per request on the data plane and that on the control plane

The average weighted hops is defined to measure the average weighted hops to get contents in physical content network on the data plane and that on the control plane, namely in (24)

and (25), respectively:

$$AWH_{CP} = \frac{total\_hops\_CP}{total\_req\_num} \tag{24}$$

$$AWH_{DP} = \frac{total\_hops\_DP}{total\_req\_num} \tag{25}$$

where *total_hops_CP* indicates the total weighted hops to meet all the user requests on the control plane, and *total_hops_DP* is the total weighted hops to forward the content in response to all the user requests on the data plane. Obviously, the average weighted hops can be regarded as the communication resources allocated for transmitting one content for every content request.

(3) The average content redundancy

As we know, in content network, caching is ubiquitous. For example, in Information-Centric Network (ICN), on-path and off-path caching are supported by name resolution request process. In on-path caching, the network places information cached along the path taken by a name resolution request, while in off-path caching, the network exploits information cached outside that path [34]. High hit rate is obtained at the expense of caching redundant content, especially when the Zipf parameter of the content is high, which leads to redundancy of the content, energy consumption for caching those duplicated copies of contents, and bandwidth to transmit the contents as well. In order to measure the redundancy of the content, we propose a metric called the average content redundancy. It is defined as the ratio of copies cached for one content in the cache space on the nodes within certain hops in content-oriented network, so it is a performance metric considering the hops between two duplication of the specific content in the content network.

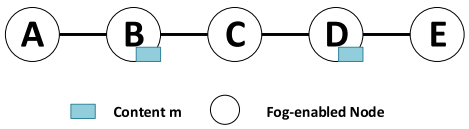

FIGURE 10. An example of defining average content redundancy.

Fig.10 reveals an example for computing average content redundancy of content m with a string topology, in which node B and node D have cached the content m, while other nodes have not. The average content redundancy of the content m within two hops is computed as follows. Assuming that the size of all contents are the same, in the worst case, when content m is cached in all of the five nodes, the total cache units for each node within two hops without considering other content constraints is computed as 3 units to cache content m (namely 0-3 copies are for content m) for node A, 4 units to cache content m (namely 0-4 copies are for content m) for node B, 5 units to cache content m (namely 0-5 copies are for content m) for node C, 4 units to cache content m (namely 0-4 copies are for content m) for node D and 3 units to cache content m (namely 0-3 copies are for content m) for node E. So the accumulative cache space for all the nodes caching content m within the range of two hops in the network is $3+4+5+4+3 = 19$, which is the denominator of the average content redundancy. According to Fig.10, the copy number of the content m for each node within two hops is computed as 1 copy for content m for node A, 2 copies for content m for node B, 2 copies for content m for node C, 2 copies for content m for node D, and 1 copies for content m for node E. So the accumulative number of copies of the content m for all the nodes within the range of two hops in the network is $1+2+2+2+1 = 8$, which is the molecular of the average content redundancy. Finally, the average content redundancy of the content m within two hops is $8/19 = 42.1\%$.

In the fog-enabled heterogeneous RAN, the average content redundancy is defined including the average content redundancy in the MBS level and that in the SBS level, which are listed in (26) and (27), respectively,

$$Redu_{MBS}(h) = \frac{\sum_{m=1}^{M}\sum_{l=1}^{S}\sum_{k\in\{hop_l\le h\}} q_m^l c_{m,k}}{\sum_{l=1}^{S}|\{hop_l \le h\}|} \tag{26}$$

$$Redu_{SBS}(h) = \frac{\sum_{m=1}^{M}\sum_{l=1}^{S}\sum_{i\in\mathbf{A}_l}\sum_{j\in\{hop_i\le h\}} q_m^l e_{m,j}^l}{\sum_{l=1}^{S}\sum_{i\in\mathbf{A}_l}|\{hop_i \le h\}|} \tag{27}$$

where $\{hop_i \le h\}$ indicates the set of the MBSs in the MBS level or the set of SBSs in the SBS level whose hops from base station i are less than h.

B. SIMULATION PARAMETERS IN THE PERFORMANCE EVALUATION

In the simulation, assuming that the size of all contents are the same, the bandwidth on the links are sufficient enough for the content transmission, M contents are provided to users, and they can be cached in vCaches of SBS vCAs and the MBS vCA. The parameters used in the simulations are listed as follows, S ranges from 7 to 19, N ranges from 700 to 1900, M ranges from 900 to 1500 [37], [38], the popularity of contents ranges from 0 to 1.8, h ranges from 1 to 3, $d = 4$.

The communication links include the links between SBS and SBS, SBS and MBS, MBS and MBS, MBS and the core network. The weight of w_1 to w_4 can be set according to the importance of the links. In order to evaluate the performance of the cooperative processing among different levels, we set $w_1 = 1$, $w_2 = 5$, $w_3 = 1$, and $w_4 = 8$ in the simulation. The performance evaluation is also presented comparing the impact of different weight value to the average weighted hops of the proposed algorithms.

