College of Engineering, Design and Physical Sciences
Electronic and Computer Engineering

Mobile Edge Cloud: Intelligent deployment and services for 5G Indoor Network

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy (PhD)

Nawar Jawad
Student Number: 0835698

Supervisor:
Professor John Cosmas

Year of Submission: 2020

Abstract

Fifth-Generation (5G) mobile networks are expected to perform according to the stringent performance targets assigned by standardization committees. Therefore, significant changes are proposed to the network infrastructure to achieve the expected performance levels. Network Function Virtualization, cloud computing and Software Defined Networks are some of the main technologies being utilised to ensure flexible network design, with optimum performance and efficient resource utilization. The aforementioned technologies are shifting the network architecture into service-based rather device-based architecture.

In this regard, this thesis provides experimental investigation, design, implementation and evaluation of various multimedia services along with integration design and caching solution for 5G indoor network.

The multimedia services are targeting the enhancement of UEs' Quality of Experience, by exploiting the intelligence offered by the synergy between SDN and NFV technologies, to design and develop new multimedia solutions with improved QoE.

The caching solution is designed to achieve a good trade-off between latency reduction and resource utilization that satisfies efficient network performance and resource utilization.

The proposed network integration design targets deploying IoRL gNB with its innovative intelligent services. It have successfully achieved lower overhead signalling compared to the traditional network architectures.

Whilst all of the proposed solutions have proven to provide enhancement to the system performance, the testing results for the multimedia services showed high QoS performance parameters in the form of zero packet loss due to route switching, very high throughput and 0.03 ms jitter.

The caching solution test results provided up to 300% server utilization improvement (based on the deployed scenario) with negligible extra delay cost (0.5ms).

As for the proposed integration design, the quantification of the performance enhancement is represented by the amount of the reduced overhead signalling. In the case of Intra-secondary gNB handover within the same Main eNB, the back-haul signalling for the AMF was reduced 100% while the overall overhead signalling is reduced by 50% compared to traditional deployment architecture.

Acknowledgments

First, I wish to express my deepest appreciation and sincere gratitude to my supervisor, Professor John Cosmas, for his valuable guidance, encouragement and great support during my PhD research. It has been an honour to work under his supervision.

Also, it has been a privilege to work with my second supervisor Professor Maozhen Li, who I sincerely grateful to him.

I would also like to express my sincere thanks to the European Union's Horizon 2020 research program for its financial support on the Internet of Radio-Light (IoRL) project H2020-ICT 761992 that made my PhD work possible.

I would like to thank all my friends and colleagues at Brunel University who supported me during my PhD research.

Finally, I would like to acknowledge the financial, academic and technical assistance given by Brunel University with special gratitude to the staff in the Student Centre, the Post-Graduate Research Office, the Library, the Accommodation Office and the Residences Office.

Dedication

I dedicate this work to

Mum and Dad for their sacrifices and their support to help me get the best education possible;

Beloved wife and children for their love, support, encouragement and for giving me the motivation to work hard during my research period and

Brothers and friends for their support and encouragement;

List of Contents

List of figures

List of abbreviation

3GPP	**3rd Generation Partnership Project**
5G	5th Generation Mobile Network
5GC	Fifth Generation Core
CAPEX	Capital Expenditure
CDN	Content Delivery Networks
CLI	Command Line Interface
COTS	Commercial off -the- Shelf
CP	Control Plane
CRAN	Cloud Radio Access Network
DC	Dual Connectivity
DP	Data Plane
eNB	Evolved Node B (3GPP
EPC	Evolved Packet Core
gNB	5th Generation Node B
HetNet	Heterogeneous Networks

HOT	Heat Orchestration Templates
IaaS	Infrastructure as a Service
IHIPGW	Intelligent Home IP Gateway
IoRL	Internet of Radio Light
IoT	Internet of Things
ISP	Internet Service Provider
LTE	Long Term Evolution
MCC	Mobile Cloud Computing
MeNB	Master eNB
MNO	Mobile Network Operator
NFV	Network Function Virtualisation
ng-eNB	Next Generation eNB
OF	Open Flow Protocol
OFS	Open Flow Switch
OPEX	Operational Expenditure
OvS	Open virtual Switch
PaaS	Platform as a Service
PDCP	Packet Data Convergence Protocol
PDU	Packet Data Unit
QoE	Quality of Experience
QoS	Quality of Service
RAN	Radio Access Network
RAT	Radio Access Technology
RRC	Radio Resource Control
RRH	Remote Radio Head
RRU	Remote Radio Unit
SaaS	Software as a Service
SDAP	Service Data Adaptation Protocol
SDN	Software Defined Networks
SDU	Service Data Unit
SeNB	Secondary eNB
SINR	Signal to Interference + Noise Ratio

UE	User Equipment
UP	User Plane
VGW	Virtual Gateway
VIM	Virtual Infrastructure Manager
VLC	Visible Light Communication
VM	Virtual Machine
VNF	Virtualised Network Function

1 Introduction

1.1 Introduction

This chapter briefly describes motivation behind the research, followed by an explanation of some challenges of the next generation mobile networks along with description of the open challenges facing the development of next generation services and the deployment of one of the next generation indoor coverage with mobile RAN. It briefly describes the main contributions of the research and shows the used research methodology. It further presents the publications based on the researched items. Finally, it summarizes the other chapters of this thesis.

1.2 Motivation

Mobile networks are witnessing an unprecedented increase in the amount of traffic traversing through their network, influenced by the increasing number of connected devices e.g. smartphones and tablets, as well as the proliferation in the data-hungry applications and high-quality videos. On the other hand, users' expectation is inclining to the point that they expect to be connected anytime everywhere.

Meanwhile, operators are trying to enhance the network performance by utilizing the existing network infrastructure, bearing in mind the inherited inflexibility due to proprietary vendor-specific interfaces and inefficient routing mechanisms due to the centralized gateways, which has caused long latency, data forwarding inefficiency, and user plane congestion.

Due to the aforementioned reasons, mobile network designers are considering new network design as well as introducing new technologies and concepts in an attempt to enhance the user experience. Networks require a significant re-design of the network's data and control plane infrastructures. Now it is widely accepted that future cellular networks will require a greater degree of service awareness and optimum use of network resources.

The motivation is to enhance the network performance with cost-effective and resource-aware solutions, by providing programmable systems, which enables operators to deploy a new service or tune an existing one in a simple, flexible manner. Alternatively, adopting simpler approach by introducing autonomous

systems that are able to respond to load fluctuations and perform load balancing and network optimization in real time without human intervention.

In the meantime, researchers have found that 70-90 percent of the traffic are generated due to mobile users within indoor environments[1][2]. Therefore, it is more urgent for mobile networks to improve the network performance for indoors. From this perspective Internet of Radio Light (IoRL) has emerged as an innovative solution for improving network performance within indoor environments by adopting new access technologies e.g. VLC and mmWave, while being driven by Intelligent Home IP Gateway (IHIPGW). The IHIPGW is designed to be flexible, cost-efficient and resource-aware solution by exploiting Software Defined Networking (SDN) and NFV technologies.

SDN continues to be the dominant topic of interest in networking by providing a centralized view of the network yet flexibly programmable, which makes the deployment of new applications and services easier. These new features make the networks more adaptable and easily configurable in order to cope with the changing demands and operating environments. NFV technology and Infrastructure as a Service (IaaS) enables transforming the underlying multi-vendor resources into pools of homogeneous resources available for utilization, which facilitates the rapid deployment of services and the efficient resource utilization. In summary, SDN and NFV technologies enables the design and deployment of networks that are highly adaptive, flexible, and scalable.

1.3 Challenges

The ever-increasing user expectations and data consumption are presenting a constant pressure on mobile networks to evolve their performance and develop new services to improve the users' Quality of Experience (QoE) to meet those demands and expectations.

It is worth mentioning that the term QoE that used throughout the thesis refers to enhancing the end-users' experience by offering new services to end users in simplified way so that enables them to enjoy new features along with traditional network services, thereby achieving higher user satisfaction.

This section summarizes the challenges of the current networks that make it difficult to meet the target performance expectations.

The current network architecture is heavily dependent on proprietary devices and solutions, which in turn utilize proprietary interfaces to communicate and to be configured. This networking paradigm lacks openness, flexibility and limits innovations. Therefore, it is very difficult to introduce a new service and very

expensive to upgrade the networks. The available services are centrally located within the core network, leading to inefficient traffic routing. Current networks have defects of long latency, data forwarding inefficiency, and user plane congestion. In the same context, there are various challenges that encounters mobile networks in different categories, which include but not limited to:

- ✓ Challenges that face developing multimedia services that merge multicasting and TV control services
    - The need for fine-tuning between TV and smartphones is always required.
    - The lack of common approach between the various remote mirroring developments, and the existence of variety solutions are numerous, and, they do not share a common approach, which makes it difficult to evolve towards a holistic approach that can serve a wide range of different mirroring and sharing services.
- ✓ Challenges that face developing multimedia multicasting among multiple end users
    - Services and mechanisms rely on end users' device capabilities, which limits the service performance and expandability.
- ✓ Challenges that face the deployment of caching solutions with mobile networks.
    - Low resource utilization for locally deployed servers.
    - Higher latency value for highly utilized servers (core deployments).
- ✓ Challenges that face the deployment of (IoRL) small cells within network RAN.
    - Need for solution maintaining the benefits of the IoRL brought by its innovative design without worsen the overall performance due to excessive overhead signalling.
    - Signal interference issues.
    - High bandwidth requirement for front-haul links.
    - Increased signaling back to the network core due to frequent handovers.
    - Lacking locally deployed intelligent services, which poses extra load on the backhaul links for traffic traversing.

Researchers in both academia and industry consider SDN/NFV based solutions for the next generation mobile networks, considering that SDN and NFV technologies as the enablers for novel approaches that are gaining great momentum and

interests, by separating the control plane from the data plane, using open interfaces to increase system programmability and creating homogeneous pools of resources. The work presented in this thesis addresses all of the aforementioned challenges, providing solutions to overcome such challenges. All solutions and proposed services have been tested and provided a numerical quantification to justify their validity and performance enhancement.

1.4 Contributions

The work presented in this work reflects the importance of Cloud computing, SDN and NFV technologies in designing new services and features for the next generation mobile networks that are cost efficient and resource aware. The contributions of the thesis specifically are:

- ✓ Design and build an intelligent platform that encompasses IoRL's Intelligent Home IP Gateway (IHIPGW), by exploiting the SDN networking, NFV technology and infrastructure as a service (IaaS) paradigm. Openstack Virtual Infrastructure Manager (VIM) are utilized to develop the overall platform.
- ✓ Designing an innovative media casting and control solution that works with TV sets around indoor premises, the novelty of the solution is that it exploits the network intelligence and location estimation accuracy to offer multimedia services that enable the end users to enjoy multicasting features without relying on their devices' capabilities, rather on network's capabilities.
- ✓ Similarly, designing another media service that is more suitable for other indoor environments e.g. museums namely Multicast Sharing Service (MSS), which is another geolocation media service that enables its clients to share media contents based on proximity or subscription. Both services are promoting the network of services approach and maximizing resources utilization.
- ✓ Designing cloud-based caching solution to support the aforementioned services deployment in a cost-efficient and resource awareness fashion. The proposed caching solution not only achieves a reasonable balance between server utilization and achieved latency, but also strengthens the affordability and practicality of the IoRL 5G gNB.
- ✓ Developing a deployment and an integration solution for IoRL that enables such an intelligent and service-full indoor gNB to coexist within the network RAN at lower signaling cost, thereby promoting the adoption of such network coverage in the next generation mobile network.

1.5 Aims and objectives

The aims of this work are to enhance the network performance of the next generation mobile network by facilitating the deployment of a novel small cell gNBs within mobile network architecture along with the deployment of smart services for the end users. Also, to facilitate a faster time to market service delivery and offer new geolocation services for the end users in indoors environment specifically.

The objectives of the research are to propose solutions to address the current challenges that facing the deployment of indoor small-cell gNBs and services over the next-generation network.

Therefore, the objectives of this work can be categorized as:

- Proposing a network architecture to deploy IoRL gNBs within mobile networks efficiently, so that it enables end-users to enjoy the intelligent services offered by the gNB, without degrading the services by the overhead signaling due to frequent handovers.
- Adopting low-latency solutions and offloading some of the intelligence from a centralized core network to distributed nodes enables a fine-grain traffic monitoring and management, which in turn enables scenario based customized services to end-users.
- Exploiting the available resources within the IoRL gNB e.g. the radio access technologies i.e. mmWave, VLC and the localization accuracy to enable geolocation multimedia services.
- Providing network architecture implementation walkthrough, by exploiting open-source Virtual Infrastructure Manager that is OpenStack. The use of OpenStack enabled cloud deployment in an industrial-like platform. Therefore, enabled the deployment of multiple network services in the form of virtualized network functions.
- Proposing a caching deployment that is dedicated for IoRL small-cell gNBs to enable the efficient deployment of multimedia services as well as promoting virtual resources slicing.

1.6 Methodology

The fast-growing demand of data by (User Equipment)s UEs and users' expectations about the next generation mobile network are the main motivators

for service providers to find solutions to cope with this performance race. In this perspective, the work presented in this thesis supports these facts, by adopting a small cell solution for indoor environment to enhance the network performance within indoors, where most of the traffic (70-90) percentage of the traffic is generated indoors [2][1]. Therefore, new multimedia services were proposed to be deployed within an innovative platform of the IoRL small cell; new network caching solution that enhances the resource optimization for supporting such services and finally a network architecture design for IoRL deployment without adding new network entity nor worsen the signaling load.

The proposed solutions and services are designed to work with IoRL-based gNBs specifically, therefore the ultimate target is to test the solutions on an actual IoRL system. During the development stages, the IoRL testbed was not fully ready for the final tests e.g. the wireless access links were in constant development process, therefore, I stepped forward and performed emulation and simulation testing to provide proof of concept for the proposed solutions. Various tools were utilized e.g. Mininet, OMNeT++ and mathematical models. Various scenarios and case studies were designed and simulated to obtain sufficient results. The obtained results were, compared, analyzed and evaluated accordingly to assess the performance of the proposed solutions.

