Chapter 4

Management and Orchestration of Network Slices in 5G, Fog, Edge and Clouds

Adel Nadjaran Toosi, Redowan Mahmud, Qinghua Chi and Rajkumar Buyya

Abstract—Network slicing allows network operators to build multiple isolated virtual networks on a shared physical network to accommodate a wide variety of services and applications. With network slicing, service providers can provide a cost-efficient solution towards meeting diverse performance requirements of deployed applications and services. Despite slicing benefits, End-to-End orchestration and management of network slices is a challenging and complicated task. In this chapter, we intend to survey all the relevant aspects of network slicing, with the focus on networking technologies such as Software-defined networking (SDN) and Network Function Virtualization (NFV) in 5G, Fog/Edge and Cloud Computing platforms. To build the required background, this chapter begins with a brief overview of 5G, Fog/Edge and Cloud computing, and their interplay. Then we cover the 5G vision for network slicing and extend it to the Fog and Cloud computing through surveying the state-of-the-art slicing approaches in these platforms. We conclude the chapter by discussing future directions, analyzing gaps and trends towards the network slicing realization.

4.1 Introduction

The major digital transformation happening all around the world these days has introduced a wide variety of applications and services ranging from smart cities and vehicle to vehicle (V2V) communication to virtual reality (VR)/augmented reality (AR) and remote medical surgery. Design and implementation of a network that can simultaneously provide the essential connectivity and performance requirements of all these applications with a single set of network functions not only is massively complex but also is prohibitively expensive. The 5G Infrastructure Public-Private Partnership (5G-PPP) has identified various use case families of enhanced mobile broadband (eMBB), massive machine-type communications (mMTC), and ultra-reliable low latency communication (uRLLC) or Critical Communications that would simultaneously run and share the 5G physical multi-service network [1]. These applications essentially have very different Quality of Service (QoS) requirements and transmission characteristics. For instance, Video-on-demand streaming applications in eMMB category require very high bandwidth and transmitting a large amount of content. While mMTC applications, such as Internet of Things (IoT), typically have a multitude of low throughput devices. The differences between these use cases show that the *one-size-fits-all* approach of the traditional networks does not satisfy different requirements of all these vertical services.

A cost-efficient solution towards meeting these requirements is slicing physical network into multiple isolated logical networks. Similar to server virtualization technology successfully used in Cloud computing era, network slicing intends to build a form of virtualization that partitions a shared physical network infrastructure into multiple end-to-end level logical networks allowing for traffic grouping and tenants' traffic isolation. Network slicing is considered as the critical enabler of the 5G network

where vertical service providers can flexibly deploy their applications and services based on the requirements of their service. In other words, network slicing provides a *Network-as-a-Service* (NaaS) model which allows service providers to build and set up their own networking infrastructure according to their demands and customize it for diverse and sophisticated scenarios.

Software Defined Networking (SDN) and Network Function Virtualization (NFV) can serve as building blocks of network slicing by facilitating network programmability and virtualization. Software-defined networking (SDN) is a promising approach to computer networking that separates the tightly coupled control and data planes of traditional networking devices. Thanks to this separation, SDN can provide a logically centralized view of the network in a single point of management to run network control functions. NFV is another trend in networking gaining momentum quickly with the aim of transferring network functions from proprietary hardware to software-based applications executing on general-purpose hardware. NFV intends to reduce the cost and increase the elasticity of network functions by building virtual network functions (VNFs) that are connected or chained together to build communication services.

With this in mind, in this chapter, we aim to review the state of the art literature on network slicing in 5G, Edge/Fog and Cloud computing, and identify the spectrum challenges and obstacles must be addressed to achieve the ultimate realization of this concept. We begin with a brief introduction of 5G, Edge/Fog, and Clouds and their interplay. Then, we outline the 5G vision for network slicing and identify a generic framework for 5G network slicing. We then review research and projects related to network slicing in Cloud computing context while we focus on SDN and NFV

technologies. Further, we explore network slicing advance in emerging Fog and Edge Cloud computing. This leads us to identify the key unresolved challenges of network slicing within these platforms. Concerning this review, we discuss the Gaps and trends towards the realization of network slicing vision in Fog and Edge and Software-defined Cloud computing. Finally, we conclude the chapter.

Table 4.1 lists various acronyms and abbreviations referenced throughout the chapter.

Table 4.1- Acronyms and Abbreviations

V2V	Vehicle to Vehicle
VR	Virtual Reality
AR	Augmented Reality
5G	5th generation mobile networks or 5th generation wireless systems
eMBB	enhanced Mobile Broadband
mMTC	massive Machine-Type communications
uRLLC	ultra-Reliable Low Latency communication
QoS	Quality of Service
IoT	Internet of Things
SDN	Software Defined Networking
NFV	Network Function Virtualization
VNF	Virtualized Network Function
MEC	Mobile Edge Computing
NaaS	Network-as-a-Service
NFaaS	Network function as a Service
SDC	Software-defined Clouds
VM	Virtual Machine
VPN	Virtual Private Network
NAT	Network Address Translation
SFC	Service Function Chaining
SLA	Service Level Agreement
CRAN	Cloud Radio Access Network
RRH	Remote Radio Head
BBU	Baseband Unit
FRAN	Fog radio access network

4.2 Background

5G: The renovation of telecommunications standards is a continuous process. Practicing this, 5th generation mobile network or 5th generation wireless system, commonly called 5G, has been proposed as the next telecommunications standards beyond the current 4G/IMT Advanced standards [3]. The wireless networking architecture of 5G follows 802.11ac IEEE wireless networking criterion and operates on millimeter wave bands. It can encapsulate Extremely high frequency (EHF) from 30 to 300 gigahertz (GHz) that ultimately offers higher data capacity and low latency communication [4].

The formalization of 5G is still in its early stage and expected to be mature by 2020. However, the main intentions of 5G include enabling Gbps data rate in a real network with least round trip latency and offering long-term communication among the large number of connected devices through high fault tolerant networking architecture [1]. Also, it targets to improve the energy usage both for the network and the connected devices. Moreover, it is anticipated that 5G will be more flexible, dynamic and manageable compared to the previous generations [5].

Cloud Computing: Cloud computing is expected to be an inseparable part of 5G services for providing an excellent backend for applications running on the accessing devices. During last decade, Cloud has evolved as a successful computing paradigm for delivering on-demand services over the Internet. The Cloud data centers adopted virtualization technology for efficient management of resources and services. Advances in server virtualization contributed to the cost-efficient management of computing resources in the Cloud data centers.