C. EVALUATION RESULTS AND DISCUSSIONS

1) THE AVERAGE HIT RATE OF USER REQUESTS

The average hit rate of user requests is evaluated including the average hit rate of user requests in the SBS level, that in the MBS level and that in the embedded IVCN slice in physical networks.

The impact of α to the average hit rate of user requests in the SBS level is investigated for the three algorithms

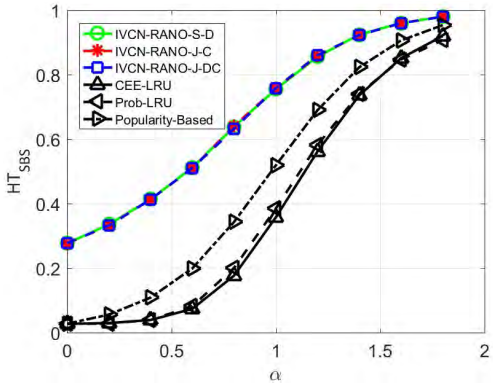

FIGURE 11. The impact of α to HT_{SBS} ($M = 900$, $S = 7$).

compared to the baselines in Fig.11, when $M = 900$ and $S = 7$. It reveals that the average hit rate in the SBS level of the proposed IVCN-RANO is better than that of the three baselines, and the performance of the average hit rate of user requests of IVCN-RANO-J-C, IVCN-RANO-J-DC and IVCN-RANO-S-D are almost the same in the SBS level.

In Fig.12, IVCN-RANO-S-D is evaluated with different M and S. When α is greater, the average hit rate in the SBS level is greater. With the same MBS number S, it can be observed that the more the content number M is, the lower the average hit rate in the SBS level is. The size S of MBS level has small impact to the average hit rate in the SBS level.

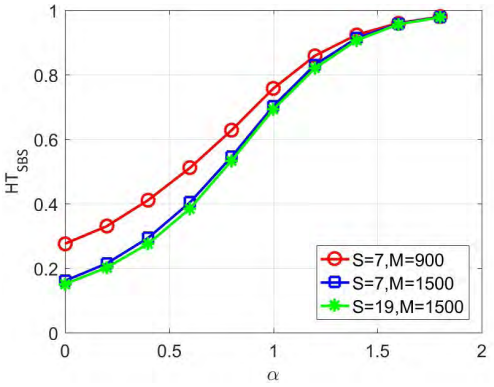

FIGURE 12. The impact of M and S to HT_{SBS} in IVCN-RANO-S-D.

In Fig.13, the average hit rate in the MBS level of the proposed IVCN-RANO is better than that of the three baselines when $M = 900$ and $S = 7$. The performance of the average hit rate in the MBS level of IVCN-RANO-J-C, IVCN-RANO-J-DC and IVCN-RANO-S-D are also nearly the same.

In Fig.14, the impact of M and S to the average hit rate in the MBS level is investigated for IVCN-RANO-S-D. When $S = 7$, the average hit rate in the MBS level is smaller than that when $S = 19$. This is because the greater S leads to bigger size of the MBS level and more cache resources for content service, which contributes to higher hit rate.

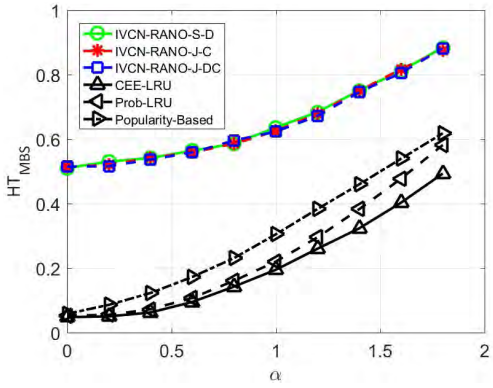

FIGURE 13. The impact of α to HT_{MBS} ($M = 900$, $S = 7$).

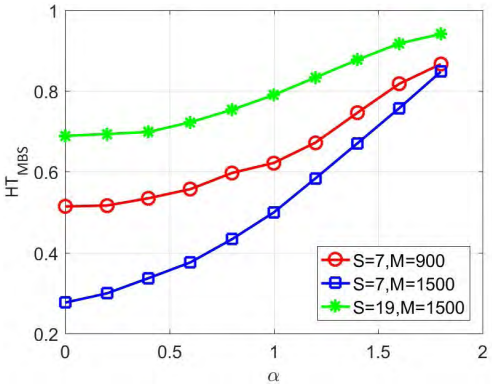

FIGURE 14. The impact of M and S to HT_{MBS} in IVCN-RANO-S-D.

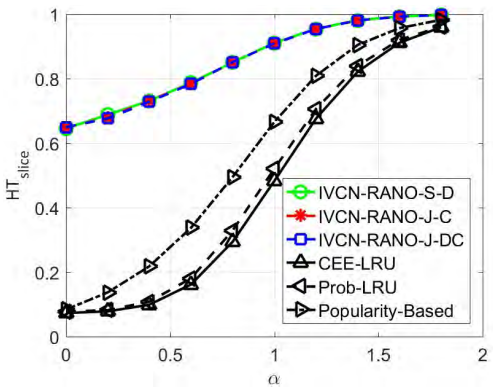

FIGURE 15. The impact of α to HT_{slice} ($M = 900$,$S = 7$).