1.7 Publications

Journals

- N. Jawad et al., "Smart Television Services Using NFV/SDN Network Management," in IEEE Transactions on Broadcasting, Special Issue on the Convergence of Broadcast and Broadband in the 5G Era, vol. 65, no. 2, pp. 404-413, June 2019. doi: 10.1109/TBC.2019.2898159.
- Nawar Jawad, Mukhald Salih and John Cosmas., " Media Casting as a Service: Industries Convergence opportunity and caching service for 5G indoor gNB," Accepted in IEEE Transactions on Broadcasting 2020
- Simone Redana, Ömer Bulakci, Christian Mannweiler, Laurent Gallo, David Navrátil, Holger Karl, Anastasius Gavras, Stephanie Parker, John Cosmas, Nawar Jawad, Mukhald Salih et al. "5GPPP Architecture Working Group View on 5G Architecture" EU CNC Valencia Spain 17th - 21st June 2019
- Yue Zhang, Hequn Zhang, Li-Ke Huang, Wei Li, Israel Koffman, John Cosmas, Nawar Jawad, Kareem Ali, Ben Meunier, Xun Zhang, Lina Xiao, Rudolf Zetik, Robert Muller, Adam Kapovits, Jintao Wang, Jian Song, Charilaos Zarakovitis "Internet of Radio and Light: 5G Building Network Radio and Edge Architecture" IEEE ITU Journal paper, 2020

Conferences

- N. Jawad et al “Indoor Unicasting/Multicasting service based on 5G Internet of Radio Light network paradigm” 2019 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB), S.Korea, 2019,
- Nawar Jawad et al., " Virtual Gateway: Local Multimedia Services and mobility management for 5G Internet of Radio Light gNB," 2020 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB), France, 2020, submitted under review.
- Co-author
- M. Salih, N. Jawad and J. Cosmas, "Software Defined Selective Traffic Offloading (SDSTO)," 2018 IEEE 23rd International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD), Barcelona, 2018, pp. 1-7. doi: 10.1109/CAMAD.2018.8514940.
- J.Cosmas, B.Meunier, K. Ali, N. Jawad et al"A 5G Radio-Light SDN Architecture for Wireless and Mobile Network Access in Buildings," 2018 IEEE 5G World Forum (5GWF), Silicon Valley, CA, 2018, pp. 135-140.doi: 10.1109/5GWF.2018.8516970

- M. Salih, N. Jawad and J. Cosmas,"Simulation and Performance Analysis of Software-Based Mobile Core Network Architecture (SBMCNA) Using OMNeT++," 2018 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB), Valencia, 2018
- J.Cosmas, B.Meunier, K. Ali, N. Jawad et al "A Scalable and License Free 5G Internet of Radio Light Architecture for Services in Homes & Businesses," 2018 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB), Valencia, 2018, pp. 1-6. doi: 10.1109/BMSB.2018.8436938
- K. Ali, A. Alkhatar, N. Jawad and J. Cosmas, "IoRL Indoor Location Based Data Access, Indoor Location Monitoring & Guiding and Interaction Applications," 2018 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB), Valencia, 2018, pp. 1-5. doi: 10.1109/BMSB.2018.8436932
- John Cosmas, Ben Meunier, Kareem Ali, Nawar Jawad et al "5G Internet of radio light services for Musée de la Carte à Jouer," 2018 Global LIFI Congress (GLC), Paris, 2018, pp. 1-6. doi: 10.23919/GLC.2018.8319095
- John Cosmas, Ben Meunier, Kareem Ali, Nawar Jawad et al "5G Internet of radio light services for supermarkets," 2017 14th China International Forum on Solid State Lighting: International Forum on Wide Bandgap Semiconductors China (SSLChina: IFWS), Beijing, 2017, pp. 69-73. doi: 10.1109/IFWS.2017.8245977
- Charilaos C Zarakovitis, N. Jawad et al., " Optimising the 3D Access Point Assignment Problem for Hybrid VLC, mmWave and WiFi Wireless Networks,"IEEE ICC'20 - NGNI Symposium,2020.

1.8 Thesis Organization

Thesis consists of seven chapters, beginning with an introductory chapter to outline the reason behind the research, the existing challenges and the methodology of the research. Each chapter starts with an introduction and ends with a summary. The chapters are independent of each other but complementary to each other. The structure of the remainder of the thesis is organized as follows:

- **Chapter 2**: conducts an extensive survey to present the relevant literature on the next generation mobile network enabling technologies, enabling concepts, network's capacity enhancement and dual connectivity.
- **Chapter 3**: provides a detailed insight about the practical implementation methodology that been used for creating flexible and intelligent platform. Openstack VIM is explained elaborately.
- **Chapter 4**: presents the concept, implementation and the components of two of the next generation services so-called: Follow Me Service (FMS) and Multicast Sharing Service (MSS).
- **Chapter 5**: presenting the IoRL-Cache, which is cloud-based solution proposed to enhance the resource utilization of the traditional local caching servers.
- **Chapter 6**: presents the Virtual Gateway (VGW), which is a virtualized entity, enables an optimized deployment for a cluster of IoRL base stations efficiently. The mobility management discussed, highlighting the role of VGW, especially in the intra-gNB handover procedure.
- **Chapter 7**: concludes the works presented in the thesis by summarizing the research finding and points out potential future work.

2 Literature Review & Research background

2.1 Introduction

Mobile networks are evolving constantly, incorporating new technologies to cope with the exponential increase in user's demands for higher data traffic. Cisco [3] predicted that mobile data traffic will grow seven-fold increase over 2017 to 2022. Mobile operators and researchers from academia and industry are developing new proposals trying to enhance the network performance to cope with the aforementioned expectations. The vast majority of the proposed architectures are considering utilizing the existing resources more efficiently i.e. exploiting SDN, Network Function Virtualization (NFV) and Mobile Cloud Computing (MCC) as the drivers for new scalable, flexible platforms that provides efficient signal processing and data forwarding. On the other hand, new access technologies along with heterogeneous networks are also proposed to enhance the performance over the air, e.g. mmWave, Visible Light Communication (VLC) … etc.

Based on the aforementioned facts and information, his chapter aims to review and survey various technologies and concepts that are collectively enables the creation of flexible services in the next generation mobile networks.

The rest of the chapter is organized into five sections namely: Enabling technologies; Enabling Concepts; Network's capacity and Quality of Experience enhancement; Dual-connectivity and finally a summary to the chapter.

2.2 5G Enabling Technologies

Traditional networks are being constructed for Internet by using various dedicated networking devices, each of which comprises of hardware resources, proprietary software and protocols, which are designed to perform specific task precisely. This paradigm has provided acceptable performance to an extent, but some of the main challenges faced using this networking paradigm is the non-efficient resource utilization and the costly and impractical network expansions. Therefore, new networking paradigm has become more appealing, based on SDN and NFV. Such networking paradigm has enabled far more scalable, flexible and resourceful platforms. This subsection presents these technologies along with their advantages they bring for the mobile networks.

2.2.1 Network Function Virtualization

Virtualization is the mechanism of abstracting the available resources and sharing it with various applications, deployed on the same platform, with a high segregation level. Applications coexist on the shared hardware, performing similarly to the ones deployed on dedicated hardware resources i.e. the resources are utilised more efficiently. Virtualization eliminates the need for extra-dedicated hardware, as it is replaced by sets of applications running in virtual environment on the same physical hardware. From this viewpoint, the ultimate benefits that can be obtained from such environment can be summarised as; power efficiency, CAPEX and OPEX reduction, efficient resource utilization … etc. Figure 2-1 depicts some of the benefits achieved by exploiting the virtualization technology. [4]

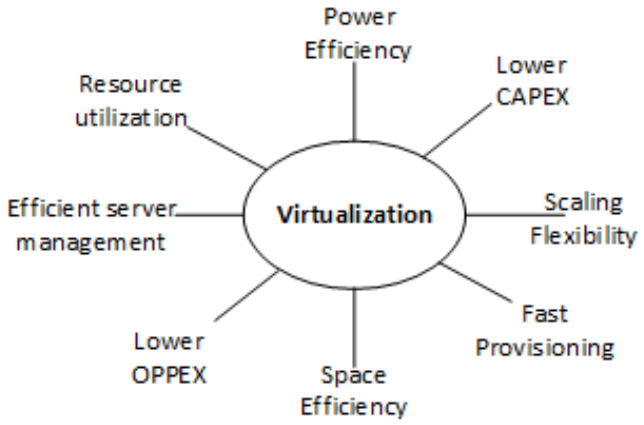

Figure 2-1 some of the benefits of resource virtualization

The NFV technology started by server virtualization, then, extended to include new forms of resource virtualization e.g. network virtualization. This new platform has directed the development of networks from device-based to service-based. In other words, hardware vendors and solution providers started to shift from the traditional manufacturing policies, where utilizing dedicated-devices to perform specific function, running on customized hardware and proprietary software, to exploiting the virtualization technology, to deploy various services, on general-purpose devices.

NFV technology exploits Commercial off -the- shelf (COTS) hardware, to create multiple copies of the abstracted resources. The main differences between COTS and the purpose-built hardware are, COTS do not have dedicated packet processing, forwarding CPUs, nor customized software or operating system that implements the network functions. Alternatively, NFV utilizes COTS to provide pool of resources, providing the “Infrastructure” layer of the NFV layered architecture. NFV also utilizes the virtualization software, to create the virtualized

resources e.g. “Virtual Compute”, “Virtual Storage”, “Virtual Networking” and so on. NFV enables the developers to exploit the obtained virtualised resources to deploy various services and functions to perform the required functionalities. Figure 2-2 depicts the high-level NFV framework.

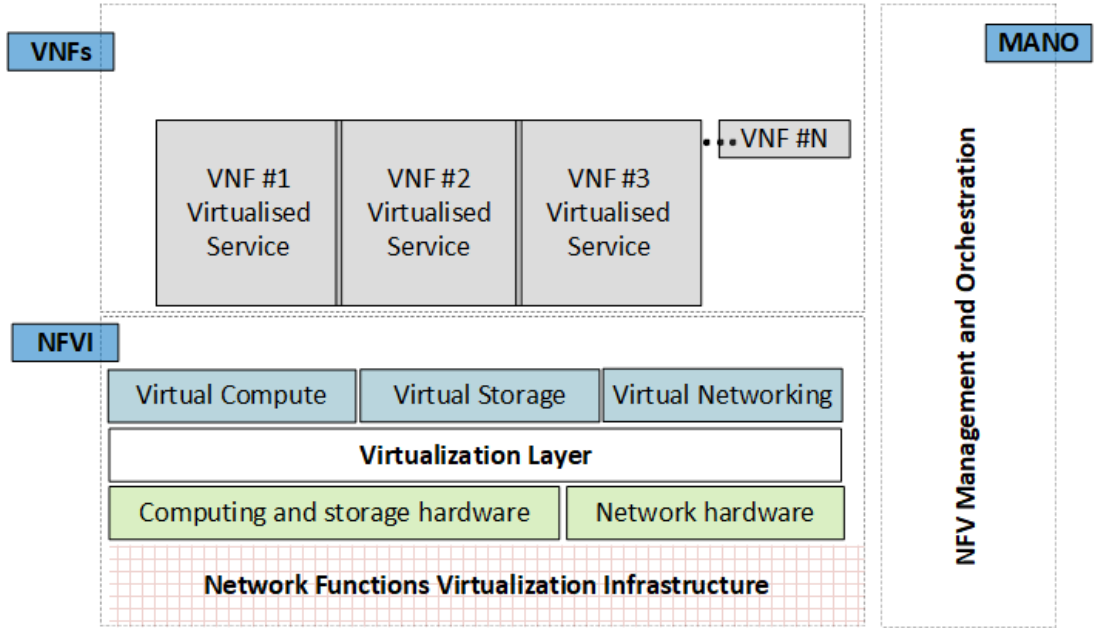

Figure 2-2 high-level NFV Framework

To achieve high performance with NFV technology inter-working with COTS, special software techniques have been developed to constitute the lack of dedicated forwarding CPUs and fast-access cache memory.

Mobile Networks reliability requires High-Availability infrastructure, so to meet the High-Availability requirement in the traditional networks, Mobile Network Operators (MNOs) design their infrastructure with redundant physical servers and overprovisioned or redundant data links, thereby in case of server failures, the redundant servers on standby replace the failed ones. This approach achieved the required performance, with high overhead cost and poor resource utilization. In contrast, NFV framework, is utilizing non-expensive COTS along with Software-based management and deployment tools, to dynamically provision, teardown, or move the virtual services. The resulting architecture overcomes the failures via re-provisioning, moving, or reconnecting the impacted services.

To perform resource abstraction and provisioning efficiently, special tools were implemented to achieve these tasks, e.g. Openstack, VMWare’s vSphere, and Kubernetes. Next chapter presents the Intelligent Home IP Gateway (IHIPGW) with the implementation steps, which involve utilizing Openstack as the Virtual Infrastructure Manager (VIM)[5], [6].

2.2.2 Software Defined Networking

SDN is a network architecture which is designed with Control Plane (CP)/Data Plane (DP) separation. CP is logically centralized, which utilizes Openflow protocol, to communicate with the DP. The deployment of CP is not bound to specific physical node, to prevent the possibility of creating single point of failure. Cluster of nodes, which are horizontally expandable, communicates using specific protocols, e.g. Border Gateway Protocol (BGP) or Path Computational Element communication Protocol (PCEP), to form a cluster of nodes that act as a centralized CP. Figure 2-3 depicts the network infrastructure transition from the traditional networking to the SDN networking [4].

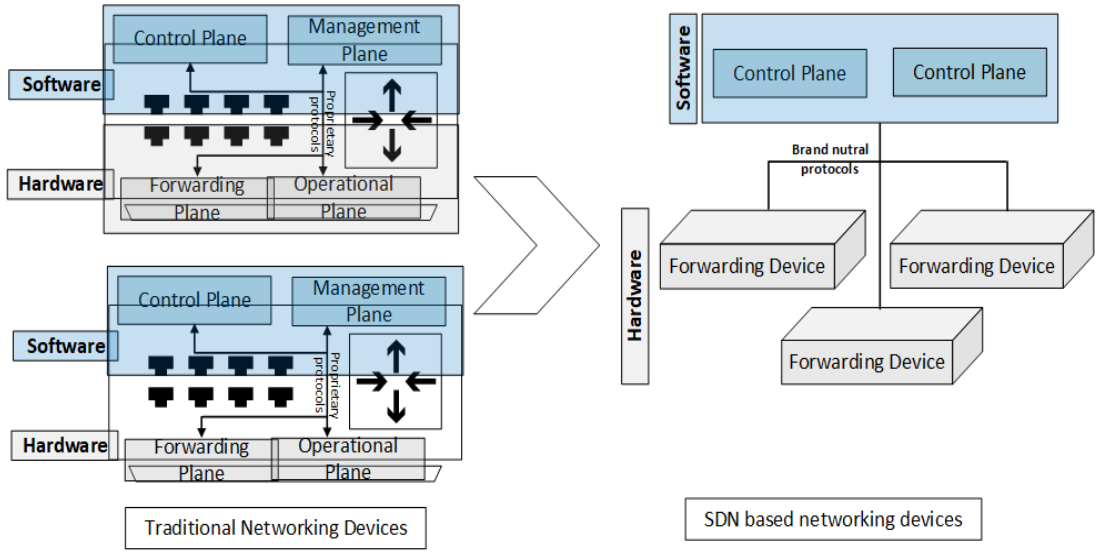

Figure 2-3 transition from traditional to SDN Architecture

The SDN architecture is depicted in Figure 2-4 it comprises of three planes namely, application plane, control & management plane and operational/Forwarding plane. Application plane encompasses a set of applications interacting with control plane and management planes, to gain a complete view of the underlying operational plane and the network topology, thereby instructing the control plane to perform the actions as received from the applications. Figure 2-4 depicts the SDN network with their respectful planes.