Recently, the virtualization notion in Cloud data centers, thanks to the advances in SDN and NFV, has extended to all resources including compute, storage, and networks which formed the concept of Software Defined Clouds (SDC) [2]. SDC aims to utilize the advances in areas of Cloud computing, system virtualization, SDN, and NFV to enhance resource management in data centers. In addition, Cloud is regarded as the foundation block for *Cloud Radio Access Network (CRAN),* an emerging cellular framework that aims at meeting ever-growing end-users demand on 5G. In CRAN, the traditional base stations are split into radio and baseband parts. The radio part resides in the base station in the form of Remote Radio Head (RRH) unit and the baseband part in placed to Cloud for creating a centralized and virtualized Baseband Unit (BBU) pool for different base stations.

Mobile Edge Computing (MEC): Among the user proximate computing paradigms, Mobile Edge Computing (MEC) is considered as one of the key enablers of 5G. Unlike CRAN [48], in MEC, base stations and access points are equipped with Edge servers that take care of 5G related issues at the edge network. MEC facilitates a computationally enriched distributed RAN architecture upon the LTE-based networking. Ongoing researches on MEC targets real-time context awareness [49], dynamic computation offloading [50], energy efficiency [51] and multi-media caching [52] for 5G networking.

Edge and Fog Computing: Edge and Fog computing are coined to complement remote Cloud to meet the service demand of a geographically distributed large number of IoT devices. In Edge computing, the embedded computation capabilities of IoT devices or local resources accessed via ad-hoc networking are used to process IoT data. Usually, Edge computing paradigm is well suited to perform light computational tasks and does

not probe global Internet unless intervention of remote (core) Cloud is required. However, not all the IoT devices are computationally enabled, or local Edge resources are computational-enriched to execute different large-scale IoT applications simultaneously. In this case, executing latency sensitive IoT applications at remote Cloud can degrade the QoS significantly [60]. Moreover, a huge amount of IoT workload sent to remote Cloud can flood the global internet and congest the network. Therefore, Fog computing is coined that offers infrastructure and software services through distributed Fog nodes to execute IoT applications within the network [54].

In Fog computing, traditional networking devices such as routers, switches, set-top boxes and proxy servers along with dedicated Nano-servers and Micro-datacenters can act as Fog nodes and create a wide area Cloud-like services both in independent or clustered manner [55]. Mobile Edge servers or Cloudlets [53] can also be regarded as Fog nodes to conduct their respective jobs in Fog enabled Mobile Cloud Computing and MEC. In some cases, Edge and Fog computing are used interchangeably although, in a broader perspective, Edge is considered as a subset of Fog Computing [56]. However, in Edge and Fog computing, the integration of 5G has already been discussed in terms of bandwidth management during computing instance migration [57] and SDN-enabled IoT resource discovery [58]. The concept of Fog radio access network (FRAN) [59] is also getting attention from both academia and industry where Fog resources are used to create BBU pool for the base stations.

Working principle of these computing paradigms largely depends on virtualization techniques. The alignment of 5G with different computing paradigms can also be analyzed through the interplay between network and resource virtualization

techniques. Network Slicing is one of the key features of 5G network virtualization. Computing paradigms can also extend the vision of 5G network slicing into data center and Fog nodes. By the latter, we mean that the vision of network slicing can be applied to the shared data center network infrastructure and Fog networks to provide an end-to-end logical network for applications by establishing a full-stack virtualized environment. This form of network slicing can also be expanded beyond a data center networks into multi-Clouds or even cluster of Fog nodes [14]. Whatever the extension may be, this creates a new set of challenges to the network, including Wide Area Network (WAN) segments, cloud data centers (DCs) and Fog resources.

4.3 Network Slicing in 5G

In recent years, numerous research initiatives are taken by industries and academia to explore different aspects of 5G. Network architecture and its associated physical and MAC layer management are among the prime focuses of current 5G research works. The impact of 5G in different real-world applications, sustainability, and quality expectations are also getting predominant in the research arena. However, among the ongoing researches in 5G, network slicing is drawing more attractions since this distinctive feature of 5G aims at supporting diverse requirements at the finest granularity over a shared network infrastructure [6][7].

Network slicing in 5G refers to sharing a physical network's resources to multiple virtual networks. More precisely, network slices are regarded as a set of virtualized networks on the top of a physical network [8]. The network slices can be allocated to specific applications/services, use cases or business models to meet their requirements.

Each network slice can be operated independently with its own virtual resources, topology, data traffic flow, management policies, and protocols. Network slicing usually requires implementation in an end-to-end manner to support co-existence of heterogeneous systems [9].

The network slicing paves the way for customized connectivity among a high number of inter-connected end-to-end devices. It enhances network automation and leverages the full capacity of SDN and NFV. Also, it helps to make the traditional networking architecture scalable according to the context. Since network slicing shares a common underlying infrastructure to multiple virtualized networks, it is considered as one of the most cost-effective ways to use network resources and reduce both capital and operational expenses [10]. Besides, it ensures that the reliability and limitations (congestion, security issues) of one slice do not affect the others. Network slicing assists isolation and protection of data, control and management plane that enforce security within the network. Moreover, network slicing can be extended to multiple computing paradigms such as Edge [11], Fog [14] and Cloud that eventually improves their interoperability and helps to bring services closer to the end user with less Service Level Agreement (SLA) violations [12].

Apart from the benefits, the network slicing in current 5G context is subjected to diversified challenges, however. Resource provisioning among multiple virtual networks is difficult to achieve since each virtual network has a different level of resource affinity and it can be changed with the course of time. Besides, mobility management and wireless resource virtualization can intensify the network slicing problems in 5G. End-to-End slice orchestration and management can also make network slicing complicated.

Recent researches in 5G network slicing mainly focus on addressing the challenges through efficient network slicing frameworks. Extending the literature [12][13], we depicted a generic framework for 5G network slicing in Figure 4.1 The framework consists of three main layers: *Infrastructure layer*, *Network Function layer*, and *Service layer*.