When $S = 7$, it is obvious that less content numbers ($M = 900$ compared with $M = 1500$) in the MBS level give rise to the higher hit rate in the MBS level.

In Fig.15, the impact of α to the average hit rate of all the content in the embedded IVCN slice in physical network is simulated when $M = 900$ and $S = 7$. The proposed IVCN-RANO performs better than that of the three baselines on the average hit rate of all the content when α changes.

In Fig.16, the impact of M and S to the average hit rate of all the content is given in the embedded IVCN slice in physical

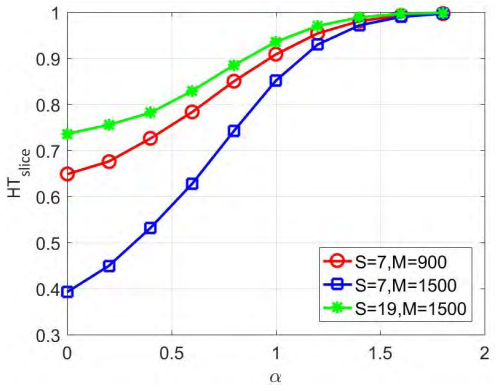

FIGURE 16. The impact of M and S to HT_{slice} in IVCN-RANO-S-D.

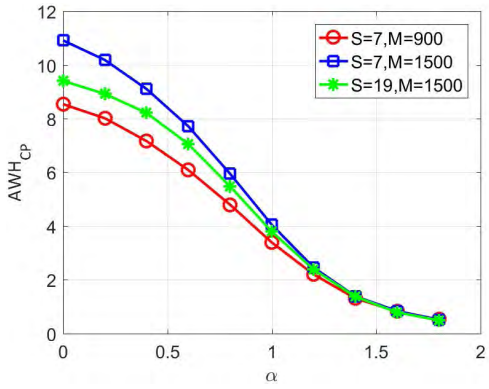

FIGURE 18. The impact of M and S to AWH_{CP} in IVCN-RANO-S-D.

network for IVCN-RANO-S-D. When $S = 7$,$M = 1500$, the average hit rate is smaller than that when $S = 19$ and $M = 1500$. The reason is that the greater S brings bigger size of the MBS level and more cache resources allocated to the slice which contributes to higher hit rate. When $S = 7$, it is visible that the hit rate is higher with less content ($M = 900$ compared with $M = 1500$) in the embedded IVCN slice in physical network.

2) THE AVERAGE WEIGHTED HOPS

In Fig.17, the impact of α to the average weighted hops on the control plane (AWH_{CP}) in IVCN-RANO is shown compared to the three baselines when $M = 900$ and $S = 7$. It indicates that the average weighted hops on the control plane of IVCN-RANO is better than that of the three baselines. The average weighted hops on the control plane in IVCN-RANO-J-C is better than that in IVCN-RANO-J-DC and IVCN-RANO-S-D, since IVCN-RANO-J-C aims to optimize the average weighted hops on the control plane without considering the data plane. It also shows that AWH_{CP} of IVCN-RANO-S-D is nearly the same as that of IVCN-RANO-J-DC.

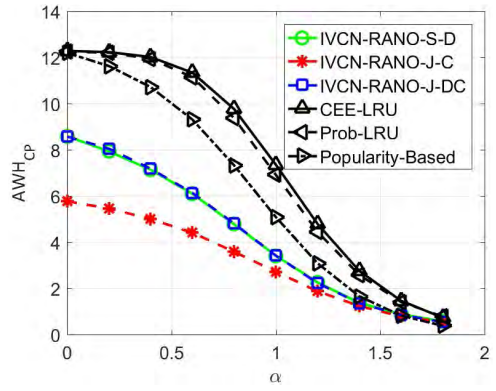

FIGURE 17. The impact of α to AWH_{CP} ($M = 900$,$S = 7$).

In Fig.18, the impact of M and S to AWH_{CP} for IVCN-RANO-S-D is investigated. When $S = 7$,$M = 1500$, the AWH_{CP} is larger than that when $S = 19$,$M = 1500$. S becomes greater means bigger size of the MBS level and more cache resources can be allocated for vCache mapping with content placement in the MBS level as well. In the case of $S = 7$, less content ($M = 900$ compared to $M = 1500$) in the embedded IVCN slice in physical network gives rise to the lower AWH_{CP}.

In Fig.19, the impact of α to the performance of average weighted hops on the data plane (AWH_{DP}) is given. The average weighted hops on the data plane of IVCN-RANO is better than that of the three baselines. Since IVCN-RANO-S-D and IVCN-RANO-J-DC aim at optimizing the weighted hops on the data plane, while IVCN-RANO-J-C optimize the weighted hops without considering the data plane, the average weighted hops on the data plane of IVCN-RANO-S-D and IVCN-RANO-J-DC is better than that of IVCN-RANO-J-C. AWH_{DP} of IVCN-RANO-S-D is almost the same as that of IVCN-RANO-J-DC.