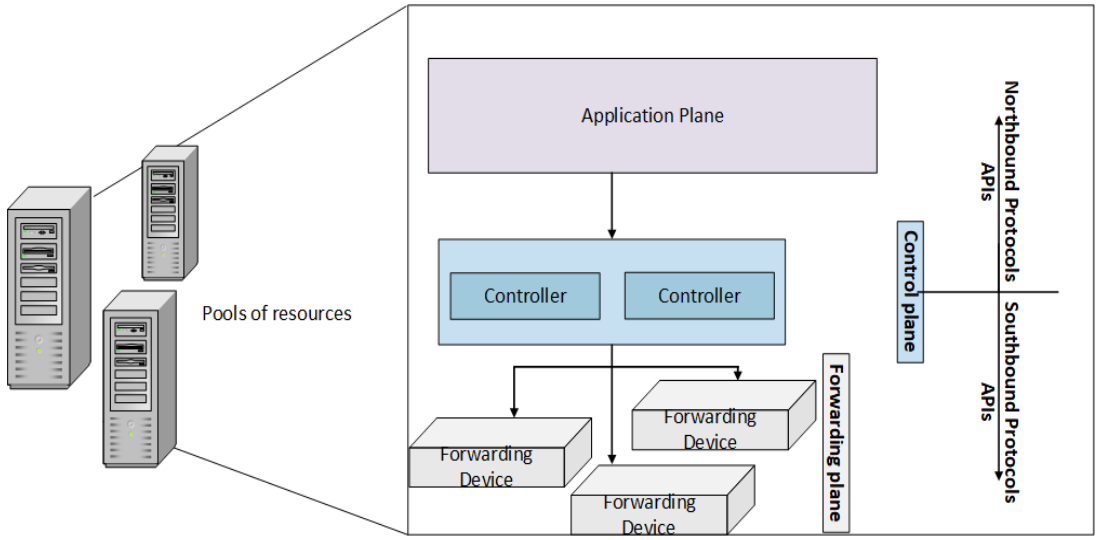

Figure 2-4 application plane and Northbound/Southbound Protocols

The northbound protocols/APIs are the protocols used between the control plane and the application plane, while the southbound protocols are referring to the protocols running between the control plane and the forwarding plane. The most common northbound API is the REST API, while Openflow is the standardised protocol for southbound communication.

2.2.3 Advantages of SDN

SDN architecture has been adopted in the networking infrastructure industry to solve the issues related to the traditional networking, especially after the massive increase in the customer demands, which revealed the inefficiency and impracticality of the solutions based on the traditional networking paradigm. SDN networking has the potential to implement flexible, scalable, open, and programmable networks. Some of the main benefits of SDN are discussed in the next sub-sections.

2.2.3.1 Programmability and Automation

Managing the traffic and controlling the network via applications is the ultimate goal of SDN networking. Eliminating or minimizing the human-machine interaction and replacing it with machine-machine interaction brings speed and overcomes human induced errors. The use of automation tools and software enables faster, scalable and agile networks deployments, which in turn is becoming a necessity to cope with current network demands. Automation and programmability are essential features that consumes the run-time information, along with predefined configuration to support the provisioning of networks. Traditional networking tools were based on proprietary software and did not have support for

configuration from external devices. SDN architecture enables the management of the network based on the received instructions from the application layer, the latter preconfigured to forward instructions to the control plane based on the current state of the network's status. Applications exist in the controller or deployed on top of the controller and interact with SDN controller using the northbound APIs. Such automation procedures enable the network to react to server failures promptly and autonomously, thereby minimizing the services downtime.

2.2.3.2 Support for Centralized Control

SDN architecture has a logically centralized the control plane, which facilitates information gathering by application developers in a simplified way, thereby developing and deploying various applications and control logics and scripts efficiently.

2.2.3.3 Multivendor and Open Architecture

SDN supports and works based on vendor-neutral platforms, by utilizing open and standardized protocols. Contrary, traditional networking paradigms utilised vendor-specific protocols and proprietary control planes, leading to issues with interoperability especially with multi-vendor platforms. However as mentioned earlier, SDN paradigm segregates the control plane from the forwarding devices, thereby enabling the control plane to manage the intelligence and instruct the forwarding devices using standardised and vendor-neutral protocol i.e. promoting interoperability in a mixed-vendor environment.

2.2.3.4 Off-Loading the Network Device

Co-locating the control plane and the forwarding plane on the physical hardware is considered as resource consuming approach, which eventually influences the device performance or requires extra resources to maintain the required performance levels. SDN approach of removing the control plane from the networking devices leads to off-loading the resources to perform the forwarding process, which simplifies the device's software and ultimately reduces the overall cost of the forwarding device [4], [7].

2.2.4 SDN/NFV in Mobile network architecture

Mobile networks are adopting the SDN and NFV technologies to provide data and control plane separation and automate network resources' allocation. There have been many works performed by researchers to present the benefits of utilising SDN

and NFV in mobile networks, e.g. authors of [5] presented Software-defined control of virtualized mobile packet core network architecture. The performed research considers utilizing OpenFlow switches that can perform sophisticated functionalities that are required in the mobile network. The core network entities are virtualized and deployed on Virtualized Network Functions (VNFs) running in a data centre, while there are SDN controllers to handle the control plane of both core network and data plane network. The proposed solution emphasised on the use of SDN and NFV technologies to offer programmability, boost network flexibility, provide on-demand dynamic provisioning, promising interoperability and reduction of both CAPEX and OPEX. Authors of [6] have proposed to change the current network architecture by exploiting SDN, NFV and Cloud computing. They referred to their proposed architecture as SoftNet and claimed that it provides Adaptability, Efficiency, Scalability, and Simplicity by utilizing Virtualized SDN-based core network and an access server to provide Unified RAN. Similarly, authors of [8] have also proposed leveraging SDN and NFV technologies within the mobile network architecture to bring automation and flexibility. Their proposed architecture consists of four planes namely; knowledge plane, network applications plane, control plane, and forwarding plane as shown in Figure 2-5. In the proposed architecture, the knowledge plane collects UEs' and network's information to be used for optimizing the network utilization and enhancing UE's QoE. The network application's plane is implemented in a centralized cloud-based infrastructure and holds the LTE control plane (in the form of virtualized functions) including virtualized SGW, virtualized PGW, virtualized MME, Routing, and RRM.
The proposed architecture facilitates implementing new applications to be deployed autonomously, which in turn reduces the cost and time-to-market of the new services. The control plane consists of Network Operating System (NOS), network virtualization block, which uses OpenFlow protocol to communicate with the forwarding plane. Forwarding plane consists of a set of OpenFlow switches.

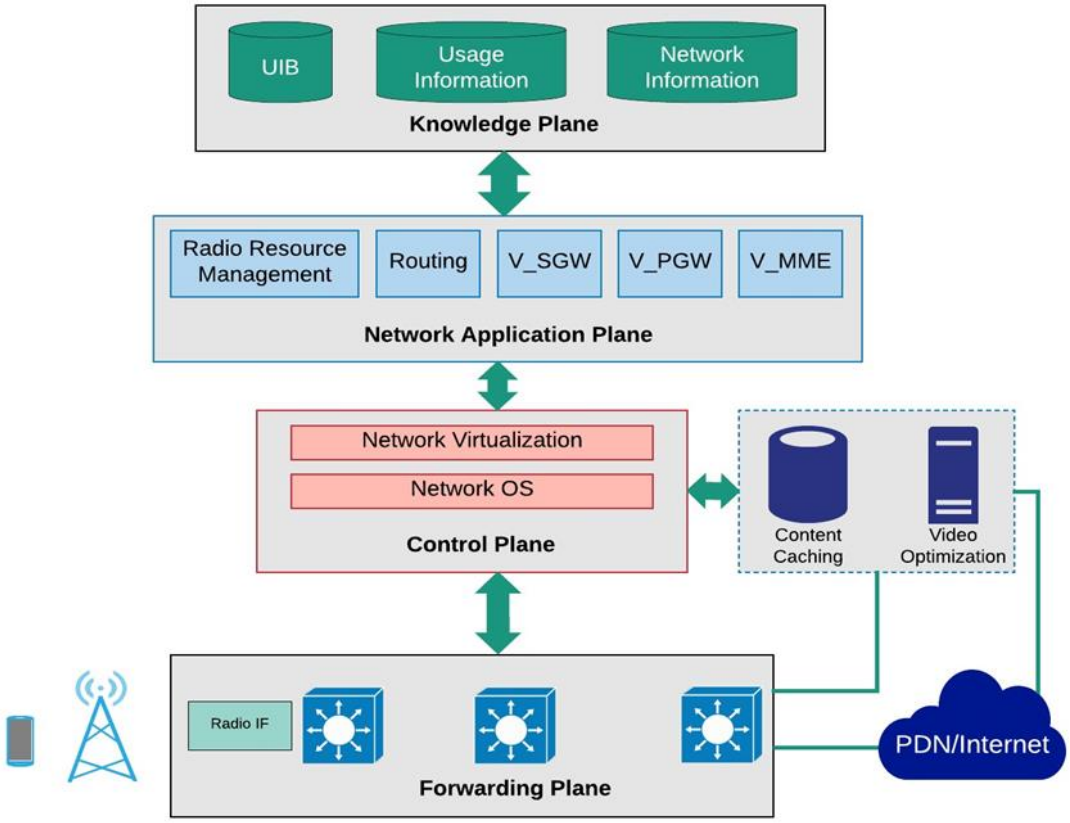

Figure 2-5 cellular software defined network proposed by [8]

2.3 5G Enabling Concepts

With the emergence of recent innovative applications e.g., augment reality, virtual reality, mixed reality … etc.; mobile networks need to cope with such delay sensitive and data intensive applications and services. Edge Cloud, Cloud Services and Mobile edge computing concepts could be exploited to provide an effective solution for meeting the low latency demands. Edge clouds enables end users to enjoy high computational power without relying on their devices' computational powers with relatively reduced latency compared with traditional remotely deployed clouds.

Cloud services concept refers to the most commonly defined cloud services as defined by the National Institute of Standards and Technology (NIST), namely, Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS) [8].

Mobile Cloud Computing concept is leveraging both of the Edge Cloud and Cloud Services concepts to enable UEs to access new services that requires extra computational power, yet without wasting devices' energy. This section reviews the aforementioned concepts, which are utilised in the proposed services within the IoRL network infrastructure.

Access network is an essential part of the system that enables successful integration between end users and cloud platforms, in IoRL the access technologies used within the access network are mmWave and VLC, these technologies offer low latency, high bandwidth and decimetre localization accuracy. To appreciate the benefits of exploiting high bandwidth technologies e.g. mmWave and VLC, let us define latency and bandwidth, as well as highlighting the relationship between the two. Latency is defined as the time spent from the source sending a packet to the destination receiving it; while bandwidth is known as the maximum throughput of a logical or physical communication path. As depicted in Figure 2-6 below [9].

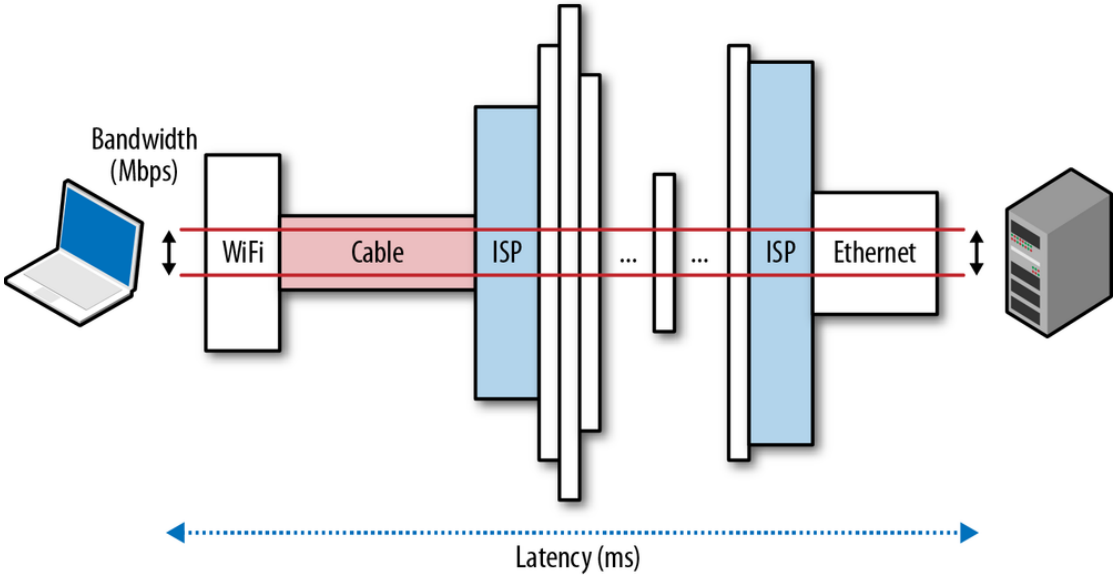

Figure 2-6 Bandwidth and latency relationship [9]

Therefore, when utilising technologies that offers high bandwidths such as mmWave and VLC, it will result in reducing the latency thereby enhancing the system performance.

2.3.1 Edge Cloud

Within the last 10 years, public clouds have experienced tremendous development in their abilities and features, which led to dramatic reshaping to the IT landscape of the enterprises. Many enterprises have first announced they were adopting a "cloud-first" strategy or that they were "all in" on the public cloud [10]. Public clouds provide services to its end uses and charge them based on resource utilization. Enterprises have utilised public clouds, private clouds and hybrid clouds. With public clouds approach, enterprises do not maintain any part of the cloud infrastructure, while the cloud provider is billing the enterprise for the use of clouds' resources, which means enterprises are only charged for the resource usage only. With public clouds, the security also represents a vulnerability for the

enterprise due to storing and processing of the data outside the premises of the enterprise. While in private clouds, an enterprise encounters high CAPEX/OPEX due to mainlining all the infrastructure on premises, which also means that private clouds are the most secure approach. Hybrid clouds represents the intelligent balance between workloads in the cloud and workloads that run on-premises, whether on traditional infrastructure or in a private cloud [10]. Table 2-1 depicts the comparison amongst the different types of cloud deployments [7].

	private	**Public**	**Hybrid**
Scalability	limited	Very high	Very high
security	Most security	Medium security	Very secure
performance	Very good	Low – Medium	Good
Reliability	Very high	Medium	Medium - high
Cost	High	Low	Medium

Table 2-1 comparison of cloud deployment

2.3.2 Cloud Services

The NIST have defined the service models of the clouds, the most common service models are being IaaS, Platform as a Service (PaaS) and SaaS. Figure 2-7 depicts the differences amongst these service models.

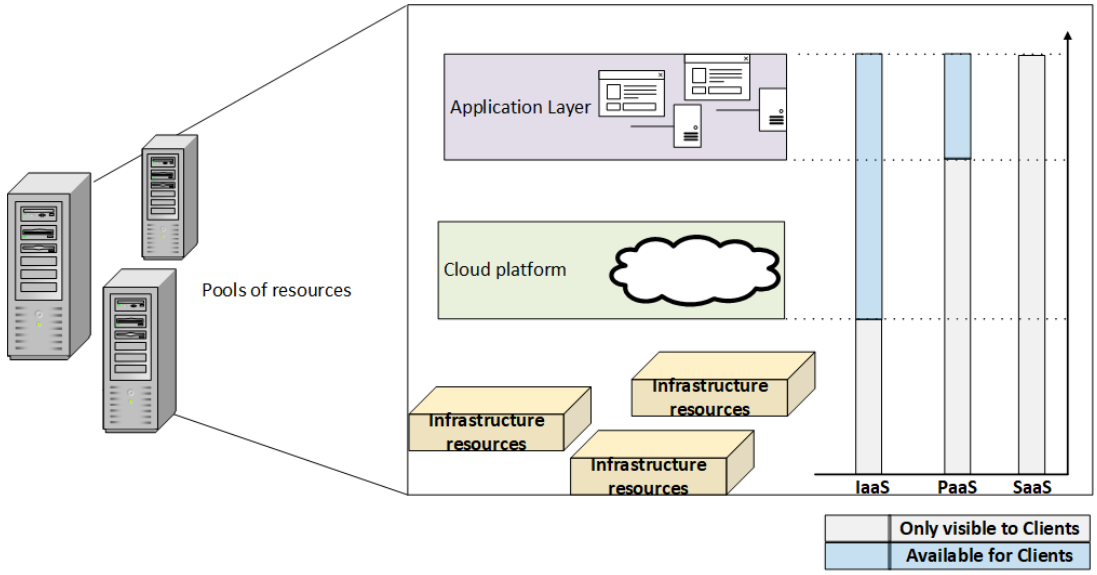

Figure 2-7 Cloud service models

2.3.2.1 Software as a Service

SaaS service model refers to enabling end users to access and exploit the final product/service in the form of available features. In other words, cloud providers have full control and visibility of their own resources and obstructed resources;

thereby deploy the required virtual machines to perform the required tasks, while the end user is able to access the resultant software or application. End users can access the cloud application using various client devices, via a simple interface such as a web browser. Enterprise (end user) is being billed for some sort of licence, which in turn makes it available for its employees to enable them to access this application. The use of SaaS eliminates the complexity of software installation, maintenance, upgrades, and patches. Examples of services at this level are Google Gmail, Microsoft 365, Salesforce and Cisco WebEx [5].