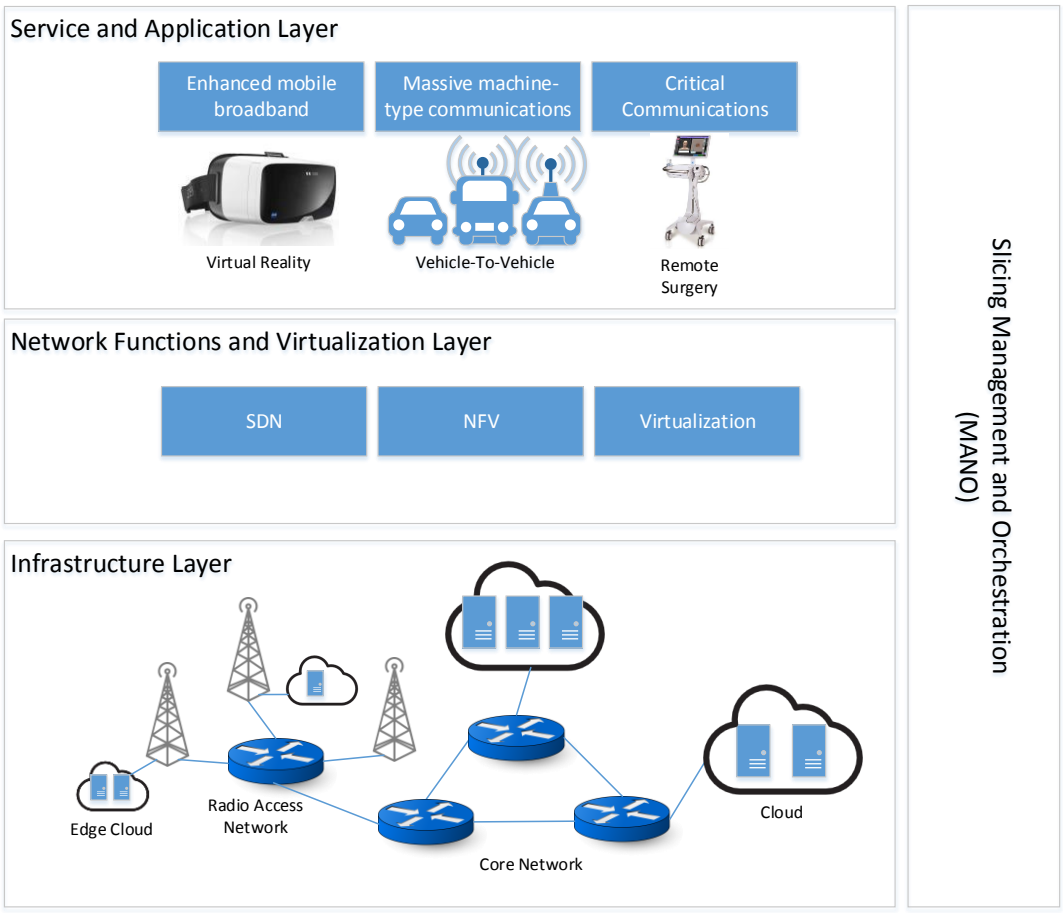

Figure 4.1: Generic 5G Slicing Framework.

Infrastructure layer: The infrastructure layer defines the actual physical network architecture. It can be expanded from Edge Cloud to remote Cloud through radio access network and the core network. Different software defined techniques are encapsulated to facilitate resource abstraction within the core network and the radio access network. Besides, in this layer, several policies are conducted to deploy, control, manage and

orchestrate the underlying infrastructure. This layer allocates resources (compute, storage, bandwidth, etc.) to network slices in such way that upper layers can get access to handle them according to the context.

Network Function and Virtualization Layer: The network function and virtualization layer executes all the required operations to manage the virtual resources and network function's life cycle. It also facilitates optimal placement of network slices to virtual resources and chaining of multiple slices so that they can meet specific requirements of a particular service or application. SDN, NFV and different virtualization techniques are considered as the significant technical aspect of this layer. This layer explicitly manages the functionality of core and local radio access network. It can handle both coarse-grained and fine-grained network functions efficiently.

Service and Application Layer: The service and application layer can be composed by connected vehicles, virtual reality appliances, mobile devices, etc. having a specific use case or business model and represent certain utility expectations from the networking infrastructure and the network functions. Based on requirements or high-level description of the service or applications, virtualized network functions are mapped to physical resources in such way that SLA for the respective application or service does not get violated.

Slicing Management and Orchestration (MANO): The functionality of the above layers are explicitly monitored and managed by the slicing management and orchestration layer. The main task of this layer includes;

1. Creation of virtual network instances upon the physical network by using the functionality of the infrastructure layer.

2. Mapping of network functions to virtualized network instances to build a service chain with the association of network function and virtualization layer.
3. Maintaining communication between service/application and the network slicing framework to manage the lifecycle of virtual network instances and dynamically adapt or scale the virtualized resources according to the changing context.

The logical framework of 5G network slicing is still evolving. Retaining the basic structure, extension of this framework to handle the future dynamics of network slicing can be a potential approach to further standardization of 5G.

According to Huawei high-level perspective of 5G network [42], Cloud-Native network architecture for 5G has the following characteristics: 1) it provides Cloud data center based architecture and logically independent network slicing on the network infrastructure to support different application scenarios. 2) It uses Cloud-RAN[1] to build radio access networks (RAN) to provide a substantial number of connections and implement 5G required on-demand deployments of RAN functions. 3) It provides simpler core network architecture and provides on-demand configuration of network functions via user and control plane separation, unified database management, and component-based functions, and. 4) In automatic manner, it implements network slicing service to reduce operating expenses.

In the following section, we intend to review the state-of-the-art related work on network slice management happening in Cloud computing literature. Our survey in this area can help researcher to apply advances and innovation in 5G and Clouds reciprocally.

[1] CLOUD-RAN (CRAN) is a centralized architecture for radio access network (RAN) in which the radio transceivers are separated from the digital baseband processors. This means that operators can centralize multiple base band units in one location. This simplifies the amount of equipment needed at each individual cell site. Ultimately, the network functions in this architecture become virtualized in the Cloud.

4.4 Network Slicing in Software Defined Clouds

Virtualization technology has been the cornerstone of the resource management and optimization in Cloud data centers for the last decade. Many research proposals have been expressed for VM placement and Virtual Machine (VM) migration to improve utilization and efficiency of both physical and virtual servers [15]. In this section, we focus on the state of the art network-aware VM/VNF management in line with the aim of the report, i.e., network slicing management for SDCs. Figure 4.2 illustrates our proposed taxonomy of network-aware VM/VNF management in SDCS. Our taxonomy classifies existing works based on the objective of the research, the approach used to address the problem, the exploited optimization technique, and finally the evaluation technique used to validate the approach. In the remaining parts of this section, we cover network slicing from three different perspectives and map them to the proposed taxonomy: Network-aware VM management, Network-aware VM migration, and VNF management.