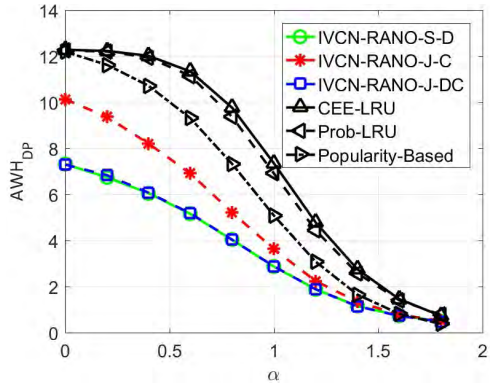

FIGURE 19. The impact of α to AWH_{DP} ($M = 900$,$S = 7$).

In Fig.20, the impact of M and S to AWH_{DP} of IVCN-RANO-S-D is evaluated. When $S = 7$,$M = 1500$, the AWH_{DP} is larger than that when $S = 19$ and $M = 1500$. This is because the greater S brings about bigger size of the MBS level, and more cache resources can be allocated for vCache mapping with content placement in MBS level. When $S = 7$, the less the number of content in the embedded IVCN slice in physical network, the lower AWH_{DP} is.

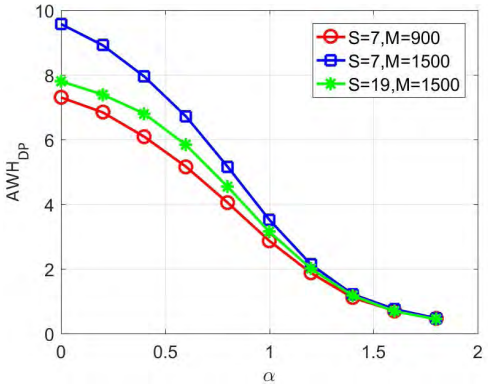

FIGURE 20. The impact of M and S to AWH_{DP} in IVCN-RANO-S-D.

The performance of average weighted hops on the control plane and that on the data plane with $w_1 = 1$, $w_2 = 5$, $w_3 = 1$, and $w_4 = 8$ is compared to the performance of average weighted hops on the control plane and that on the data plane with $w_1 = 1$, $w_2 = 2$, $w_3 = 5$, and $w_4 = 8$ in Fig.21 and Fig.22, respectively.

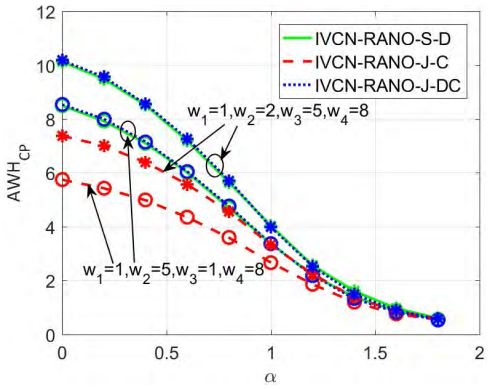

FIGURE 21. The impact of link weight to AWH_{CP}.

In Fig.21, the AWH_CP with $w_1 = 1$, $w_2 = 5$, $w_3 = 1$, and $w_4 = 8$ is lower than that with $w_1 = 1$, $w_2 = 2$, $w_3 = 5$, and $w_4 = 8$. Also, in Fig.22, the AWH_DP with $w_1 = 1$, $w_2 = 5$, $w_3 = 1$, and $w_4 = 8$ is a little lower than that with $w_1 = 1$, $w_2 = 2$, $w_3 = 5$, and $w_4 = 8$.

3) THE AVERAGE CONTENT REDUNDANCY

The impact of α to the average content redundancy of all the content in the SBS level ($Redu_{SBS}$) is shown in Fig.23 when $M = 900$,$S = 7$, and $h = 2$. When the Zipf parameter α increases, the average content redundancy of all the content in the SBS level increases as well, because more popular content cached in the vCache of the SBS vCAs causes higher average content redundancy. That is to say, α is one of the important factors for the duplication copies in the content network. Compared with the three baselines, the average content redundancy of all the content in the SBS level of IVCN-RANO is lower.

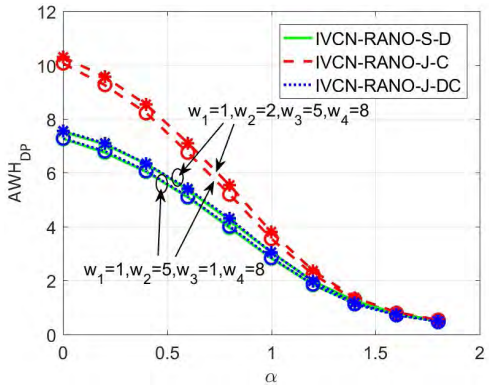

FIGURE 22. The impact of link weight to AWH_{DP}.