2.3.2.2 Platform as a Service

With PaaS, end users are able to exploit the cloud's resources to deploy their applications directly, without involving the resource provisioning and virtual machines deployment. A PaaS service platform provides essential software building blocks, plus a number of development tools, such as programming language tools, runtime environments, and other tools, which are utilized in deploying new applications. In other words, PaaS clients are utilizing cloud operating systems for software deployment yet being billed for the use of the computing resources only. Microsoft Azure is one of the most well-known PaaS examples [5].

2.3.2.3 Infrastructure as a Service

IaaS service model enables the end user to access the resources of the underlying cloud infrastructure. With IaaS, the hardware resources are provided in the form of virtual machines, which can be deployed with various operating systems. IaaS offers the end users processing, storage, networks, and other essential resources, which thereby enables them to deploy and run various software and applications. With IaaS, end users have the means to provision the required infrastructure, via either Graphical User Interface (GUI) or Command Line Interface (CLI). This management interface provides higher flexibility to the end user to deploy and manage the required resources as needed. Examples of IaaS are Amazon Elastic Compute Cloud (Amazon EC2), Microsoft Windows Azure and Google Compute Engine (GCE)[5].

2.3.3 Mobile Cloud Computing

Cloud Computing refers to the concept of relocating the IT-operations from local infrastructure to an Internet-connected infrastructure for the purpose of achieving higher reliability and ultimate availability in the form of ready to consume data.
The essential characteristics of the cloud computing service are as the following:

- Broad network access: enabling end user to access the cloud capabilities via heterogeneous platforms e.g. mobile phones, laptops … etc.
- Rapid elasticity: Cloud computing enables end-users to expand and reduce resources as required, concurrently.
- Measured service: Cloud systems autonomously control the resource usage to provide transparency and the required billing to both ends.
- On-demand self-service: end-user is able to provision computing capabilities, such as server time and network storage, as required, without the need of human involvement at the back end.
- Resource pooling: the available resources are being abstracted to provide a pool of resources that is being sliced amongst various end users as required.

From this viewpoint, Mobile Cloud Computing (MCC) was introduced as a promising approach. MCC is implementing the cloud-computing concept in the mobile network infrastructure, which offers to MNOs and UEs similar features of cloud computing e.g. adaptability, scalability, availability and self-awareness. From the network architecture perspectives, MCC acts as an IaaS for data storage and processing, thereby enhancing the end user's device computing and processing abilities remotely [11][10]. There are various use cases of MCC in the mobile network architecture as shown in the next subsection. Authors of [12] have employed MCC concept in their proposed architecture, they have proposed lightweight Mobile Cloud Offloading Architecture (MOCA). MOCA exploits SDN with MCC platform to perform traffic offloading. The proposed platform enhances the network performance especially with delay sensitive applications e.g. mobile gaming. The proposed MOCA architecture is depicted in Figure 2-8 .

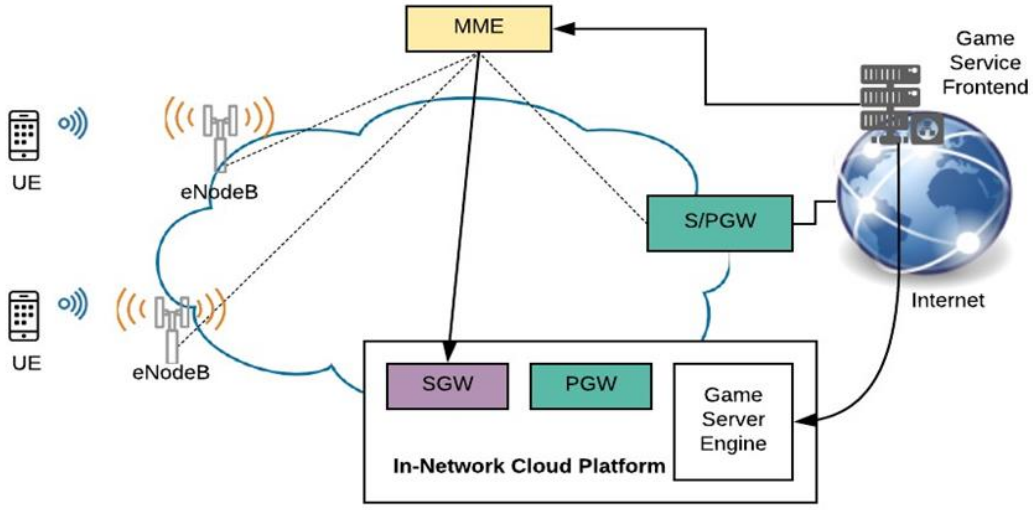

Figure 2-8 mobile Cloud Offloading Architecture (MOCA) proposed by [13]

The authors assumed that the network platform is capable of deploying new Virtual Machine (VM) as needed to perform traffic redirection tasks. Similarly, authors of

[14] have proposed cloud-based traffic offloading to enhance the traditional routing mechanism in the LTE architecture. The proposed Software-Defined Networking Mobile Offloading Architecture (SMORE) is depicted in Figure 2-9. The architecture includes SMORE controller, SMORE monitor, Database, SDN switch, and MCC platform to the LTE architecture, thereby reducing the end-to-end delay of the user's traffic.

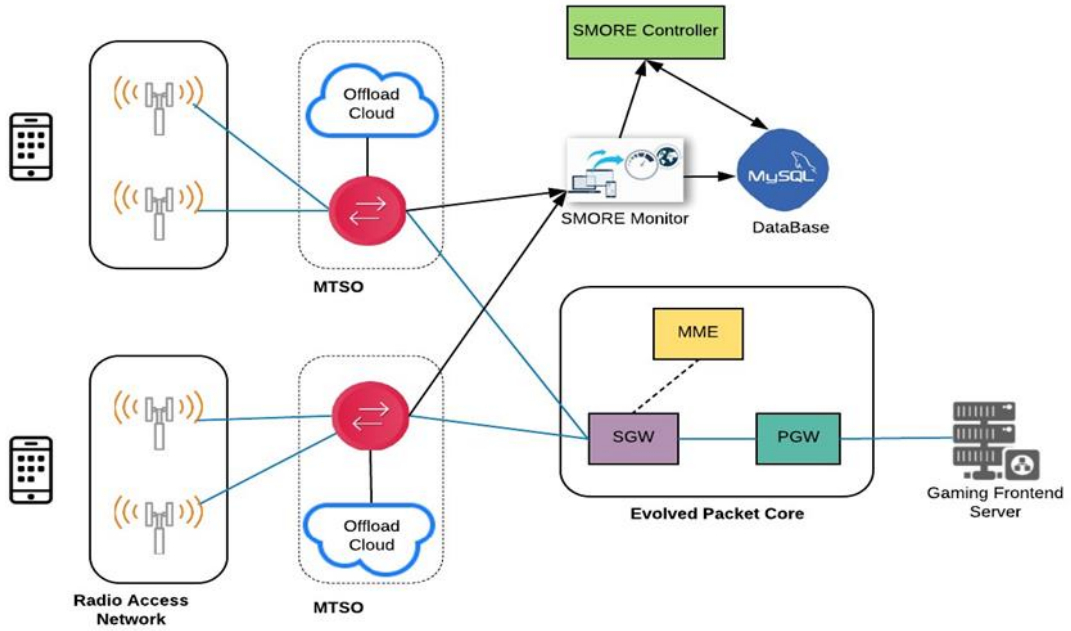

Figure 2-9 Software-Defined Networking Mobile Offloading Architecture [15]

In SMORE, the SDN switches are deployed in network edges, while the SMORE-monitor monitors the LTE/EPC control plane signalling and triggers the SMORE controller about the UEs activities e.g. UE initial attachment and handover procedures. The SMORE controller configures the SDN switches to route the traffic between the UE and the cloud. Authors presented the design of the individual entities of the proposed architecture with detailed explanations about the interception and the rerouting of the traffic to offloading servers located inside the cellular core.

2.3.4 Mobile Cloud Computing Use Cases and Applications

MCC has the potential for enhancing the network performance by enabling MNOs to offer new services to the UEs. MCC provides IT resources at the UE proximity, thereby allowing UEs to exploit these resources and achieve better Quality of Experience. In this subsection, the main MCC use cases are given along with related research works.

2.3.4.1 Computation Offloading

With computation offloading technique, UEs are being able to perform computation-intensive task by utilizing MNO's cloud resources, thereby reducing the power consumption of the UE while achieving the required performance. For instance, multimedia applications entails encoding process, which is a resource intensive task, therefore, research work [13] has proposed a specific communication protocol to enable offloading the encoding procedure to the mobile cloud. Another resource-intensive use case is the mobile gaming, where mobile gaming involves lots of rendering procedures, which requires high computational power; therefore, by offloading the rendering procedure to the cloud, UEs can enjoy the game experience regardless of their device's computational power. However, the availability of the cloud resources have to be taken into consideration, in this regard, researchers in [15] have proposed a distributed-computation offloading model, to achieve fair distribution between the users and the available cloud resources by using game theoretic approach in their analysis.

2.3.4.2 Web Performance enhancement

Web performance-enhancement use case is referring to reducing the web page latency as well as enhancing the overall browsing experience. In this context researchers of [16] have proposed an Edge Accelerated Web Browsing (EAB) solution, where in the proposed solution, an edge server is fetching Web contents prior to performing an evaluation and task rendering, thereby accelerating browsing speed. Similarly, authors of [17] have introduced Content Delivery Network-based solution for caching specific contents that constitutes large percentage of the web contents, thereby accelerating the web loading speed.

2.3.4.3 Internet of Things

MCC platform is capable of performing Internet of Things (IoT) related tasks, e.g. data aggregation and analysis for sets of IoT reporting events and measurements. In this regard, work presented in [18] describes an edge IoT architecture that comprises of hierarchy of edge cloud platforms, where within the proposed platform, data computation procedures are performed and the data streams distributed to the corresponding application server utilizing SDN networking paradigm. Researchers of [19] have provided a study about edge-cloud medical cyber-physical systems, and emphasised on the role of edge-cloud–computing in enabling low latency intelligent interaction between the computational elements and the medical devices, thereby providing a cost-efficient platform.

2.3.4.4 Smart City Services

Smart city services include multimedia services, transportation services, location-based services and so on. The aforementioned services make use of MCC to provide intelligent storage and computational power in the user proximity. Within this context, authors in [20] have provided a video analytics service executed on a MCC platform, which highlights the importance of edge clouds in eliminating the need for high bandwidth in the transport links.

2.3.4.5 Environment-aware MCC applications

As the MCC is deployed closer to the end user, it also brings the intelligence closer to the Radio Access Network. Authors in [21] presented a platform that comprises of MCC capabilities combined with Dynamic Adaptive Streaming over HTTP (DASH). The MCC provides the DASH streamer with real-time network status; thereby DASH streamer drives the suitable size of media segments to ensure optimum QoE. Other applications at the mobile network edge, were proposed as MCC-based solutions, which take advantage of UEs' proximity, low latency and radio analytics for performance optimization [22].

2.4 Network's Capacity and Users' Quality of Experience Enhancement

In 2014, the number of connected mobile devices exceeded the number of people on Earth for the first time. Connected devices have witnessed rapid increase, from zero to 7.6 billion connected devices and 3.7 billion unique subscribers in only three decades [23]. This has fundamentally transformed the way we communicate and access information. Nowadays, most of the data services reside in the Internet, far away from the user where the speed of light becomes one of the main factors limiting latency. To address this problem, processing will have to move closer to the user into a cloud-computing infrastructure as part of the network. Meanwhile in telecommunications, capacity refers to the amount of received information per UE, so increasing the capacity requires increasing the bandwidth per UE; this can be achieved either by using more frequency resources, or by densifying the available frequency resources. To understand the different aspects involved in capacity calculation, we refer back to Shannon–Hartley theorem [24]:

$$C = Blog_2\left(1 + {}^{S}/_{N}\right) \qquad 2.1$$

C is the amount of transmitted information over a communication channel. *B* is the channel bandwidth, while *S/N* refers to signal to interference noise ratio [24].

Capacity *C* could be scaled by increasing the bandwidth B per user, the signal-to-noise ratio *S/N*, or in a multi-user network the signal-to interference- plus-noise ratio (SINR). Addressing the bandwidth is a more promising approach, because it results in a linear scaling compared to the logarithmic scaling when increasing spectral efficiency by improving the SINR. Increasing the capacity by increasing the bandwidth requires increasing the carrier frequencies, due to the limitation of the bandwidth in the lower frequency carriers (mm-Wavelengths). However, increasing carrier frequency leads to increased path loss, therefore, higher frequencies are allocated to small cell base stations to gain the capacity without the need for higher transmission power. Small cells are the answer to the technological challenges of creating a wireless access network that connects the mobile devices, machines and objects to a processing cloud engine. It is important to note that the costs of a small cell BS only accounts for approximately 20% of the total deployment costs associated with outdoor small cells in 2015. The majority of the costs are site leasing (26%), backhaul (26%), planning (12%), and installation (8%) [25]. This information leads to change to the deployment model from operator deployment to a "drop and forget" user deployment, and reusing of existing power and backhaul infrastructure, that reduces small cells cost to a fifth of the conventional cost. The small cell concept covers a number of different options from relatively high power outdoor micro cells to indoor Pico-cells and very low power Femtocells. The number of commercial small cell deployments is more than 13 million cells, which already exceeds double of the number of conventional macro-cells worldwide [25]. Today, it is widely accepted that small cells are an essential component of future networks, representing the foundation of all ultra-high capacity (>1Gbps) wireless networks going forward [25].

To summarize, increasing the magnitude of network capacity is becoming a necessity for MNOs. Increasing network capacity, data rate or in any other QoS parameter, leads to enhance the level of services offered to end user, which consequently enhances end-users' QoE level or user satisfaction as described earlier in section 1.3. Relying on increasing the capacity by increasing the bandwidth means utilizing higher frequency carriers. mmWave frequencies suffer from path loss and lacks the penetration ability, therefore, small cell-oriented networks have been adopted. Small cells improve network performance in two folds. First, it offers cell densification, second it enables the use of mmWave frequencies at low power, thereby increasing small cell capacity linearly with bandwidth [26].

2.4.1 Brief history of small cells

Early deployments of smallcells were not considered as a solution to increase the capacity and improve network performance, due to their planning and management mechanisms, where early small cells had to be planned, managed, and interfaced with the network the same way as macrocell BSs, which meant the cost associated with deploying many small cells outweighed the benefits [26].