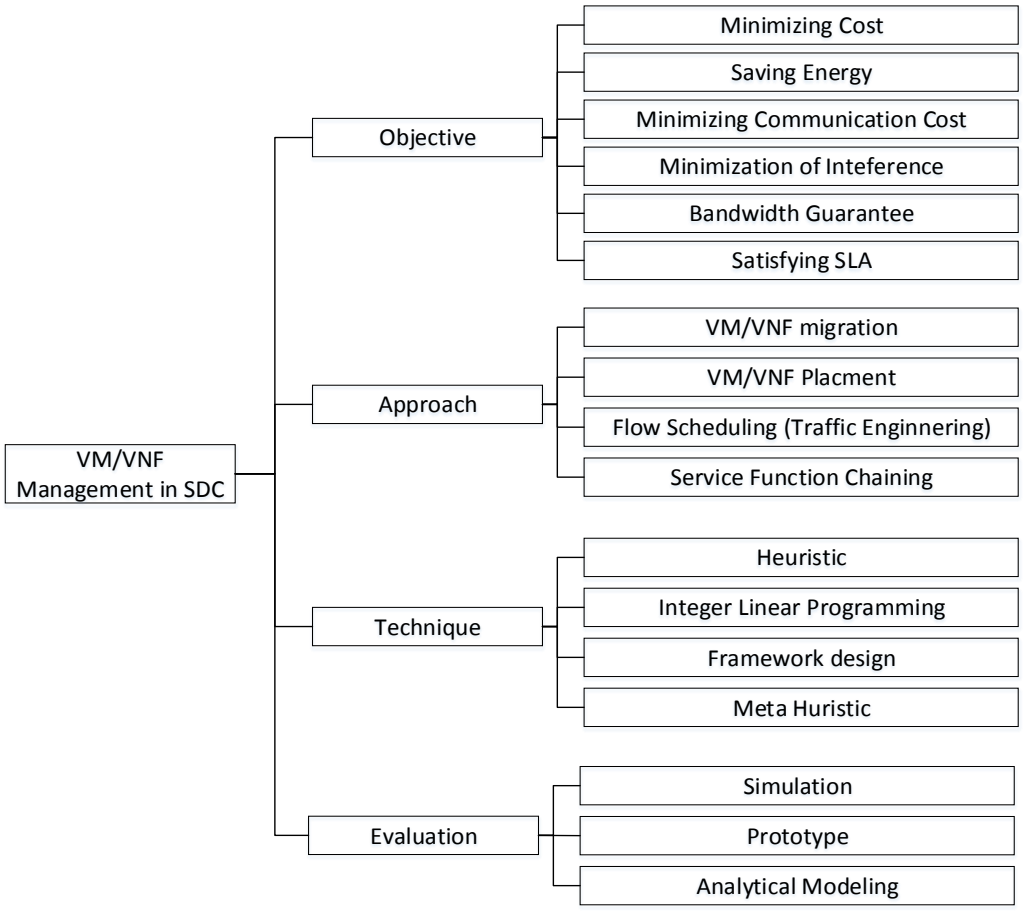

Figure 4.2: Taxonomy of network-aware VM/VNF Management in software-defined Clouds

4.4.1 Network-aware Virtual Machines Management

Cziva et al. [15] present an orchestration framework to exploit time-based network information to live migrate VMs and minimize the network cost. Wang et al. [16] propose a VM placement mechanism to reduce the number of hops between communicating VMs, save energy, and balance the network load. Remedy [17] relies on SDN to monitor the state of the network and estimate the cost of VM migration. Their technique detects congested links and migrates VMs to remove congestion on those links.

Jiang et al. [18] worked on joint VM placement and network routing problem of data centers to minimize network cost in real-time. They proposed an online algorithm to optimize the VM placement and data traffic routing with dynamically adapting traffic loads. VMPlanner [19] also optimizes VM placement and network routing. The solution includes VM grouping that consolidates VMs with high inter-group traffic, VM group placement within a rack, and traffic consolidation to minimize the rack traffic. Jin et al. [21] studied joint host-network optimization problem. The problem is formulated as an integer linear problem which combines VM placement and routing problem. Cui et al. [20] explore the joint policy-aware and network-aware VM migration problem and present a VM management to reduce network-wide communication cost in data center networks while considering the policies regarding the network functions and middleboxes. Table 4.2 summarizes the research projects on network-aware VM management.

Table 4.2 - Network-aware Virtual Machines Management

Project	Objectives	Approach/Technique	Evaluation
Cziva et al. [15]	Minimization of the network communication cost	VM migration – Framework Design	Prototype
Wang et al. [16]	Reducing the number of hops between communicating VMs and network power consumption	VM placement – Heuristic	Simulation
Remedy [17]	Removing congestion in the network	VM migration – Framework Design	Simulation
Jiang et al. [18]	Minimization of the network communication cost	VM Placement and Migration – Heuristic (Markov approximation)	Simulation
VMPlanner [19]	Reducing network power consumption	VM placement and traffic flow routing - Heuristic	Simulation
PLAN [20]	Minimization of the network communication cost while meeting network policy requirements	VM Placement - Heuristic	Prototype/Simulation

4.4.2 Network-aware Virtual Machine Migration Planning

A large body of literature focused on improving the efficiency of VM migration mechanism [22]. Bari et al. [23] propose a method for finding an efficient migration plan. They try to find a sequence of migrations to move a group of VMs to their final destinations while migration time is minimized. In their method, they monitor residual bandwidth available on the links between source and destination after performing each step in the sequence. Similarly, Ghorbani et al. [24] propose an algorithm to generate an ordered list of VMs to migrate and a set of forwarding flow changes. They concentrate on imposing bandwidth guarantees on the links to ensure that link capacity is not violated during the migration. The VM migration planning problem is also tackled by Li et al. [25] where they address the workload-aware migration problem and propose methods for

selection of candidate virtual machines, destination hosts, and sequence for migration. All these studies focus on the migration order of a group of VMs while taking into account network cost. Xu et al. [26] propose an interference-aware VM live migration plan called *iAware* that minimizes both migration and co-location interference among VMs. Table 4.3 summarizes the research projects on VM migration planning.

Table 4.3 -Virtual Machine Migration Planning

Project	Objectives	Approach/Technique	Evaluation
Bari et al. [23]	Finding sequence of migrations to while migration time is minimized	VM migration – Heuristic	Simulation
Ghorbani et al. [24]	Finding sequence of migrations while imposing bandwidth guarantees	VM migration – Heuristic	Simulation
Li et al. [25]	Finding sequence of migrations and destination hosts to balance the load	VM migration – Heuristic	Simulation
iAware [26]	Minimization of migration and co-location interference among VMs	VM migration – Heuristic	Prototype/Simulation

4.4.3 Virtual Network Functions Management

Network Functions Virtualization (NFV) is an emerging paradigm where network functions such as firewalls, Network Address Translation (NAT), Virtual Private Network (VPN), etc. are virtualized and divided up into multiple building blocks called Virtualized Network Functions (VNFs). VNFs are often chained together and build Service Function Chains (SFC) to deliver a required network functionality. Han et al. [27] present a comprehensive survey of key challenges and technical requirements of NFV where they present an architectural framework for NFV. They focus on the efficient instantiation, placement and migration of VNFs and network performance. VNF-P is a model proposed by Moens and Turck [28] for efficient placement of VNFs. They

propose a NFV burst scenario in a hybrid scenario in which the base demand for network function service is handled by physical resources while the extra load is handled by virtual service instances. Cloud4NFV [29] is a platform following the NFV standards by European Telecommunications Standards Institute (ETSI) to build Network Function as a Service using a Cloud platform. Their VNF Orchestrator exposes RESTful APIs allowing VNF deployment. A Cloud platform such as OpenStack supports management of virtual infrastructure at the background. vConductor [30] is another NFV management system proposed by Shen et al. for the end-to-end virtual network services. vConductor has simple graphical user interfaces (GUIs) for automatic provisioning of virtual network services and supports the management of VNFs and existing physical network functions. MORSA [31] proposed as part of vConductor to perform virtual machine (VM) placement for building NFV infrastructure in the presence of conflicting objectives of involving stakeholders such as users, Cloud providers, and telecommunication network operators.