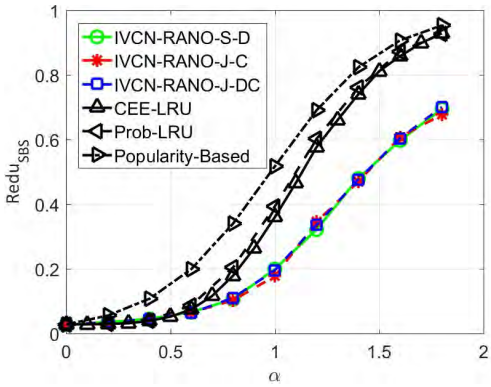

FIGURE 23. The impact of α to $Redu_{SBS}$ ($M = 900$, $S = 7$, $h = 2$).

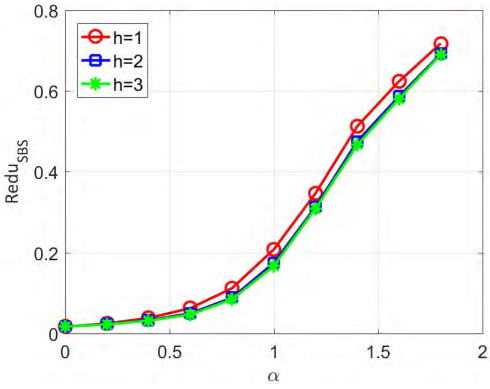

FIGURE 24. The impact of h to $Redu_{SBS}$ in IVCN-RANO-S-D ($M = 1500$, $S = 19$).

In Fig.24, the impact of distance range h to $Redu_{SBS}$ is evaluated for IVCN-RANO-S-D when $M = 1500$ and $S = 19$. It reveals that the impact of h to $Redu_{SBS}$ is slight with the same α.

In Fig.25, the impact of M and S to $Redu_{SBS}$ is investigated for IVCN-RANO-S-D when $h = 2$, which indicates that when M and S becomes large, the average content redundancy in the SBS level increases slightly.

In Fig.26, the impact of α to $Redu_{MBS}$ of IVCN-RANO is evaluated. The proposed IVCN-RANO outperforms in

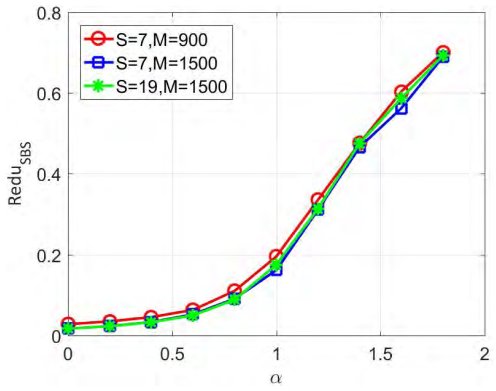

FIGURE 25. The impact of M and S to $Redu_{SBS}$ in IVCN-RANO-S-D ($h = 2$).

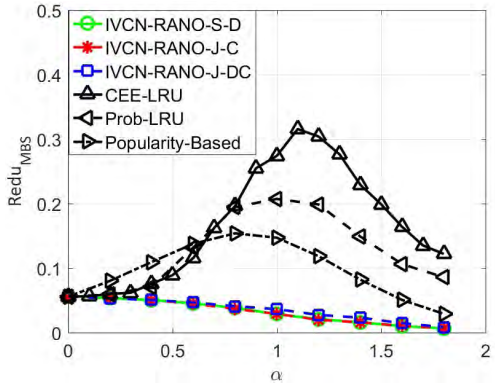

FIGURE 26. The impact of α to $Redu_{MBS}$ ($M = 900, S = 7, h = 2$).

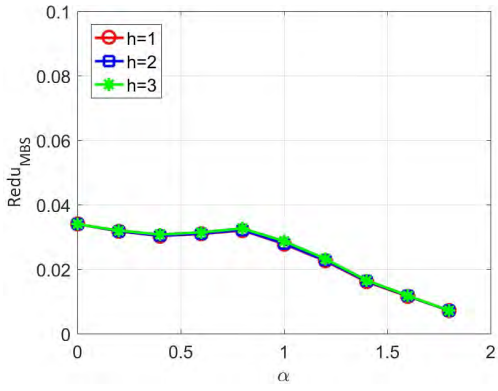

FIGURE 27. The impact of h to $Redu_{MBS}$ in IVCN-RANO-S-D ($M = 1500, S = 19$).

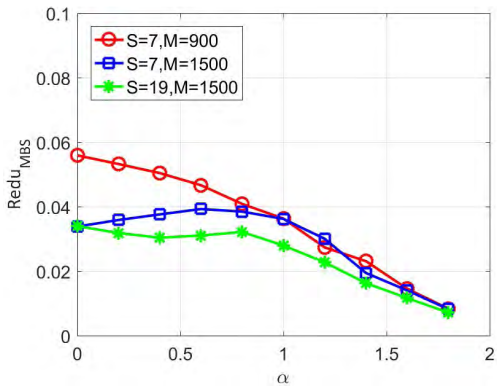

FIGURE 28. The impact of M and S to $Redu_{MBS}$ in IVCN-RANO-S-D ($h = 2$).