Later-on cost-effective small cells have emerged, which paved the road for later versions of small cells such as femto-cells [26].

Femtocells are small cells with extended capabilities, such as self-optimization and auto-configuration, which offered a low-cost solution to enhance network coverage and performance [27][28][29]. The auto-configuration feature enabled the deployment in plug-and-play manner for home environments. Femtocells are considered as the enablers for hetrogenious network deployment model. Femtocells gained increased interest from researchers and research bodies, for instance, the European Union has started funding research on femtocells [30][31].

Several operators around the world have started deploying commercial Femtocell deployments for residential environments, such as Sprint (US), Vodafone (EU), Softbank (Japan)…etc.[32]. By 2015, 71 carriers operated in-building small cells, known as pico- or microcells, in enterprise or public buildings [33]. Their design is tailored to reduce planning and deployment costs, decrease the need for large customer support teams, and eliminate the need for massive reprovisioning. The generally good availability of Internet protocol (IP) backhaul such as Ethernet in indoor spaces is an important deployment advantage. In-building small cell deployments are equipped with quality of service (QoS) and per-call analytics, as well as self-organizing network (SON) features.

In United States and Europe, operators continued to deploy small cells within their network architecture. By 2016, 13.3 million small cells were shipped, and the small cell market size increased to more than $1 billion annually [33]. Based on the required capacity by end users, it is predicted that small cell densification solutions, will keep on going for residential and enterprise environments, and the market revenue will reach $6.7 billion by 2020 [33].

In addition, small cells were also predicted to play a big role in 5G, especially in terms of low latency and high data rates due to the close proximity to the user.

2.4.2 Small cell Deployments

There are several deployment options for small cells within the mobile network architecture. Figure 2-10 depicts three deployment options. Figure 2-10a, presents

all-in-one small cell deployment, small cell base station is visible to the core network, and it uses S1 interface to communicate with LTE core and X2 interface to communicate with macro base station and other small cell base stations. Figure 2-10b, presents the Cloud Radio Access Network (CRAN) deployment; it utilizes a centralized baseband processing unit at the macro-cell base station and remote low-power Radio Frequency (RF) heads. With this deployment, small cells are not visible to the CN, rather they are considered to be as another sector of the macro-cell base stations. RF heads use Open Base Station Architecture Initiative (OBSAI) interface to communicate with the macro-cell base stations. However, this interface requires high bandwidth and low latency. Figure 2-10c, shows small cells with DC option. This type of small cell deployment is supported in 3GPP specification release 12. UEs under this deployment maintain two simultaneous connections with macro-cell and small cell base stations. Small cell base station houses the functionalities with low latency requirements, Layer 1, Medium Access Control (MAC) and Radio Link Control (RLC), while keeping the Packet Data Convergence Protocol (PDCP) layer at the macro-cell base station. There are several deployment variants under DC deployment option, where small cell can have direct link for the CP and User-Plane (UP) with CN directly, or can have only direct connection for the UP with CN. DC options are explained in further details in Dual Connectivity (DC) section.

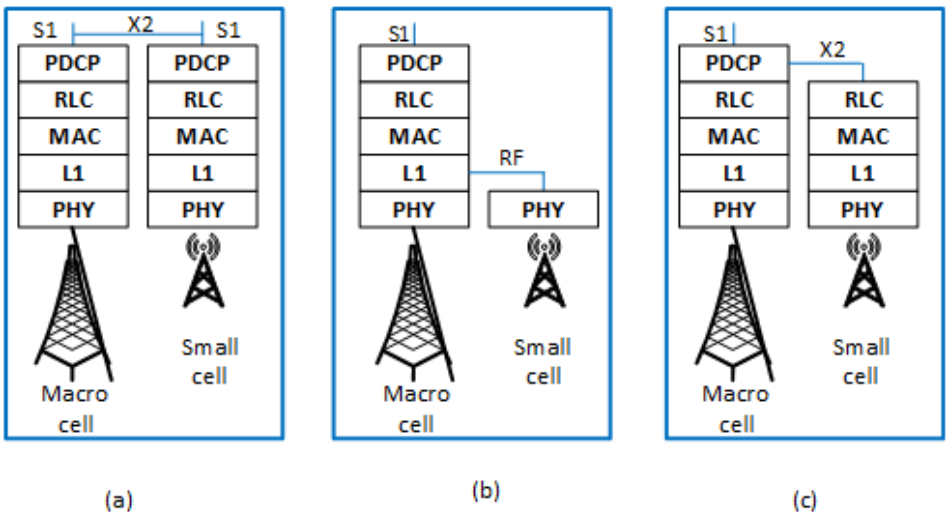

Figure 2-10 small cell deployment options. a) All-in-one base station, b) Cloud RAC (CRAN) Low power RF head, c) Release 12 Dual Connectivity

The requirement of the transport link for the all-in-one base-station is more relaxed, which simplifies the practical deployment. However, the visibility of the all-in-one base station by the core, negatively affects this deployment, since it would increase the signalling to the CN, due-to inter-site handovers.

The transport link between the macro-cell base station and the CN is referred to as backhaul link, since it is located behind the baseband unit, while the transport link between the RF units and the macro-cell base station is referred to as fronthaul links. The requirements for the fronthaul is much higher than the requirements for the backhaul. In case of multiple small cells share the same transport link, it is possible to have total transport rate less than the sum of the peak rates, because different cells are not typically fully loaded at the same time, which is called trunking gain. The RF head requires also low latency in addition to the high data rate. The one-way delay must be clearly below 0.5 ms to run the low latency functions including retransmissions from the baseband, which means IP addresses cannot be used for routing the traffic between the macro-cell base station and RF heads. Also, the jitter must not exceed the order of nanoseconds, which implies that, fronthaul communication requires optical fibre links, to maintain the delay and jitter requirements [21]. It worth mentioning the term jitter could refer to multiple definitions, in this research, jitter is referring to the variation of speed in packets causing them to arrive to the destination with irregular delays.

2.4.2.1 Indoor Small Cells

Indoor cells provide a sustainable coverage and capacity solution in dense indoor traffic hotspots such as train stations, airports, shopping malls or enterprise buildings, consequently the next subsections review two types of network configurations for indoor mobile coverage, Distributed Antenna Systems (DAS) and Femtocells.

A. Distributed Antenna Systems

The traditional solution for the lack of mobile coverage inside indoor environments is the use of DAS. DAS is used to provide coverage to highly populated indoor buildings e.g. high-rise buildings, train stations, museums …etc. DAS system is depicted in Figure 2-11, where it comprises of high-power base station, coaxial cables, splitters, and small indoor antennas. The downside of DAS system is the installation cost, but there are many benefits once installed, firstly, it provides sustainable mobile coverage sufficient to access UE and eliminate the need to switching back and forth to the external mobile coverage, secondly, it facilitates sharing the system by multiple bands, technologies and operators.

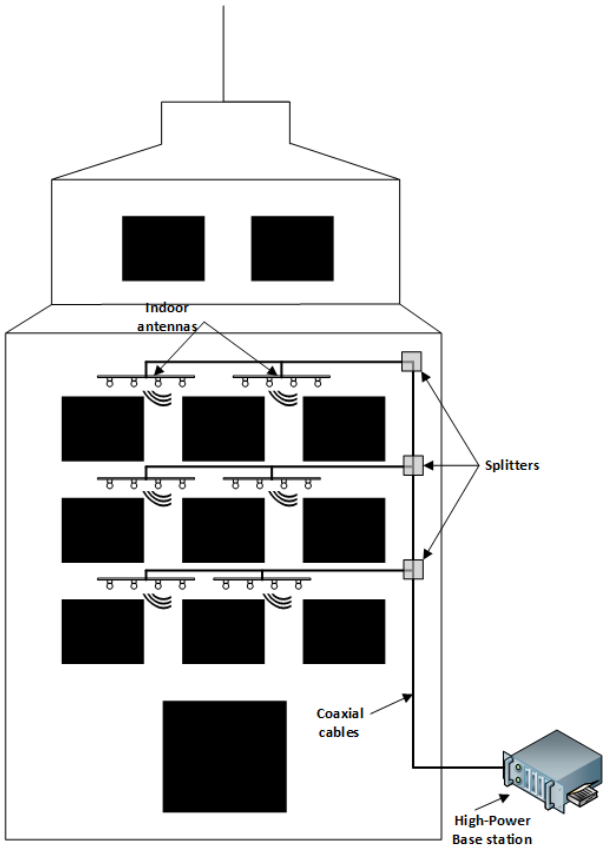

Figure 2-11 distributed antenna systems (DAS) architecture

The traditional DAS infrastructure includes single coaxial cable, which is sufficient for providing good signal levels and high data rate, however, a parallel coaxial cable is required to accommodate Multiple Input Multiple Output (MIMO) configuration. The capacity upgrade for DAS system is feasible when adding more frequencies, while it is more challenging when it requires adding more sectors, since it may require major redesign and cable installations. The system utilizes passive, active and hybrid distribution trends. Passive distribution refers to the process of using passive components to distribute the wireless signal to the antennas for transmission without any amplification. Active distribution involves converting the RF signal to digital signal, transmit it over optical fibre link and then convert it back to RF signal to be transmitted through the antennas. Hybrid distribution utilizes a combination of the previously mentioned passive and active distributions.

B. Femtocells

Femtocells represent low power base stations designed for indoor environments. Femtocells are connected to MNOs using broadband connection links. Femtocells are deployed in large numbers within MNOs since the third generation of Mobile Networks. The large number of femtocells energized lots of work to implement efficient ways for deployment. There are two main network architectures for

Femtocells deployment namely, Femtocell with Femto-gateway and Femtocells without Femto-gateway. Figure 2-12 depicts both network architectures. Deploying Femto-gateway in between the CN and Femtocells is shown in Figure 2-12a. Each architecture has its own pros and cons.

Femtocells with Femto-gateway: Femto-gateway is capable of covering large number of Femtocells, by terminating their connection at the Femto-gateway, while maintaining individual links with MME and SGW. All the Femtocells appear as a single Femtocell to the CN and reduces the mobility-induced signalling to the CN (by handling it locally by the Femto-gateway). The extra security node that can act as a firewall to protect the CN from malicious traffic from the Femtocell users. It also Optimizes the paging mechanisms at the Femto-gateway and reduces the scaling requirements for the SGW and MME.

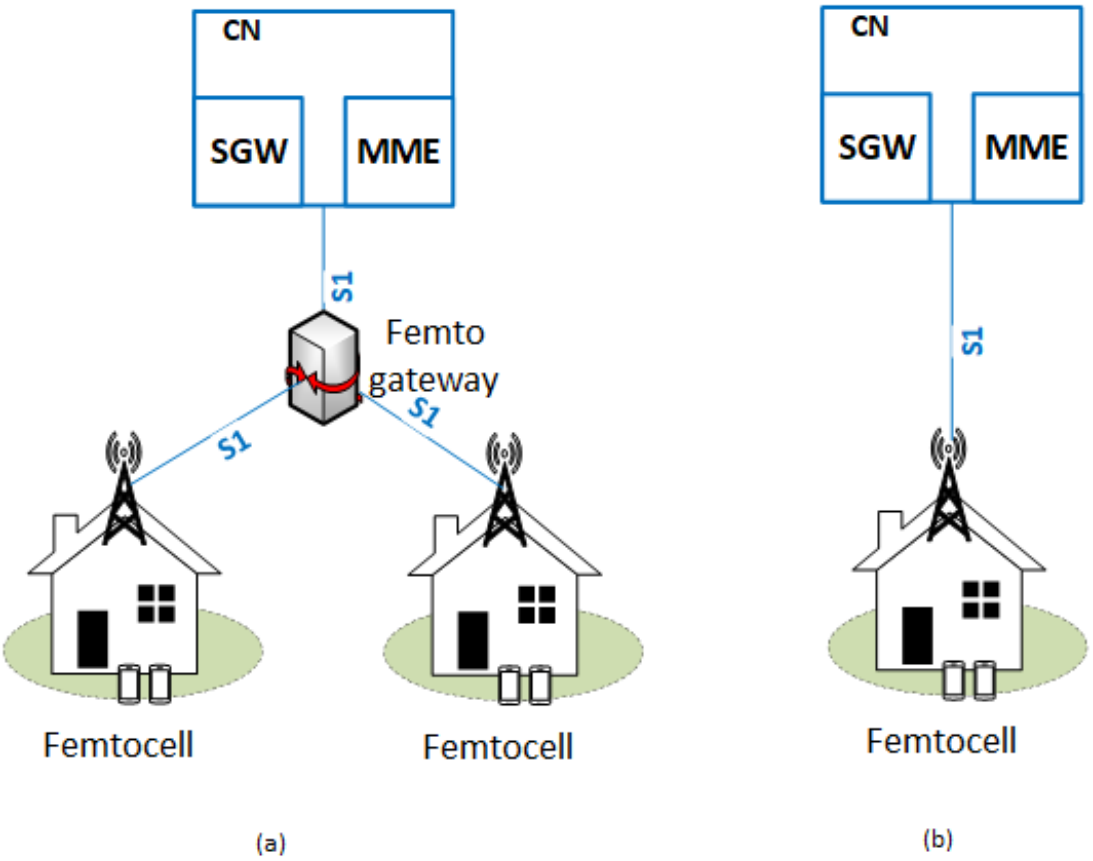

Figure 2-12 Femto architecture options a) with Femto gateway, b) without Femto gateway

Femtocells without Femto-gateway: not including Femto-gateway in the network architecture reduces an extra node in the network, eliminates the added latency by the Femto-gateway and removes the single point of failure for the connected Femtocells.

Furthermore, Femtocells offer local breakout as the process of routing some of the data towards local caches or towards the Internet connection without going through the CN. Figure 2-13 depicts the network architecture with Internet breakout links [28]. The local breakout has the benefit that the user can access local content and devices, for example, in-home multimedia, data backups or enterprise intranet. The benefit for the operator is that the local traffic is offloaded from the operator's core network. Femtocells are designed with automation mechanisms for quick spawning in plug and play mode, in order to reduce the cost of deployment.