Service chain is a series of VMs hosting VNFs in a designated order with a flow goes through them sequentially to provide desired network functionality. Tabular VM migration (TVM) proposed by [32] aims at reducing the number of hops in service chain of network functions in Cloud data centers. They use VM migration to reduce the number of hops (network elements) the flow should traverse to satisfy Service level agreements (SLAs). SLA-driven Ordered Variable-width Windowing (SOVWin) is a heuristic proposed by Pai et al. [33] to address the same problem, however, using initial static placement. Similarly, an orchestrator for the automated placement of VNFs across the resources proposed by Clayman et al. [34].

Table 4.4 - Virtual Network Functions Management Projects

Project	Objectives	Approach/Technique
VNF-P	Handling burst in network services demand while minimizing the number of servers	Resource Allocation - Integer linear programming (ILP)
Cloud4NFV	Providing Network Function as a Service	Service provisioning – Framework Design
vConductor	Virtual network services provisioning and management	Service provisioning – Framework Design
MORSA	Multi Objective placement of virtual services	Placement - Multi-objective Genetic Algorithm
TVM	Reducing number of hops in service chain	VNF Migration - Heuristic
SOVWin	Increasing user requests acceptance rate and minimization of SLA violation	VNF Placement - Heuristic
Clayman et al.	Providing automatic placement of the virtual nodes	VNF Placement - Heuristic
T-NOVA	Building a Marketplace for VNF	Marketplace – Framework Design
UNIFY	Automated, dynamic service creation and service function chaining	Service provisioning– Framework Design

The EU-funded T-NOVA project [35] aims to realize the NFaaS concept. It designs and implements integrated management and orchestrator platform for the automated provisioning, management, monitoring and optimization of VNFs.
UNIFY [36] is another EU-funded FP7 project aims at supporting automated, dynamic service creation based on a fine-granular SFC model, SDN, and Cloud virtualization techniques. For more details on SFC, interested readers are referred to the literature survey by Medhat et al. [37]. Table 4.4 summarizes the state of the art projects on VNF management.

4.5 Network Slicing Management in Edge and Fog

Fog computing is a new trend in Cloud computing that intends to address the quality of service requirements of applications requiring real-time and low latency processing. While Fog acts as a middle layer between Edge and core Clouds to serve applications

close to the data source, core Cloud data centers provide massive data storage, heavy-duty computation, or widearea connectivity for the application.

One of the key visions of Fog computing is to add compute capabilities or general purpose computing to Edge network devices such as mobile base stations, gateways and routers. On the other hand, SDN and NFV play key roles in prospective solutions to facilitate efficient management and orchestration of network services. Despite natural synergy and affinity between these technologies, there exist not many research on the integration of Fog/Edge computing and SDN/NFV as both are still in their infancy. In our view, intraction between SDN/NFV and Fog/Edge computing is crucial for emerging applications in IoT, 5G and stream analytics. However, the scope and requirements of such interaction is still an open problem. In the following, we provide an overview of the state-of-the-art within this context.

Lingen et al. [45] define a model-driven and service-centric architecture that addresses technical challenges of integrating NFV, Fog and 5G/MEC. They introduce an open architecture based on NFV MANO proposed by the European Telecommunications Standards Institute (ETSI) and aligned with the OpenFog Consortium (OFC) reference architecture[2] that offers uniform management of IoT services spanning through Cloud to the Edge. A two-layer abstraction model along with IoT-specific modules and enhanced NFV MANO architecture is proposed to integerate Cloud, network, and Fog. As a pilot study, they presented two use cases for physical security of Fog nodes and sensor telemetry through street cabinets in the city of Barcelona.

Truong et al. [43] are among the earliest who have proposed an SDN-based architecture to support Fog Computing. They have identified required components and

[2] OpenFog Consortium, https://www.openfogconsortium.org/

specified their roles in the system. They also showed how their system can provide services in the context of Vehicular Adhoc Networks (VANETs). They showed benefits of their proposed architecture using two use-cases in data streaming and lane-change assistance services. In their proposed architecture, the centeral network view by the SDN Controller is utilized to manage resources and services and optimize their migration and replication.

Bruschi et al. [44] propose network slicing scheme for supporting multi-domain Fog/Cloud services. They propose SDN-based network slicing scheme to build an overlay network for geographically distributed Internet services using non-overlapping OpenFlow rules. Their experimental results show that the number of unicast forwarding rules installed in the overlay network significantly drops compared to the fully-meshed and OpenStack cases.

Inspired by Open Network Operating System (ONOS)[3] SDN controller, Choi et al. [46] propose a Fog operating system architecture called *FogOS* for IoT services. They identified four main challenges of Fog computing as: 1) *scalability* for handling significant number of IoT devices, 2) *complex inter-networking* caused by diverse forms of connectivity, e.g., various radio access technologies, 3) *dynamics and adaptation* in topology and quality of service (QoS) requirements, and finally 4) *diversity and heterogeneity* in communications, sensors, storage, and computing powers, etc. Based on these challenges, their proposed architecture consists of four main components: 1) Service and device abstraction, 2) Resource management, 3) Application management, 4) Edge resource: registration, ID/addressing, and control interface. They also demonstrate a

[3] ONOS, https://onosproject.org/

preliminary proof-of-concept demonstration of their system for a drone-based surveillance service.

In a recent work, Diro et al. [47] propose a mixed SDN and Fog architecture which gives priority to critical network flows while takes into account fairness among other flows in the Fog-to-things communication to satisfy QoS requirements of heterogeneous IoT applications. They intend to satisfy QoS and perfromance measures such as packet delay, lost packets and maximize throughput. Results show that their proposed method is able to serve critical and urgent flows more efficiently while provides allocation of network slices to other flow classes.

4.6 Future Research Directions

In this section, we discuss open issues in software-defined Clouds and Edge computing environments along future directions.

4.6.1 Software Defined Clouds

Our survey on network slicing management and orchestration in SDC shows that community very well recognized the problem of joint provisioning of hosts and network resources. In the earlier research, a vast amount of attention has been given to solutions for the optimization of cost/energy only focusing on either host [38] or network [39], not both. However, it is essential for the management component of the system to take into account both network and host cost at the same time. Otherwise, optimization of one can exacerbate the situation for the other. To address this issue, many research proposals have also focused on the joint host and network resource management. However, most of the

proposed approaches suffer from high computational complexity, or they are not optimal. Therefore, the development of algorithms that manage joint hosts and network resource provisioning and scheduling is of great interest. In joint host and network resource management and orchestration, not only finding the minimum subset of hosts and network resources that can handle a given workload is crucial, but also SLA and users' QoS requirements (e.g., latency) must be satisfied. The problem of joint host and network resource provisioning becomes more sophisticated when SDC supports VNF and SFC.