$Redu_{MBS}$ compared to that of the three baselines. With the increase of the Zipf parameter α, the average content redundancy in the MBS level decreases slightly because most popular contents have been cached in the SBS level and the remaining contents cached in the MBS level is less popular. The remaining contents are cached only one copy in the MBS level, thus brings about lower average content redundancy in the MBS level (lower that 6%). The low popularity of the content cached in the MBS level leads to the small decrease of $Redu_{MBS}$ with the increasing of Zipf parameter in IVCN-RANO, while the $Redu_{MBS}$ become much larger when α ranges from 0.8 to 1.5 by using the three baselines.

In Fig.27, the impact of distance range h to average content redundancy in the MBS level is addressed for IVCN-RANO-S-D when $M = 1500$ and $S = 19$. It indicates that the impact of h to $Redu_{MBS}$ is slight (It just ranges from 0.8% to 3.3%). The reason is that the contents cached in the MBS level are only one copy in the vCache of the MBS vCA, which give rise to the low value of $Redu_{MBS}$.

In Fig.28, the impact of M and S to $Redu_{MBS}$ is given for IVCN-RANO-S-D when $h = 2$. The impact of M and S to $Redu_{MBS}$ is slight (It just ranges from 0.8% to 5.5%). The reason is that the contents cached in the MBS level are only one copy in the vCache of the MBS vCA, which results in the low value of $Redu_{MBS}$.

VI. CONCLUSION

In this paper, an information-centric virtual content network slicing framework is proposed called IVCN slicing framework based on MANO architecture, which includes vCA, vCache, vLinks and virtual content items. The VNF component based function composition orchestration and networking orchestration are addressed. The service function chaining and optimization of IVCN are investigated both on the data plane and on the control plane. Taking user mobility into consideration, the optimization of IVCN is formulated for networking orchestration as a use case in the fog-enabled heterogeneous RAN. An optimization method is proposed called IVCN-RANO including three optimization algorithms, aiming at minimizing the average forwarding hops to get the contents required by mobile users based on optimal virtual function mapping and content placement in the physical infrastructure. The optimization problem is solved by two heuristic algorithms based on ant colony optimization algorithm. The performance of IVCN-RANO is evaluated by performance metrics including hit rate, average weighted hops as well as the proposed average content redundancy, comparing with CEE-LRU, Prob-LRU and Popularity-Based schemes.

Regarding future research work, there is still some challenging work lie ahead. The first is to investigate

optimization of IVCN slicing considering virtual network function orchestration. Since the optimization problem and evaluation in this paper focuses on the optimization slicing of information-centric radio access networking and caching, the modelling of air interface is not considered in detail, therefore, the second future work is to formulate the optimization problem considering the modelling of air interface, for example, spectrum usage and interference management[45]. The third is that the system model can be enhanced by introducing D2D communication [46], relay and power harvesting techniques [47], for example, offloading traffic via wireless backhaul in order to improve coverage and increase rate for content delivery by unmanned aerial vehicles [48]. The fourth is to formulate the optimization problem of IVCN slicing considering user mobility pattern in detail.

REFERENCES

[1] E. Dahlman, G. Mildh, S. Parkvall, J. Peisa, J. Sachs, Y. Selén, and J. Sköld, "5G wireless access: Requirements and realization," *IEEE Commun. Mag.*, vol. 52, no. 12, pp. 42–47, Dec. 2014.

[2] C. Mouradian, D. Naboulsi, S. Yangui, R. H. Glitho, M. J. Morrow, and P. A. Polakos, "A comprehensive survey on fog computing: State-of-the-art and research challenges," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 1, pp. 416–464, 1st Quart., 2018.

[3] M. Peng, S. Yan, K. Zhang, and C. Wang, "Fog-computing-based radio access networks: Issues and challenges," *IEEE Netw.*, vol. 30, no. 4, pp. 46–53, Jul./Aug. 2016.

[4] S. Retal, M. Bagaa, T. Taleb, and H. Flinck, "Content delivery network slicing: QoE and cost awareness," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Paris, France, May 2017, pp. 1–6.

[5] S. Abdelwahab, B. Hamdaoui, M. Guizani, and T. Znati, "Network function virtualization in 5G," *IEEE Commun. Mag.*, vol. 54, no. 4, pp. 84–91, Apr. 2016.

[6] J. G. Herrera and J. F. Botero, "Resource allocation in NFV: A comprehensive survey," *IEEE Trans. Netw. Service Manage.*, vol. 13, no. 3, pp. 518–532, Sep. 2016.

[7] K. Samdanis, X. C. Perez, and V. Sciancalepore, "From network sharing to multi-tenancy: The 5G network slice broker," *IEEE Commun. Mag.*, vol. 54, no. 7, pp. 32–39, Jul. 2016.