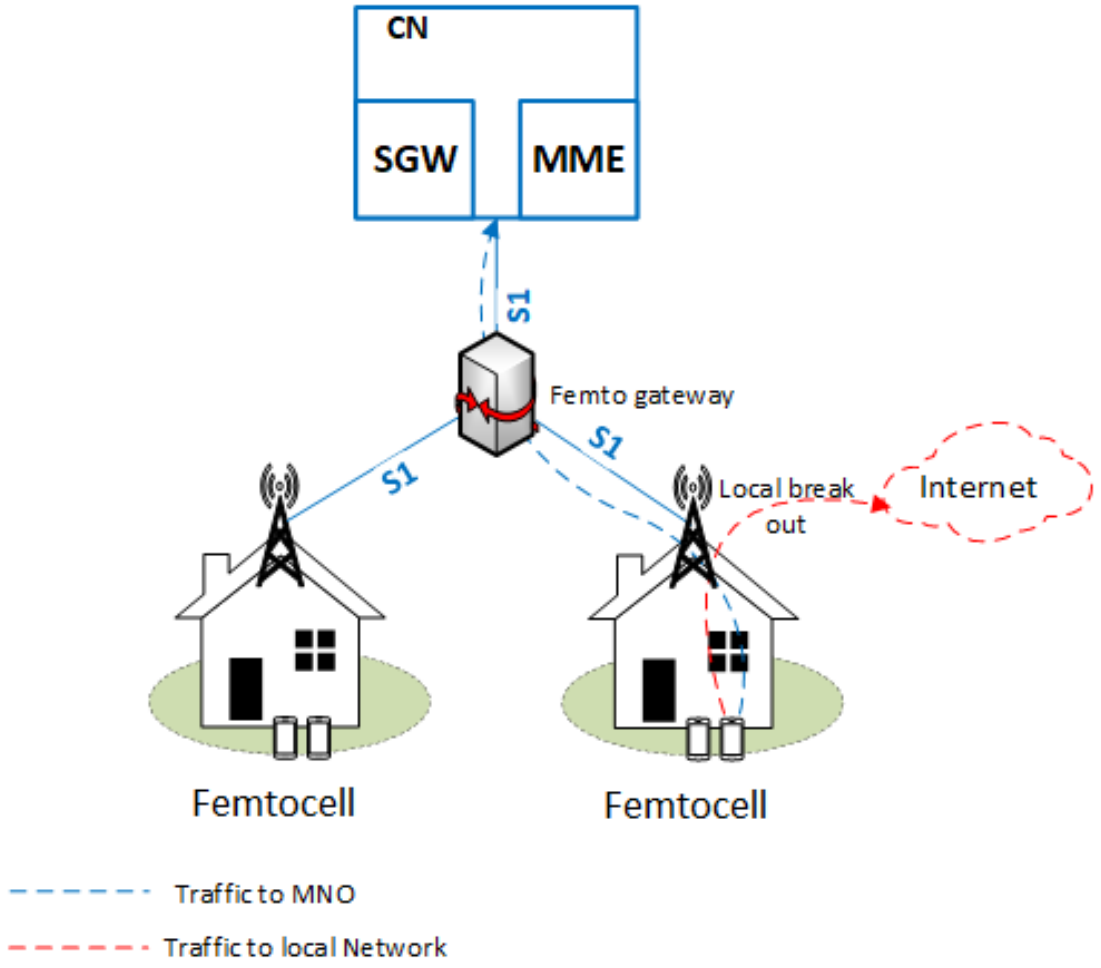

Figure 2-13 local breakout concept with femtocells

2.4.3 Location-based services

Femtocell based infrastructures are becoming one of the key components in the HetNet deployments [34]. Femtocells are utilised to provide better coverage, capacity and reduce traffic loads from the macrocell. Furthermore, Femtocell supports the deployment of many services. The type of the service is related to the deployment nature (Indoor/Outdoor). In general, Femtocell services can be classified into four categories as illustrated in Figure 2-14. Carrier services,

Internet Services, Femtozone Services, and Connected Home Services. Carrier Services are the common and basic services offered by Mobile Operators. Examples of these services are Voice Calls, Short Message Service (SMS), and push-to-talk. Internet Services such as e-mail, Web browsing, ftp, and mobile TV are accessible by any Internet access method such as mobile broadband and are based on third-party platforms. Femtozone services represent the services offered by the Femtocell itself. Femtozone services are realized by exploiting user information (e.g. location information), such as I am Home Service. This service leverages the Femtocell knowledge of its clients' location to inform other members about the arrival of users at home. While Connected Home Service, represents the services offered to UEs while they are within the home premises, such as automatic music synchronization. Connected Home Services enables the user to use its own device as a remote media controller [35].

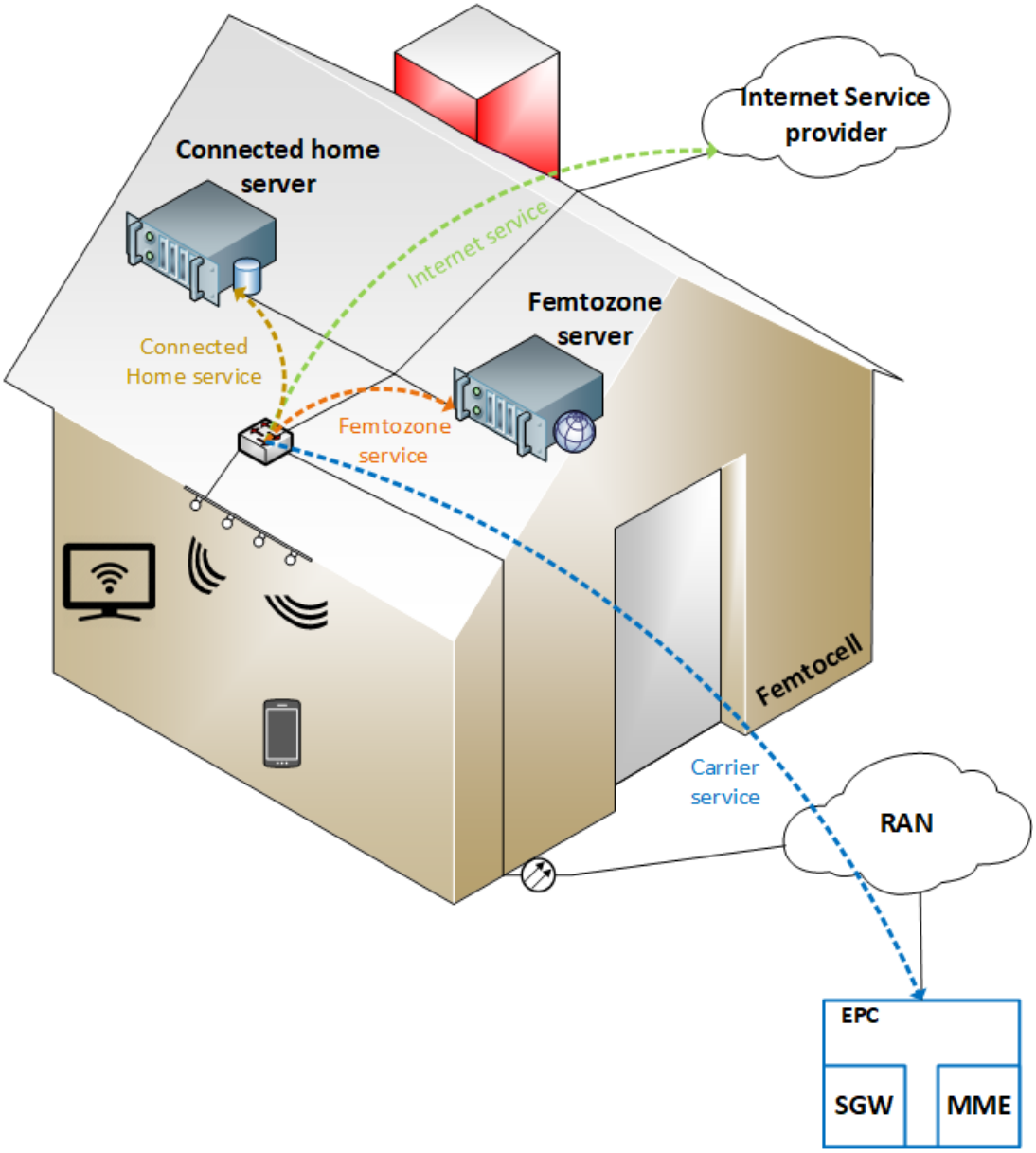

Figure 2-14 types of Femtocell services

2.4.4 Caching and content delivery networks

The demand for content delivery in mobile networks have had explosive growth in recent years, owing to an increasing number of mobile devices, high-speed wireless broadband access, as well as the boom of mobile applications. Internet Service Providers (ISP), MNOs and other service providers are always looking for mechanisms and technologies to enhance the network performance. One of these technologies is the use of Content Delivery Network (CDN). CDN is a set of proxy servers and data centres networked together and distributed over a specified geographical area. The distribution of the servers over the geographical area is related to the location of the end users. CDNs serve a large portion of the Internet content today, including web objects (text, graphics and scripts), downloadable

objects (media files, software, documents), applications (e-commerce, portals), live streaming media, on-demand streaming media, and social media sites [4]. At the most fundamental level, a CDN is about the efficient movement of digital content across the Internet middle mile on a large scale. Figure 2-15 depicts the usefulness of CDN with video streaming example. To provide video contents to each end user in coverage area A, there must be sufficient bandwidth at links 1, 2 and 3, which requires costly rentals for high bandwidth links. When the content server is located in a remote central location, then the performance for end users will not be satisfactory (lots of buffering and jitter). A CDN can be used to tackle such scenario where it helps to minimize the cost of video content delivery, ensures that network resources are utilized efficiently, and optimizes end-user experience. Without a CDN, bandwidth is overused due to the frequent content retransmissions. In summary, using CDN helps optimizing bandwidth usage to the point where it would appear as if a single end-user has requested unique content only once. No additional investment is required by the content provider to increase bandwidth capacity, as media content is packaged for delivery by the CDN infrastructure [36].

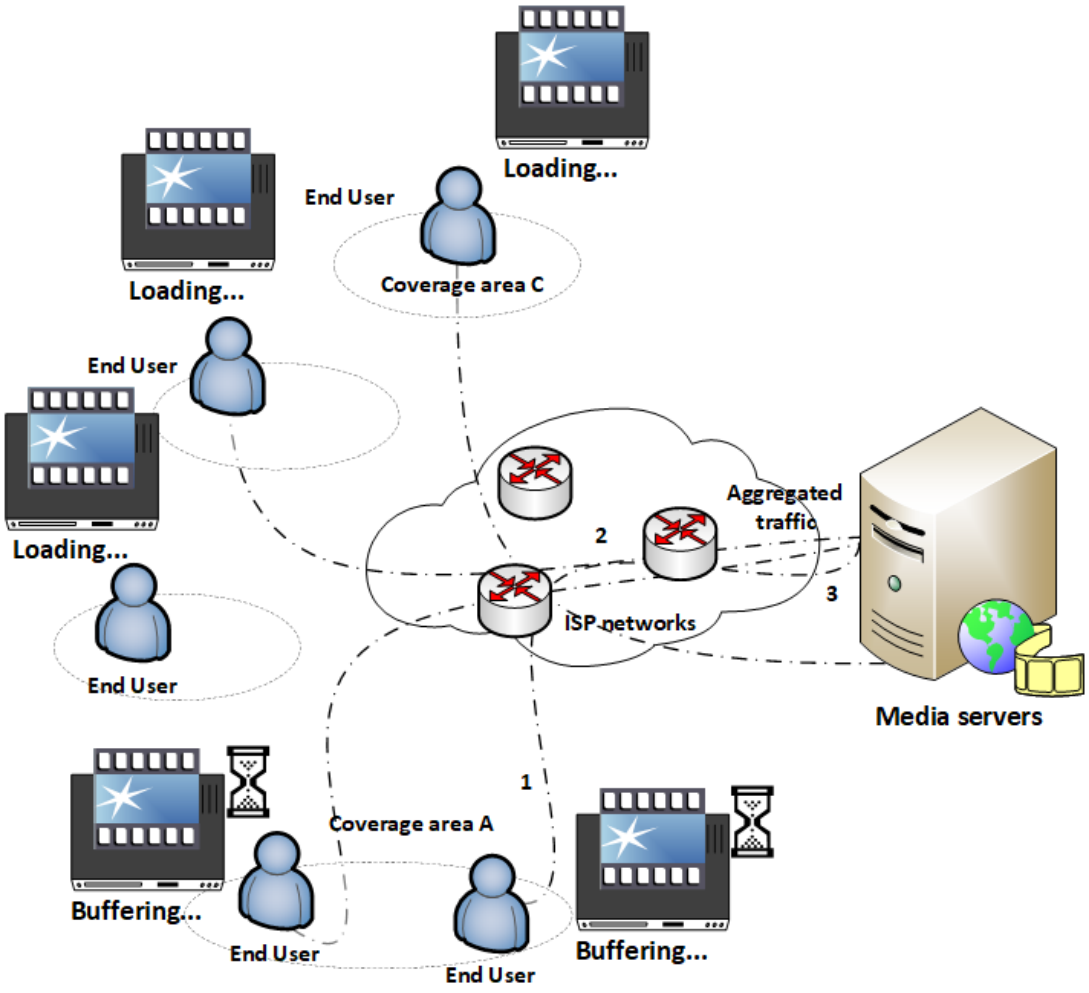

Figure 2-15 CDN in video streaming use case scenario

Deploying caching servers at UE proximity within the Radio Access Network, enhances the network performance in terms of reducing the content's latency and ultimately improving the UE's QoE. Distributed servers also reduce significantly the need for high bandwidth backhaul links, which in turn reduces the OPEX of the MNOs [37]. Research works of [38][39][40][41][42][43] confirm the aforementioned network enhancements achieved by utilizing Content Delivery Networks and Caching servers. Authors in [44] provided an adaptive video streaming service. The proposed service exploits the cloud resources and provisions them dynamically as required, thereby achieving optimum resource utilization and cost efficiency. In [45] authors proposed edge caching service with big data analytics to be deployed at the edge cloud, to achieve better UE'S QoE and network backhaul offloading. On the other hand, authors of [43] and [46] have presented a content placement and delivery approach based on the popularity prediction, with regard to wireless caching deployment, authors have highlighted the trade-offs between spectral efficiency, energy efficiency, and cache size. In [47], hybrid caching is presented as a solution for enabling data fetching from nearby cache, thereby eliminating the need for fetching from servers located at remote clouds. This approach is significantly important for delay sensitive applications e.g. augmented reality, virtual reality and so on.

2.4.4.1 CDN Types

CDNs are classified into several types based on their deployment option and targeted market, as listed below:

- **Pure-play CDN**: it is a CDN type, which utilizes its own infrastructure or ISP's infrastructure to deliver the content; however, ISP does not control or distribute the content.
- **Carrier/Telco CDN**: it is considered as a broadband provider CDN, it is exploited by MNOs and ISPs to reduce the OPEX, by eliminating the need for high bandwidth backhaul links.
- **Managed CDN**: this type of CDN is being built for the carriers, by the help, support and management of the "pure-play's" professional service group. This approach leverages the expertise, infrastructure, and software of a pure-play CDN.
- **Licensed CDN**: similar to the "Managed CDN", Licenced-CDN is being built, tested and deployed by the support of the (pure-play's) professional service group; however, the network operator manages it.

- **Federated CDN**: it refers to multiple interconnected CDNs. Federated CDNs are created to compete more directly against pure-play CDNs [36].

Each CDN type has its own challenges. Therefore, researchers are constantly working on finding solutions, for instance, authors of [48] studied the use of Artificial Intelligence (AI) with carrier/Teleco CDN. Specifically for Mobile Social Network (MSN), they proposed the use of Machine Learning techniques to overcome the challenges that faces CDN in mobile networks, such as: UEs uneven distribution, online/offline temporal dynamics, dynamics of video types request...etc. Authors of [36] have described various deployments of CDN for mobile networks, referring to different studies, which highlights the significant benefits in bringing the content inside the mobile networks. The improvement is achieved by combining information including user preference, device preference, location, and flow information as part of mobile CDN delivery.