SFC is a hot topic attainting a significant amount of attention by the community. However, little attention has been paid to VNF placement while meeting the QoS requirements of the applications. PLAN [20] intends to minimize the network communication cost while meeting network policy requirements. However, it only considers traditional middleboxes, and it does not take into account the option of VNF migration. Therefore, one of the areas requires more attention and development of novel optimization techniques is the management and orchestration of SFCs. This has to be done in a way that the placement and migration of VNFs are optimized while SLA violation and cost/energy are maximized.

Network-aware virtual machines management is a well-studied area. However, the majority of works in this context consider VM migration and VM placement to optimize network costs. The traffic engineering and dynamic flow scheduling combined with migration and placement of VMs also provide a promising direction for the minimization of network communication cost. For example, using SDN, management and orchestration module of the system can install flow entries on the switches of the shortest path with the lowest utilization to redirect VM migration traffic to an appropriate path.

The analytical modeling of SDCs has not been investigated intensely in the literature. Therefore building a model based on priority networks that can be used for analysis of the SDCs network and validation of results from experiments conducted via simulation.

Auto-scaling of VNFs is another area that requires more in-depth attention by the community. VNFs providing networking functions for the applications are subject to performance variation due to different factors such as the load of the service or overloaded underlying hosts. Therefore, development of auto-scaling mechanisms that monitor the performance of the VMs hosting VNFs and adaptively adds or remove VMs to satisfy the SLA requirements of the applications is of paramount importance for management and orchestration of network slices. In fact, efficient placement of VNFs [41] on hosts near to the service component producing data streams or users generating requests minimizes latency and reduces the overall network cost. However, placing it on a more powerful node far in the network improves processing time [40]. Existing solutions mostly focus on either scaling without placement or placement without scaling. Moreover, auto-scaling techniques of VNFs, they typically focus on auto-scaling of a single network service (e.g., firewall), while in practice auto-scaling of VNFs must be performed in accordance with SFCs. In this context, node and link capacity limits must be considered, and the solution must maximize the benefit gained from existing hardware using techniques such as dynamic pathing. Therefore, one of the promising avenues for future research on auto-scaling of VNFs is to explore the optimal dynamic resource allocation and placement.

4.6.2 Edge and Fog Computing

In both Edge and Fog computing, the integration of 5G so far has been discussed within a very narrow scope. Although 5G network resource management and resource discovery in Edge/Fog computing have been investigated, many other challenging issues in this area are still unexplored. Mobility-aware service management in 5G enabled Fog computing and forwarding large amount of data from one Fog node to another in real-time overcoming communication overhead can be very difficult to ensure. In addition, due to decentralized orchestration and heterogeneity among Fog nodes, modelling, management and provisioning of 5G network resources are not as straight-forward as other computing paradigms.

Moreover, compared to Mobile Edge servers, Cloudlets and Cloud datacenters, the number of Fog nodes and their probability of being faulty are very high. In this case, implementation of SDN (one of the foundation blocks of 5G) in Fog computing can get obstructed significantly. One the other hand, Fog computing enables traditional networking devices to process incoming data and due to 5G, this data amount can be significantly huge. In such scenario, adding more resources in traditional networking devices will be very costly, less secured and hinders their inherent functionalities like routing, packet forwarding, etc. which in consequence affect the basic commitments of 5G network and NFV.

Nonetheless, Fog infrastructures can be owned by different providers that can significantly resist developing a generalized pricing policy for 5G-enabled Fog computing. Prioritized network slicing for forwarding latency-sensitive IoT data can also contribute additional complications in 5G enabled Fog computing. Opportunistic

scheduling and reservation of virtual network resources is tough to implement in Fog as it deals with a large number of IoT devices and their data sensing frequency can change with the course of time. Balancing load on different virtual networks and their QoS can degrade significantly unless efficient monitoring is imposed. Since Fog computing is a distributed computing paradigm, centralized monitoring of network resources can intensify the problem. In this case, distributed monitoring can be an efficient solution, although it can be failed to reflect the whole network context in a body. Extensive research is required to solve this issue. Besides, in promoting fault-tolerance of 5G-enabled Fog computing, topology-aware application placement, dynamic fault detection and reactive management can play a significant role which is subjected to uneven characteristics of the Fog nodes.

4.7 Conclusion

In this paper, we intended to investigate research proposals for the management and orchestration of network slices in different platforms. We discussed emerging technologies such as Software-defined networking SDN and NFV. We explored the vision of 5G for network slicing and discussed some of the ongoing projects and studies in this area. We surveyed the state of the art approaches to network slicing in Software-defined Clouds and application of this vision to the Cloud computing context. We disscussed the state of the art literature on network slices in emerging Fog/Edge computing. Finally, we identified gaps in this context and provided future directions towards the notion of network slicing.

Acknowledgments

This work is supported through Huawei Innovation Research Program (HIRP). We also thank Zhouwei for his comments and support for the work.

4.8 References

[1] J. G. Andrews *et al*., "What Will 5G Be?," in *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1065-1082, June 2014.

[2] R. Buyya, R. N. Calheiros, J. Son, A. V. Dastjerdi and Y. Yoon, "Software-Defined Cloud Computing: Architectural elements and open challenges," *2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, New Delhi, 2014, pp. 1-12.

[3] D. Ott, N. Himayat,and S. Talwar, 5G: Transforming the User Wireless Experience. Towards 5G: Applications, Requirements and Candidate Technologies, pp.34-51, 2017.

[4] J. Zhang, X. Ge, Q. Li, M. Guizani, and Y. Zhang. "5G millimeter-wave antenna array: Design and challenges." IEEE Wireless Communications 24, no. 2 (2017): 106-112.

[5] S. Chen and J. Zhao, "The requirements, challenges, and technologies for 5G of terrestrial mobile telecommunication," IEEE Commun. Mag., vol. 52, no. 5, pp. 36–43, May 2014.

[6] T. D. P.Perera, D. N. K. Jayakody, S. De, and Maksim Anatoljevich Ivanov. "A Survey on Simultaneous Wireless Information and Power Transfer." In Journal of Physics: Conference Series, vol. 803, no. 1, p. 012113. IOP Publishing, 2017.

[7] P., Pekka. "A brief overview of 5G research activities." In 5G for Ubiquitous Connectivity (5GU), 2014 1st International Conference on, pp. 17-22. IEEE, 2014.