[8] C. Liang, F. R. Yu, and X. Zhang, "Information-centric network function virtualization over 5G mobile wireless networks," *IEEE Netw.*, vol. 29, no. 3, pp. 68–74, May 2015.

[9] Y. Liu, J. C. Point, K. V. Katsaros, V. Glykantzis, M. S. Siddiqui, and E. Escalona, "SDN/NFV based caching solution for future mobile network (5G)," in *Proc. IEEE Eur. Conf. Netw. Commun. (EuCNC)*, Oulu, Finland, Jun. 2017, pp. 1–5.

[10] X. Li, X. Wang, C. Zhu, W. Cai, and V. C. M. Leung, "Caching-as-a-service: Virtual caching framework in the cloud-based mobile networks," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, Hong Kong, China, Apr. 2015, pp. 372–377.

[11] I. Benkacem, T. Taleb, M. Bagaa, and H. Flinck, "Optimal VNFs placement in CDN slicing over multi-cloud environment," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 3, pp. 616–627, Mar. 2018.

[12] S. Fu, J. Liu, and W. Zhu, "Multimedia content delivery with network function virtualization: The energy perspective," *IEEE Multimedia*, vol. 24, no. 3, pp. 38–47, Jul./Sep. 2017.

[13] N. T. Jahromi, R. H. Glitho, A. Larabi, and R. Brunner, "An NFV and microservice based architecture for on-the-fly component provisioning in content delivery networks," in *Proc. IEEE Annu. Consum. Commun. Netw. Conf. (CCNC)*, Las Vegas, NV, USA, Jan. 2018, pp. 1–7.

[14] T. Soenen, W. Tavernier, G. Xilouris, S. Kolometsos, F. Vicens, E. M. Uriarte, and S. Siddiqui, "Service specific management and orchestration for a content delivery network," in *Proc. IEEE Conf. Netw. Soft. (NetSoft)*, Montreal, QC, Canada, Jun. 2018, pp. 326–328.

[15] N. Herbaut, D. Negru, D. Dietrich, and P. Papadimitriou, "Service chain modeling and embedding for NFV-based content delivery," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Paris, France, May 2017, pp. 1–7.

[16] N. Herbaut, D. Negru, D. Dietrich, and P. Papadimitriou, "Dynamic deployment and optimization of virtual content delivery networks," *IEEE MultimediaMag.*, vol. 24, no. 3, pp. 28–37, Aug. 2017.

[17] N. Herbaut, D. Negru, Y. Chen, P. A. Frangoudis, and A. Ksentini, "Content delivery networks as a virtual network function a win-win ISP-CDN collaboration," in *Proc. IEEE Globecom*, Washington, DC, USA, Dec. 2016, pp. 1–6.

[18] L. Yala, P. A. Frangoudis, G. Lucarelli, and A. Ksentini, "Cost and availability aware resource allocation and virtual function placement for CDNaaS provision," *IEEE Trans. Netw. Service Manag.*, vol. 15, no. 4, pp. 1334–1348, Dec. 2018.

[19] N. T. Jahromi, S. Kianpisheh, and R. H. Glitho, "Online VNF placement and chaining for value-added services in content delivery networks," 2018, *arXiv:1806.04580*. [Online]. Available: https://arxiv.org/abs/1806.04580

[20] M. Dieye, S. Ahvar, J. Sahoo, E. Ahvar, R. Glitho, H. Elbiaze, and N. Crespi, "CPVNF: Cost-efficient proactive VNF placement and chaining for value-added services in content delivery networks," *IEEE Trans. Netw. Service Manage.*, vol. 15, no. 2, pp. 774–786, Jun. 2018.

[21] K. Wang, F. R. Yu, and H. Li, "Information-centric virtualized cellular networks with device-to-device communications," *IEEE Trans. Veh. Technol.*, vol. 65, no. 11, pp. 9319–9329, Nov. 2016.

[22] Y. Zhou, F. R. Yu, J. Chen, and Y. Kuo, "Resource allocation for information-centric virtualized heterogeneous networks with in-network caching and mobile edge computing," *IEEE Trans. Veh. Technol.*, vol. 66, no. 12, pp. 11339–11351, Dec. 2017.

[23] T. D. Tran and L. B. Le, "Joint resource allocation and content caching in virtualized content-centric wireless networks," *IEEE Access*, vol. 6, pp. 11329–11341, 2018.

[24] T. D. Tran and L. B. Le, "Joint resource allocation and content caching in virtualized multi-cell wireless networks," in *Proc. IEEE Globecom*, Singapore, Dec. 2017, pp. 1–6.

[25] H. Jin, H. Lu, and C. Zhao, "Content-oriented network slicing optimization based on cache-enabled hybrid radio access network," *EURASIP J. Wireless Commun. Netw.*, vol. 2018, no. 2, pp. 1–24, Jan. 2018.

[26] R. Huo, F. R. Yu, T. Huang, R. Xie, J. Liu, V. C. M. Leung, and Y. Liu, "Software defined networking, caching, and computing for green wireless networks," *IEEE Commun. Mag.*, vol. 54, no. 11, pp. 185–193, Nov. 2016.