2.5 Dual connectivity

MNOs have the responsibility of increasing the network capacity to meet the ever-increasing user's data demands. Referring to Shannon's capacity formula [24].
Network capacity as shown in equation 2.1, can be increased by one or a combination of the following factors, spectrum, cell densification and spectral efficiency. Small cells deployment is the way to implement cell densification, while Dual-Connectivity (DC) is one of the most promising techniques for increasing the capacity through intelligent cell densification.
The cost performance ratio of small cells is known to be minimized by having them tightly integrated with the macro layer. Integration of macro and small cells comes in many forms depending on the radio frequency deployment, the type of small cells, and the corresponding inter-node connectivity architecture. Examples of small cell deployment options are outlined in [35]. Deploying a central baseband processing unit, which is connected via optical fibre link to small and macro-cells' Remote Radio Head (RRH), is called Cloud Radio Access Network (CRAN), and it is found to be the fastest integration method.
DC is another feature of mobile networks; it is introduced in Release 12 of the 3GPP specification [37], as an enabler technology for increasing per-user capacity (data throughput). DC is realized by utilizing the radio resources of two base stations, by enabling UEs to communicate to both base stations simultaneously. The macro-cell base station is considered as the primary base station or Master eNB (MeNB), while the small cell base station is considered the secondary base station (SeNB). The UE

is connecting to the MeNB to maintain the mobile network communication and to the SeNB to gain access to extra capacity [37].
In dual connectivity deployments, eNBs use the traditional backhaul network with X2 interface, however, this type of connection is found to be lower capacity and higher latency with comparison with optical-fibre links. Therefore, X2-based backhauls are also referred to as non-ideal backhauls.

2.5.1 User-Plane with DC

UEs can transmit and receive UP data only during Radio Resource Control (RRC) connected mode. In DC network infrastructure, UEs are configured to have simultaneous connections with MeNB and SeNB, therefore, for UP options with DC, there are multiple UP connection options as discussed below.
According to the 3GPP standards, there are two types of DC implementations: DC with UP split at the CN and DC with UP split in the MeNB [36].
UP split options with DC is depicted in Figure 2-16. In the case of implementing UP split at the CN, each eNBs has a separate connection with the S-GW of the CN. In addition, data is sent over a separate set of data bearers to/from each eNB, as shown in Figure 2-16 b. differently, with the second type of DC, the split occurs at the MeNB .i.e. there is only one connection with S-GW at the CN for both eNBs, and data is sent only to MeNB, as depicted in Figure 2-16 c.

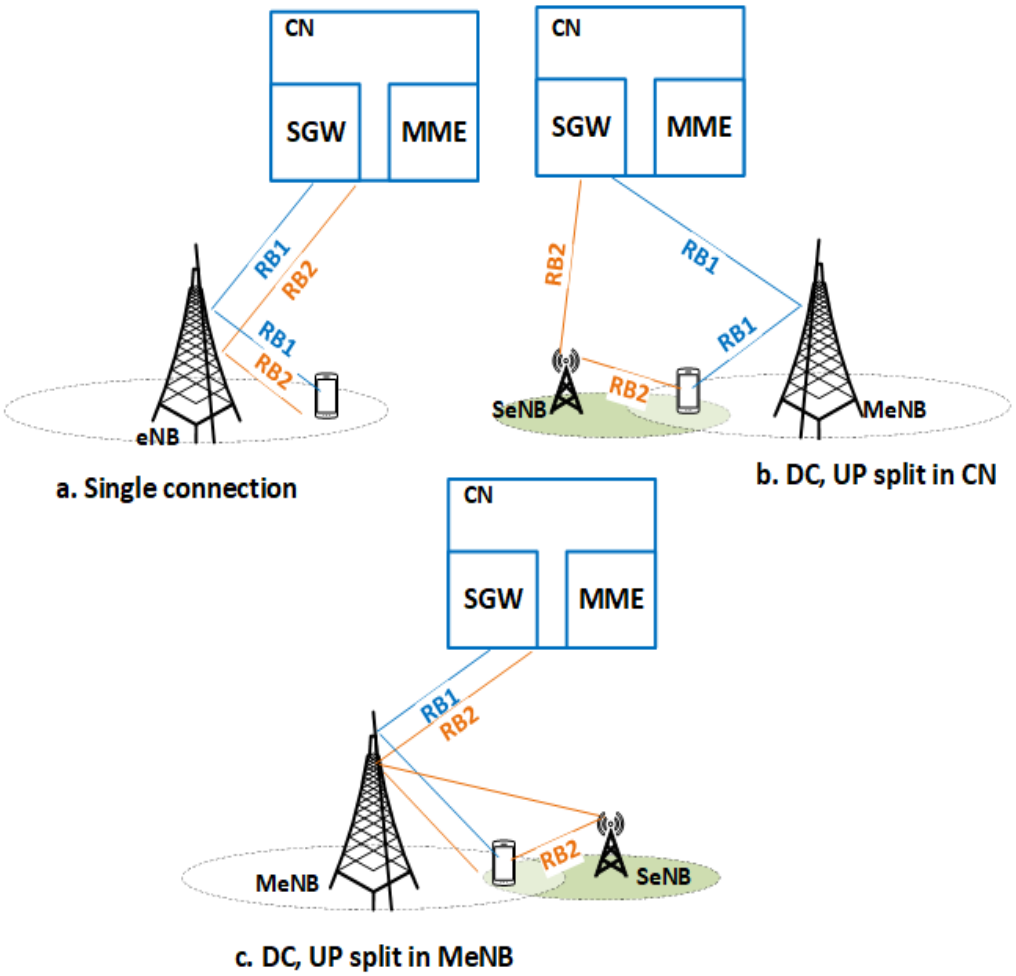

Figure 2-16 Dual connectivity architecture and U-Plane options; a. single connection, b. Dual Connectivity with User Plane split in Core Network, c. Dual Connectivity with User Plane split in the MeNB

With DC, there are three sets of radio bearers, Master Cell Group (MCG), Secondary Cell Group (SCG) and Split Bearer (SB). MCG are the set of bearers that are served by MeNB alone, SCG are the sets of bearers served by SeNB alone, while SB are served by both MeB and SeNB

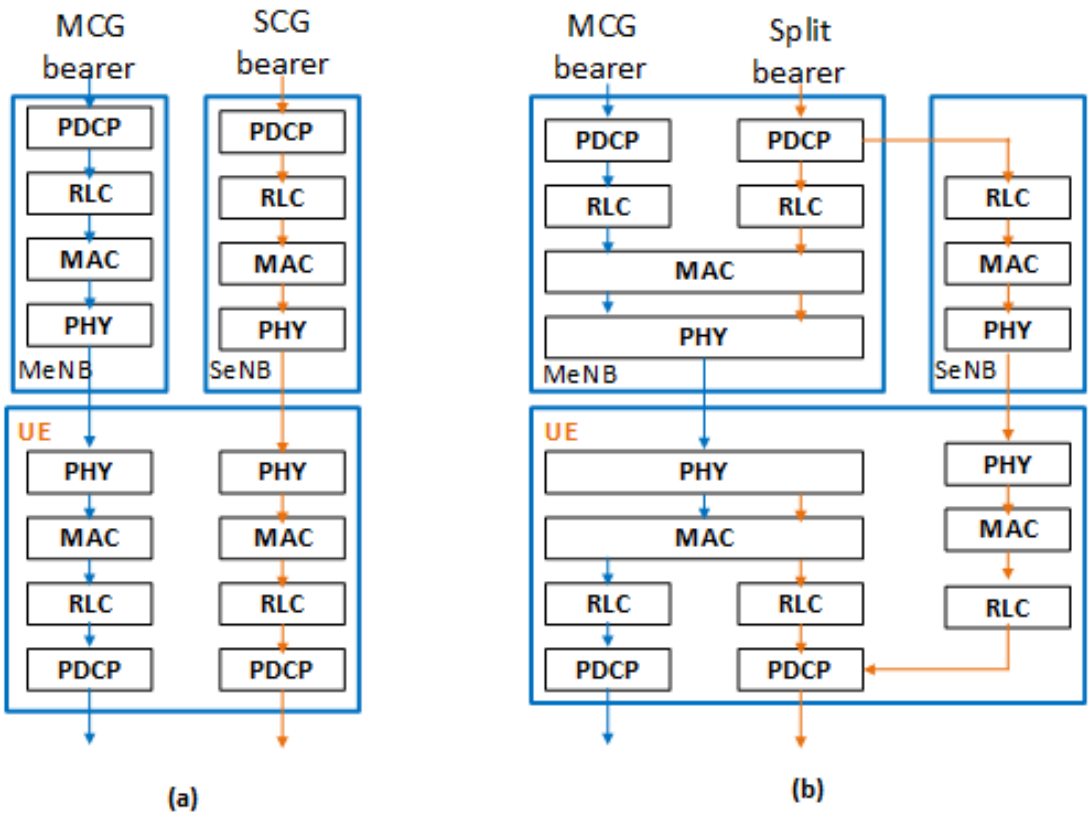

Figure 2-17 Dual Connectivity with MCG, SCG and split bearers a) split at CN, b) split at MeNB

Figure 2-17 depicts the User-Plane protocol stack at the MeNB, SeNB and the UE. Figure 2-16 a depicts the stacks in case of UE configured with one MCG and one SCG bearer. Figure 2-16 b depicts the stack of layers in case of UE configured with one MCG and one split bearer. The arrows illustrate the flow of U-plane data in downlink. The X2 specifications are enhanced, enabling the SeNB to send radio bearer level information on its available buffer size to the MeNB, to help formulate the load balancing decisions.

2.5.2 Control-Plane with DC

In LTE networks, the CP signalling with DC is always terminated at the MeNB to simplify operation and reduce complexity. The system monitors the radio link by performing Radio Link Monitoring (RLM). RLM is a monitoring procedure that monitors the radio links continuously and takes an appropriate action when it detects Radio Link Failure (RLF). With DC deployment, if the RLM detects RLF at the MeNB, it will trigger RRC connection re-establishment procedure, while it will not trigger such procedure in case of RLF detected at the SeNB, because the RRC connection towards the MeNB is still maintained even if the link to the SeNB has failed.

2.5.3 DC with 5G

The 3GPP has identified multiple architecture options for the 5G deployment [38]. It is widely accepted that, a fully 5G network in the form of Stand Alone (SA) is the ultimate goal for the MNO, meanwhile, there will be intermediary stages to go through. In these intermediary stages, DC network architecture is adopted in the Non-Stand Alone (NSA) deployment, to enable the LTE/LTE-Advanced (LTE-A) to offer the coverage and mobility support for the UEs, while the 5G provides the high data rate coverage, which enhances the capacity and facilitates the deployment of data-hungry applications and delay-sensitive data.

DC is designed to enable simultaneous connectivity of the UE with Macrocell and small cell, however, the fact that small cells can be implemented to cover outdoor or indoors, therefor I propose to use DC with IoRL UEs as the IoRL is considered as a small cell dedicated to indoors environment. j

In the section below, an overview about the possible deployment options for 5G networks will be provided, which include NSA and SA deployments [39].

2.5.3.1 Stand –Alone New Radio:

This network architecture is the ultimate goal for the MNO, as it represents the complete 5G mobile network, including 5G Core (5GC) and 5G-New Radio (5G-NR) RAN. The network architecture is depicted in Figure 2-18. With Stand-Alone New Radio (SA-NR) there is no LTE components, as it is a complete 5G deployment with gNB communicating with 5G-Core (5GC). This architecture is defined in 3GPP release 15.

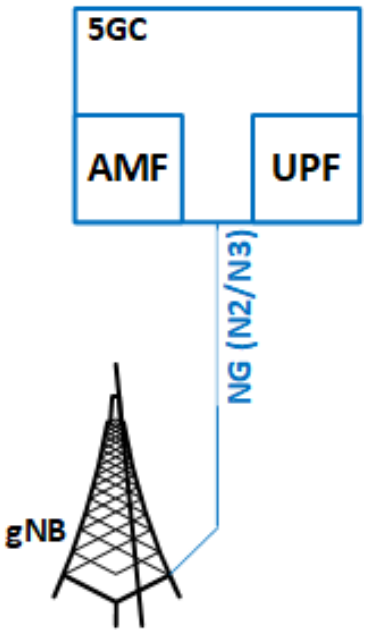

Figure 2-18 SA-NR deployment

2.5.3.2 Non-Stand-Alone New Radio in Evolved Packet System:

NSA-EPC is a combination of LTE and 5G in DC orientation. NSA-EPC is considered the most realistic deployment option, as it does not cost the MNO as much as the SA-NR, since it leverages the existing EPC along with LTE eNB. Network architecture is depicted in Figure 2-19. With this network architecture, eNB provides the anchor for mobility management and network coverage, while NG-gNBs provides higher data rates and lower latencies. Within this deployment, UE is connected to an eNB, which acts as a Master Node (MN) and to a NG-gNB that acts as a Secondary Node (SN). There are two possible connections for the small cell; first option is shown in 2-18.a, the UP is terminated at the MN, while the second option 2-18.b, UP is being terminated at the CN by a direct connection with the CN.

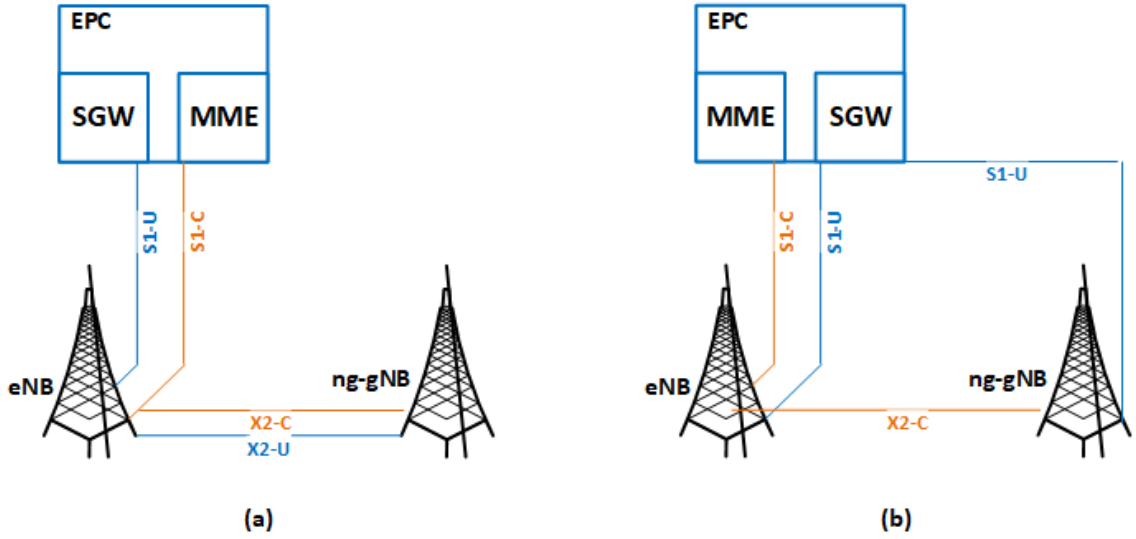

Figure 2-19 NSA-EPC deployment a) UP terminated at eNB, b) UP terminated at CN

2.5.3.3 Non-Stand-Alone LTE in 5GC:

This network deployment is proposed as a continuation of collaboration between 5G and 4G. Considering a realistic approach for the next generation, MNOs do not want to waste their current infrastructure due the emerging 5G, but rather prefer to adopt network architecture that leverages technologies from both generations intelligently; therefore, NAS-NR is a second network architecture that incorporates both 4G and 5G in one heterogeneous network. With NSA-NR UEs connected to the gNB as the MN, while NG-eNB acts as SN. The network architecture is depicted inFigure 2-20. Figure 2-20a, presents the UP of the secondary node, which terminates at gNB, while Figure 2-20b, presents the case when UP of the secondary node is being terminated at 5GC. This option includes 5GC deployment. gNB and NG-eNB are connected to each other utilizing Xn interface. The two options for UP termination are shown below.