[8] A. Nakao, P. Du, Y. Kiriha, F. Granelli, A. A. Gebremariam, T. Taleb, and M. Bagaa. "End-to-end network slicing for 5g mobile networks." Journal of Information Processing 25 (2017): 153-163.

[9] K. Samdanis, S. Wright, A. Banchs, A. Capone, M. Ulema, and K. Obana. "5G Network Slicing: Part 1–Concepts, Principales, and Architectures [Guest Editorial]." IEEE Communications Magazine 55, no. 5 (2017): 70-71.

[10] S. Sharma, R. Miller, and A. Francini. "A Cloud-Native Approach to 5G Network Slicing." IEEE Communications Magazine 55, no. 8 (2017): 120-127.

[11] W. Shi, J. Cao, Q. Zhang, Y. Li and L. Xu, "Edge Computing: Vision and Challenges," in *IEEE Internet of Things Journal*, vol. 3, no. 5, pp. 637-646, Oct. 2016.

[12] X. Foukas, G. Patounas, A. Elmokashfi, and M. K. Marina. "Network Slicing in 5G: Survey and Challenges." IEEE Communications Magazine 55, no. 5 (2017): 94-100.

[13] X. Li, M. Samaka, H. A. Chan, D. Bhamare, L. Gupta, C. Guo, and R. Jain. "Network slicing for 5g: Challenges and opportunities." IEEE Internet Computing 21, no. 5 (2017): 20-27.

[14] F. Bonomi, R. Milito, J. Zhu, and S. Addepalli. 2012. "Fog computing and its role in the internet of things". In Proceedings of the first edition of the MCC workshop on Mobile Cloud computing (MCC '12). ACM, New York, NY, USA, 13-16.

[15] R. Cziva, S. Jouët, D. Stapleton, F. P. Tso and D. P. Pezaros, "SDN-Based Virtual Machine Management for Cloud Data Centers," in *IEEE Transactions on Network and Service Management*, vol. 13, no. 2, pp. 212-225, June 2016.

[16] S. H. Wang, P. P. W. Huang, C. H. P. Wen and L. C. Wang, "EQVMP: Energy-efficient and QoS-aware virtual machine placement for software defined datacenter networks," *The International Conference on Information Networking 2014 (ICOIN2014)*, Phuket, 2014, pp. 220-225.

[17] V. Mann, A. Gupta, P. Dutta, A. Vishnoi, P. Bhattacharya, R. Poddar, and A. Iyer, 2012. "Remedy: Network-aware steady state VM management for data centers." *NETWORKING 2012*, pp.190-204.

[18] J. W. Jiang, T. Lan, S. Ha, M. Chen and M. Chiang, "Joint VM placement and routing for data center traffic engineering," *2012 Proceedings IEEE INFOCOM*, Orlando, FL, 2012, pp. 2876-2880.

[19] W. Fang, X. Liang, S. Li, L. Chiaraviglio, N. Xiong, "VMPlanner: Optimizing virtual machine placement and traffic flow routing to reduce network power costs in Cloud data centers", *In Computer* Networks, Volume 57, Issue 1, 2013, pp. 179-196.

[20] L. Cui, F. P. Tso, D. P. Pezaros, W. Jia and W. Zhao, "PLAN: Joint Policy- and Network-Aware VM Management for Cloud Data Centers," in *IEEE Transactions on Parallel and Distributed Systems*, vol. 28, no. 4, pp. 1163-1175, April 1 2017.

[21] H. Jin, T. Cheocherngngarn, D. Levy, A. Smith, D. Pan, J. Liu and N. Pissinou., "Joint Host-Network Optimization for Energy-Efficient Data Center Networking," *2013 IEEE 27th International Symposium on Parallel and Distributed Processing*, Boston, MA, 2013, pp. 623-634.

[22] W. Voorsluys, J. Broberg, S. Venugopal, and R. Buyya, 2009. "Cost of Virtual Machine Live Migration in Clouds: A Performance Evaluation". CloudCom, 9, pp.254-265.

[23] M. F. Bari, M. F. Zhani, Q. Zhang, R. Ahmed and R. Boutaba, "CQNCR: Optimal VM migration planning in Cloud data centers," *2014 IFIP Networking Conference*, Trondheim, 2014, pp. 1-9.

[24] S. Ghorbani, S. and M. Caesar, "Walk the line: consistent network updates with bandwidth guarantees". *In Proceedings of the first workshop on Hot topics in software defined networks*, 2012 (pp. 67-72). ACM.

[25] X. Li, Q. He, J. Chen, K. Ye and T. Yin, "Informed live migration strategies of virtual machines for cluster load balancing". Network and Parallel Computing, 2011 pp.111-122.

[26] F. Xu, F. Liu, L. Liu, H. Jin, B. Li and B. Li, "iAware: Making Live Migration of Virtual Machines Interference-Aware in the Cloud," in *IEEE Transactions on Computers*, vol. 63, no. 12, pp. 3012-3025, Dec. 2014.

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

[28] H. Moens and F. D. Turck, "VNF-P: A model for efficient placement of virtualized network functions," *10th International Conference on Network and Service Management (CNSM) and Workshop*, Rio de Janeiro, 2014, pp. 418-423

[29] J. Soares, M. Dias, J. Carapinha, B. Parreira and S. Sargento, "Cloud4NFV: A platform for Virtual Network Functions," 2014 IEEE 3rd International Conference on Cloud Networking (CloudNet), Luxembourg, 2014, pp. 288-293.

[30] W. Shen, M. Yoshida, T. Kawabata, K. Minato and W. Imajuku, "vConductor: An NFV management solution for realizing end-to-end virtual network services," *The 16th Asia-Pacific Network Operations and Management Symposium*, Hsinchu, 2014, pp. 1-6.

[31] M. Yoshida, W. Shen, T. Kawabata, K. Minato and W. Imajuku, "MORSA: A multi-objective resource scheduling algorithm for NFV infrastructure," *The 16th Asia-Pacific Network Operations and Management Symposium*, Hsinchu, 2014, pp. 1-6.

[32] Y. F. Wu, Y. L. Su and C. H. P. Wen, "TVM: Tabular VM migration for reducing hop violations of service chains in Cloud datacenters," *2017 IEEE International Conference on Communications (ICC)*, Paris, 2017, pp. 1-6.

[33] Yuan-Ming Pai, C. H. P. Wen and Li-Ping Tung, "SLA-driven Ordered Variable-width Windowing for service-chain deployment in SDN datacenters," *2017 International Conference on Information Networking (ICOIN)*, Da Nang, 2017, pp. 167-172.

[34] S. Clayman, E. Maini, A. Galis, A. Manzalini and N. Mazzocca, "The dynamic placement of virtual network functions," *2014 IEEE Network Operations and Management Symposium (NOMS)*, Krakow, 2014, pp. 1-9.