[27] C. Liang and F. R. Yu, "Enhancing mobile edge caching with bandwidth provisioning in software-defined mobile networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Paris, France, May 2017, pp. 1–6.

[28] Z. Hu, Z. Zheng, T. Wang, and L. Song, "Caching as a service: Small-cell caching mechanism design for service providers," *IEEE Trans. Wireless Commun.*, vol. 15, no. 10, pp. 6992–7004, Oct. 2016.

[29] 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.

[30] *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.

[31] *Network Functions Virtualisation (NFV); Management and Orchestration; Network Service Templates Specification; Release 2*, document ETSI GS NFV-IFA 014 V2.4.1, 2018.

[32] R. Langar, N. Bouabdallah, and R. Boutaba, "A comprehensive analysis of mobility management in MPLS-based wireless access networks," *IEEE/ACM Trans. Netw.*, vol. 16, no. 4, pp. 918–931, Aug. 2008.

[33] M. E. J. Newman, "The structure and function of complex networks," *SIAM Rev.*, vol. 45, no. 2, pp. 167–256, 2003.

[34] H. Jin, D. Xu, C. Zhao, and D. Liang, "Information-centric mobile caching network frameworks and caching optimization: A survey," *EURASIP J. Wireless Commun. Netw.*, vol. 2017, no. 33, pp. 1–32, Feb. 2017.

[35] C. Berthet, "Approximation of LRU caches miss rate: Application to power-law popularities," 2017, *arXiv:1705.10738*. [Online]. Available: https://arxiv.org/abs/1705.10738

[36] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and Zipf-like distributions: Evidence and implications," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, New York, NY, USA, Mar. 1999, pp. 126–134.

[37] H. Nakayama, S. Ata, and I. Oka, "Caching algorithm for content-oriented networks using prediction of popularity of contents," in *Proc. IEEE Int. Symp. Integr. Netw. Manag. (IM)*, Ottawa, ON, Canada, May 2015, pp. 1171–1176.

[38] P. Yang, N. Zhang, S. Zhang, L. Yu, J. Zhang, and X. S. Shen, "Content popularity prediction towards location-aware mobile edge caching," *IEEE Trans. Multimedia*, vol. 21, no. 4, pp. 915–929, Apr. 2019.
[39] N. Ben Hassine, D. Marinca, P. Minet, and D. Barth, "Popularity prediction in content delivery networks," in *Proc. IEEE Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun. (PIMRC)*, Hong Kong, China, Aug./Sep. 2015, pp. 2083–2088.
[40] T. Hou, G. Feng, S. Qin, and W. Jiang, "Proactive content caching by exploiting transfer learning for mobile edge computing," in *Proc. IEEE Globecom*, Singapore, Dec. 2017, pp. 1–6.
[41] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon, "I tube, you tube, everybody tubes: Analyzing the world's largest user generated content video system," in *Proc. ACM Int. Meas. Conf. (IMC)*, New York, NY, USA, Oct. 2007, pp. 1–14.
[42] X. Cheng, C. Dale, and J. Liu, "Statistics and social network of YouTube videos," in *Proc. Int. Workshop Qual. Service*, Enschede, The Netherlands, Jun. 2008, pp. 229–238.
[43] C. Baolin, *Theory and Algorithms for Optimization*, 2nd ed. Beijing, China: Tsinghua Univ. Press, 2005.
[44] B. C. Mohan and R. Baskaran, "A survey: Ant colony optimization based recent research and implementation on several engineering domain," *Expert Syst. Appl.*, vol. 39, no. 4, pp. 4618–4627, Mar. 2012.
[45] N. Zhao, X. Liu, F. R. Yu, M. Li, and V. C. M. Leung, "Communications, caching, and computing oriented small cell networks with interference alignment," *IEEE Commun. Mag.*, vol. 54, no. 9, pp. 29–35, Sep. 2016.
[46] Z. Zhou, M. Peng, and Z. Zhao, "Joint data-energy beamforming and traffic offloading in cloud radio access networks with energy harvesting-aided D2D communications," *IEEE Trans. Wireless Commun.*, vol. 17, no. 12, pp. 8094–8107, Dec. 2018.
[47] Z. Zhou, M. Peng, Z. Zhao, W. Wang, and R. S. Blum, "Wireless-powered cooperative communications: Power-splitting relaying with energy accumulation," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 969–982, Apr. 2016.
[48] N. Zhao, F. Cheng, F. R. Yu, J. Tang, Y. Chen, G. Gui, and H. Sari, "Caching UAV assisted secure transmission in hyper-dense networks based on interference alignment," *IEEE Trans. Commun.*, vol. 66, no. 5, pp. 2281–2294, May 2018.

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

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.

YI JIN received the B.Eng. degree from Northeastern University, 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 wireless networks, mobile cloud computing systems, 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, where he is currently a Professor. His current research interests include emerging technologies of short-range wireless communication, cognitive radios, mobile edge computing, and the Internet of Things.