[35] G. Xilouris *et al.*, "T-NOVA: A marketplace for virtualized network functions," *2014 European Conference on Networks and Communications (EuCNC)*, Bologna, 2014, pp. 1-5.

[36] B. Sonkoly, R. Szabo, D. Jocha, J. Czentye, M. Kind and F. J. Westphal, "UNIFYing Cloud and Carrier Network Resources: An Architectural View," *2015 IEEE Global Communications Conference (GLOBECOM)*, San Diego, CA, 2015, pp. 1-7.

[37] 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," in *IEEE Communications Magazine*, vol. 55, no. 2, pp. 216-223, February 2017.

[38] A. Beloglazov, J. Abawajy, R. Buyya, "Energy-aware resource allocation heuristics for efficient management of data centers for Cloud computing", In Future Generation Computer Systems, Volume 28, Issue 5, pp. 755-768, 2012.

[39] B. Heller, S. Seetharaman,, P. Mahadevan, Y. Yiakoumis, P. Sharma, S. Banerjee, and N. McKeown, "ElasticTree: Saving Energy in Data Center Networks". In *Nsdi*, Vol. 10, pp. 249-264, April 2010.

[40] S. Dräxler, H. Karl, and Z. A. Mann. "Joint Optimization of Scaling and Placement of Virtual Network Services". In *Proceedings of the 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing* (CCGrid '17). IEEE Press, Piscataway, NJ, USA, 365-370, 2017.

[41] A. Fischer, J. F. Botero, M. T. Beck, H. de Meer and X. Hesselbach, "Virtual Network Embedding: A Survey," in *IEEE Communications Surveys & Tutorials*, vol. 15, no. 4, pp. 1888-1906, Fourth Quarter, 2013.

[42] Huawei Technologies' white paper, "5G Network Architecture A High-Level Perspective", 2016 Available online: *http://www.huawei.com/minisite/hwmbbf16/insights/5G-Nework-Architecture-Whitepaper-en.pdf*

[43] N. B. Truong, G. M. Lee and Y. Ghamri-Doudane, "Software defined networking-based vehicular Adhoc Network with Fog Computing," *2015 IFIP/IEEE International Symposium on Integrated Network Management (IM)*, Ottawa, ON, 2015, pp. 1202-1207.

[44] R. Bruschi, F. Davoli, P. Lago and J. F. Pajo, "A scalable SDN slicing scheme for multi-domain fog/Cloud services," *2017 IEEE Conference on Network Softwarization (NetSoft)*, Bologna, 2017, pp. 1-6.

[45] F. van Lingen and M. Yannuzzi and A. Jain and R. Irons-Mclean and O. Lluch and D. Carrera and J. L. Perez and A. Gutierrez and D. Montero and J. Marti and R. Maso and a. J. P. Rodriguez "The Unavoidable Convergence of NFV, 5G, and Fog: A Model-Driven Approach to Bridge Cloud and Edge," *in IEEE Communications Magazine*, vol. 55, no. 8, pp. 28-35, 2017.

[46] N. Choi, D. Kim, S. J. Lee and Y. Yi, "A Fog Operating System for User-Oriented IoT Services: Challenges and Research Directions," in *IEEE Communications Magazine*, vol. 55, no. 8, pp. 44-51, 2017.

[47] A.A. Diro, H.T. Reda, and N. Chilamkurti, "Differential flow space allocation scheme in SDN based fog computing for IoT applications" *Journal of Ambient Intelligence and Humanized Computing*, pp.1-11, 2018.

[48] M. Afrin, M.A. Razzaque, I. Anjum, M.M. Hassan, and A. Alamri, 2017. Tradeoff between User Quality-Of-Experience and Service Provider Profit in 5G Cloud Radio Access Network. Sustainability, 9(11), p.2127.

[49] Nunna, S., Kousaridas, A., Ibrahim, M., Dillinger, M., Thuemmler, C., Feussner, H. and Schneider, A., 2015, April. Enabling real-time context-aware collaboration through 5G and mobile edge computing. In 12th International Conference on Information Technology-New Generations (ITNG), 2015 (pp. 601-605). IEEE.

[50] Ketykó, I., Kecskés, L., Nemes, C. and Farkas, L., 2016, June. Multi-user computation offloading as multiple knapsack problem for 5G mobile edge computing. In European Conference on Networks and Communications (EuCNC), 2016 (pp. 225-229). IEEE.

[51] Zhang, K., Mao, Y., Leng, S., Zhao, Q., Li, L., Peng, X., Pan, L., Maharjan, S. and Zhang, Y., 2016. Energy-efficient offloading for mobile edge computing in 5G heterogeneous networks. IEEE Access, 4, pp.5896-5907.

[52] Ge, C., Wang, N., Skillman, S., Foster, G. and Cao, Y., 2016, September. QoE-driven DASH video caching and adaptation at 5G mobile edge. In Proceedings of the 3rd ACM Conference on Information-Centric Networking (pp. 237-242). ACM.

[53] Mahmud, M., Afrin, M., Razzaque, M., Hassan, M.M., Alelaiwi, A. and Alrubaian, M., 2016. Maximizing quality of experience through context- aware mobile application scheduling in Cloudlet infrastructure. Software: Practice and Experience, 46(11), pp.1525-1545.

[54] Dastjerdi, A.V. and Buyya, R., 2016. Fog computing: Helping the Internet of Things realize its potential. Computer, 49(8), pp.112-116.

[55] Mahmud, R., Luiz Koch, F. and Buyya, R. 2018. Cloud-Fog Interoperability in IoT-enabled Healthcare Solutions. In Proceedings of the 19th International Conference on Distributed Computing and Networking (ICDCN '18). ACM, New York, NY, USA, Article 32.

[56] Mahmud, R., Kotagiri, R. and Buyya, R., 2018. Fog computing: A taxonomy, survey and future directions. In Internet of Everything (pp. 103-130). Springer, Singapore.

[57] Amendola, D., Cordeschi, N. and Baccarelli, E., 2016, October. Bandwidth management VMs live migration in wireless fog computing for 5G networks. In 5th IEEE International Conference on Cloud Networking (Cloudnet), 2016 (pp. 21-26). IEEE.

[58] Afrin, M., & Mahmud, M.R. (2017). Software Defined Network-based Scalable Resource Discovery for Internet of Things. EAI Endorsed Transaction on Scalable Information Systems, 4, e4.

[59] Peng, M., Yan, S., Zhang, K., & Wang, C. (2016). Fog-computing-based radio access networks: issues and challenges. IEEE Network, 30(4), 46-53.

[60] Afrin, M., Mahmud, M.R. and Razzaque, M.A., 2015, December. Real time detection of speed breakers and warning system for on-road drivers. In Electrical and Computer Engineering (WIECON-ECE), 2015 IEEE International WIE Conference on (pp. 495-498). IEEE.