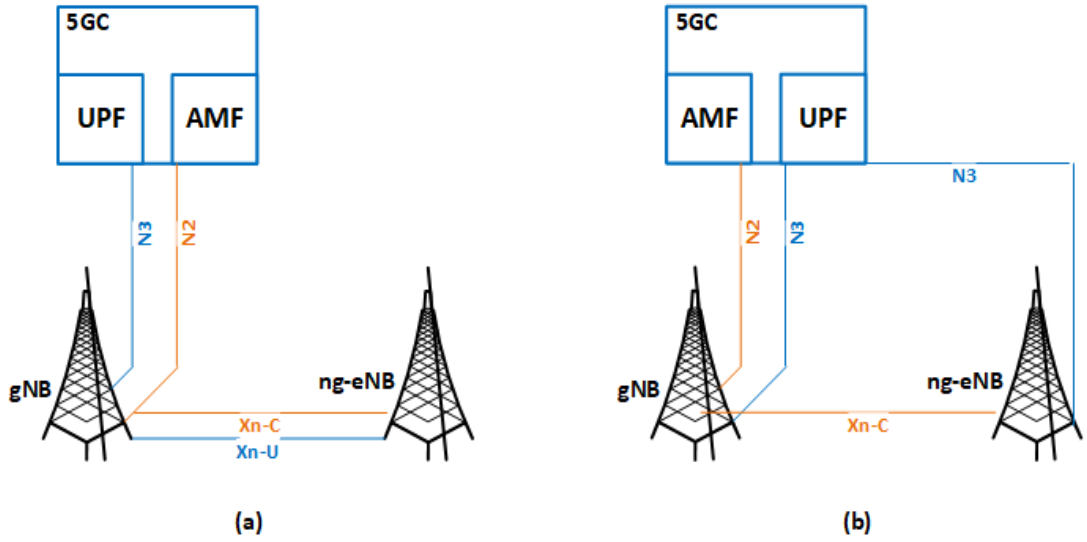

Figure 2-20 NSA-NR deployment a) UP terminated at gNB, b) UP terminated at 5GC

2.6 Summary

MNOs are facing an exponential increase in user's demands for higher data rate traffic. This demand is contributed by, service providers and application developers on one end, with Mobile devices manufacturers and users' expectation on the other end. On the other hand, traditional networking have its own vulnerabilities in control and management planes e.g. i) Interoperability amongst various vendors and even release versions of the same vendor. ii) Limitation of the available resources at the control entities. iii) The use of proprietary software and protocols limits the ability to establish a unified management platform for multivendor hardware resources and the lack of on-demand resource provisioning ability, which prevents optimal resource utilization.

Based on the aforementioned issues, this chapter presented a comprehensive survey about various technologies and concepts, which represent the basis for the mobile network evolution towards fifth generation mobile network. The survey referenced books and researched papers about enabling technologies e.g. NFV, SDN and their utilization within mobile networks, as well as the adopted concepts in mobile network architectures e.g. Edge Cloud and MCC.

As mentioned earlier that this chapter provided the research background and surveyed related concepts and technologies, the next chapter presents in-depth practical implementation walkthrough about private cloud infrastructure in the form of infrastructure as a service, in order to create hosting platform for agile service implementation and deployment.

Furthermore, small cells are getting an increased attention and research activities as a solution for enhancing network capacity and users' QoE, where small cells play

an important role in the development of the next generation mobile network architecture, therefore, extensive researched details were presented in this chapter about small cells' design and deployment options, especially their integration with MNO's RAN.

Finally, the DC mode was introduced as one of the most promising techniques for increasing the capacity through intelligent cell densification. Therefore, a thorough explanation was provided about control plane and user plane at various split options including layered architecture for each split option. Also, the importance of DC for 5G mobile networks was highlighted by providing the possible 5G deployment options within the DC context.

3 Infrastructure as a Service: Practical implementation model

3.1 Introduction

In the previous chapter, various technologies and concepts were presented as the drivers for the emergence of the fifth-generation mobile networks. In addition, the limitations and vulnerabilities of the current networking paradigms were highlighted, therefore, to bring the concepts and technologies mentioned in the previous chapter, the aims behind this chapter is to provide a detailed insight about the practical implementation methodology that has been used for creating a flexible and intelligent platform. The lack of flexibility and efficient resource utilization were tackled by creating cloud-computing based platform for deploying various services for end users [49]. Openstack platform is adopted platform for this task. Openstack is designed to enable automation, well-defined integration with various systems and hardware resources and autonomous service deployments. These advantages were achievable by multiple well-defined function methods, where for every function there are three possible methods for implementation, first, REST API, second Command Line Interface (CLI) and third Horizon (web-based) interface command[50].

3.2 OpenStack

Openstack is an open source cloud platform, which is utilized by service providers e.g., Rackspace [51] to provide shared infrastructure platform for multi-tenant environment. Similarly, end users utilize Openstack to create scalable public and private clouds. Thereby, many of the enterprise organizations exploits Openstack to create clouds as an underlying IaaS layer, PaaS or Hybrid Cloud deployments [50]. Openstack foundation [52] is the governing body of Openstack platform. The foundation is comprised of thousands of members; however, technical and financial issues are dealt with by special members whom are elected and assigned by sponsors to work in the technical committee and board of directors respectively.

3.2.1 Private cloud platform VS traditional data center virtualization

One of the main features in Openstack platform is to enable the coexistence of multiple end users on the same platform yet providing them with completely separate domains. This feature is achieved by labeling each virtual object that exists in one of these separate domains with a specific tag, to distinguish them from each other. User's objects belong to the same private domain referred to as tenant. Opensatck utilizes special identity service to authenticate and authorize each tenant separately. Tenants enjoy high level of resource segregation, which provides them with the experience and performance of sole resource ownership. Conversely, with traditional data center virtualization, there is no such mechanism to enable the coexistence of multiple tenants, therefore all the resources are being trapped for the disposal of single user [50].

3.2.2 OpenStack Switching

Network switching is enabling communication between networking devices within the data link layer of the OSI model, it is facilitated by networking device named a switch. On the other hand, virtual switching is enabling virtual machines to communicate with other virtual machines via virtual network links. Openstack platform utilizes Neutron project to handle the networking bit within the platform. Neutron supports various types of virtual switching mechanisms e.g. linux bridge, Open virtual Switch (OvS) and so on. In our deployment, linux bridge and OvS have both been utilized to provide communication. End users are not exposed to such details as they are being taken care of by the cloud administrator, while users are enabled to configure overlay networks and create new subnets and virtual routers. In the deep configuration process, there are different types of tunnelling e.g. L2-inL3, which are being utilised to connect VMs running on different hosts in single broadcasting domains. The details of the deployed Openstack platform is provided in later subsections of this chapter.

3.2.3 OpenStack Routing

OpenStack networking provides routing and NAT capabilities through the use of IP forwarding, IP tables and network namespaces. Within each network namespace, there is its own routing table, interfaces, and IPtables processes. The virtual routers are OvS based. Exploiting network namespaces, enables the neutron service to offer non-overlapping subnets to end users as well as allowing VMs attached to virtual subnets to interact with external/internal networks smoothly.

3.2.4 OpenStack Load Balancing

OpenStack offers Load Balancing as a Service (LBaaS) within its platform, which is essential for distributing client requests across multiple instances or servers. OpenStack's Load Balancer is configurable with connection limits to take care of balancing as preconfigured. Also, OpenStack Networking is equipped with a plugin for LBaaS that utilizes High Availability Proxy in the open source reference implementation.

3.2.5 OpenStack Firewalling

OpenStack networking offers two methods for network security purposes, for VM security precisely. That are: security groups and FireWall as a Service (FWaaS). The former is related to Nova compute project, which places the VM within groups that share common functionalities and rule sets, while the latter is related to neutron project. Both of the methods are enforcing the security policies at the port level, with one difference between the two implementations, that is, security groups implemented at the port level of the VMs only, while FWaaS could be implemented at both VM and routers' ports to enforce the security rules.

3.3 OpenStack Components

This section provides brief overview about most of OpenStack components and their role in the overall platform. OpenStack platform is updating constantly, thereby there will be more components deigned by developers to expand the platform capabilities to deal with new technologies.

3.3.1 Nova

Nova is the compute component of Openstack, it manages the interaction with underlying hypervisor in order to provision the virtual machines (instances). Nova supports various types of hypervisors e.g. KVM (Linux based), QEMU, Hyper-V (Microsoft products) and so on [53][50].

3.3.2 Swift

Swift is the object storage component in Openstack, which is technically cloud-based binary storage. Swift deals with underlying physical hardware and creates an abstraction of that hardware, thereby end users are able to store their data safely without dealing with specific physical hard-disc. Binary form of data storage enables Swift to make replicas of that data to store it in multiple nodes; therefore, in the instance of single physical node failure, the data will be safely available. Safe deployment recommendation requires three physical nodes minimum. Data stored using Swift is made accessible to applications via Swift proxy, which keeps track of each object storage nodes and ensures the data availability as requested.

3.3.3 Neutron

Neutron is the networking component in Openstack, which is SDN based. The main advantages provided by SDN to Openstack are firstly, creating one logical broadcast domain on a physical infrastructure that are many routers apart, secondly separating traffic belongs to various tenants, while being sent over the same infrastructure. It thereby enables cloud users to create logical networks (with one broadcast domain) using physical infrastructure that is possibly within different physical broadcast domains. The deployed VMs are behaving as expected according to the network configuration by the end user. Neutron in Openstack is driving the networking of the cloud, so that there are two types of networks, overlay and underlay networks. Underlay network is the physical network that collectively provide the physical resources to the cloud, while, overlay network is the SDN network that is offered to the end users to create their own broadcast domains, while its resources are being abstracted from the underlay network.

3.3.4 Glance

Glance is the image service within Openstack. Cloud users are interested in getting faster deployments for their solutions; this includes fast deployment of VMs rather going through the lengthy steps of operating systems' installation e.g. hard disk partitioning and so on. Glance provides the solution in the Openstack environment. Glance provides some sort of image service for the end users. Images can be downloaded or created with a predefined configuration, then the flavours service is utilised to select the logical hard disk, RAM and storage of the VM, to deploy the required VM with the requested specifications. All Linux distributions have ready-to-deploy cloud images; "Cirros" image is the lightest image with 13 MB size only, which is used for testing purposes.

3.3.5 Cinder

Cinder is the persistent block storage service within Openstack. Without cinder service, any information written to the VM would be bound to the current physical host that is currently hosting the VM, i.e. when moving the VM amongst the physical nodes (cloud computing concept) or in case of node failures, the information will be erased. This is against cloud's flexibility concept. Therefore, cinder provides a necessary storage service for VMs in cloud environment. When using cinder service, a VM will have a separate disk path for the storage e.g. /dev/vda and /dev/vdb, the former refers to the glance storage and the latter refers to the cinder storage. Using cinder storages provide flexible and persistent storage for the VMs.

3.3.6 Horizon

Horizon is the web user-interface service with Openstack, horizon available for both of the end users and cloud administrators. It enables tenants' administrators to deploy their VMs on their logical networks and cloud administrators to configure the cloud environment as required e.g. configuring quotas for tenants, configuring provider networks and so on. Almost every feature and service within Openstack could be configured using command line or horizon dashboard.

3.3.7 Keystone

Keystone is the identity service component within Openstack, which enables users to access various services within tenants' environment. Users are identified by keystone service using authentication tokens. It is an essential component of Openstack environment. It enables users from various tenants to work with various services of Openstack. Without Keystone everything drops and nothing can be connected anymore. Keystone also stores its information in a database. Typically MariaDB database on a modern Linux distribution.

3.3.8 Sahara

Sahara represents the integration interface of Openstack with big data. Therefore, Hadoop is integrated with Openstack using sahara project. Sahara is very well integrated with horizon of Openstack. Using horizon interface, the user can configure and launch Hadoop cluster consisting of master node and few worker nodes easily.

3.3.9 Trove

Trove is the Data Base as a Service (DBaaS) within Openstack. It is utilized to automate the allocation management of SQL databases. It enables end users to select, provision and operate data management infrastructure. Trove cooperates with various database interfaces e.g. MariaDB, PostgreSQL, MongoDB, Cassandra, and many more. Trove provides the end users with simple management interface to use in the process of deploying a pre-packaged database, thereby saving the time and efforts of their network administrator and eliminating the need for creating a database from a simple VM.

3.3.10 Designate

Designate is a DNS as a service component of Openstack. It is utilised to create DNS records for a singular tenant and multiple tenants. The site records are stored in pool arrangements; there are private and public pools, the former refers to internal sites, while the latter refers to external sites. Designate enables load balancing by spreading the load across the name servers; thereby users are not bound to a specific name server, rather interacting with a pool. It is also enables the network administrator to separate the internal DNS records from the external ones. Mini-DNS is part of the project, which it exploited to send the DNS request to the appropriate pool.

3.3.11 Heat

Heat is the orchestration service within Openstack platform. The main benefit of using heat orchestrator is to simplify and expedite service deployments. Services that are constructed from stack of multiple VMs are configured to perform various tasks, which are need to be deployed together at the same time to achieve the required tasks. Heat works by following the configuration information written in Heat Orchestration Templates (HOTs). Each HOT comprises of four elements namely, resources, properties, parameters and output. Resources refer to the VMs or objects that need to be created, while the properties specify the flavour specifics, furthermore, parameters refer to the properties of individual VM within the deployed stack of resources and finally the output, which represents the final solution that user has requested.

3.3.12 Ceilometer

Ceilometer is the telemetry project within Openstack. Cloud service-providers exploit ceilometer service to keep track of their user's usage. In cloud environment users' charges are based on their actual consumption of resources, thereby ceilometer enables cloud service-providers to bill their customers accurately.

Most of the mentioned services are managed via either command line interface or horizon web interface. Services and projects are constantly being upgraded and updated to fix previous bugs as well as to reflect the latest advancements in technologies. Also new projects and services are being developed to provide new services for the Openstack platform.

Other important services need to be configured correctly to enable Openstack to operate effectively, for instance, MariaDB is a standalone database solution for storing information. Since each service requires a backend storage service to store

data related to each of these services, MariaDB is one of these data base options. Another important aspect of Openstack, is the message broker. Message broker service is vital service for Openstack as it takes care of the messaging and communication among the various Openstack projects to deliver the required tasks, which enables the communication between various service even if they are existing on various hosts. There are various types of message brokers e.g. RabbitMQ, ZeroMQ, Qupid and so on. It is worth mentioning that MariaDB is a valid option for database service for Openstack deployment, however, it is not so flexible, to enable cloud database solution for Openstack, thus to allow services to access database service from anywhere in the cloud, then it is essential to exploit Trove project. As mentioned earlier, Trove provides cloud database solution, which is very flexible option for Openstack project to have such a cloud database.

3.4 Practical implementation walkthrough

The next subsections provide detailed insight about the (industrial-like) implementation of a virtualised environment for the IoRL's IHIPGW platform. OpenStack is utilized as the open source VIM. The physical resources are comprised of DELL R730xd server depicted in Figure 3-1. The server's basic specs are as following:

- CPU: 2x Intel(R) Xeon(R) CPU E5-2620 v3 @ 2.40GHz
- Memory: 192 GB RAM
- Disk 1: 240 GB SSD SATA
- Disk 2: 1 TB SATA
- Network I/F:
    - 2x 10GbE
    - 1x GbE

Figure 3-1Actual picture of the Dell R730xd server

The physical host (Dell server) is running an Ubuntu Linux 16.04LTS, while OpenStack (Queens Version)[54] is used as the cloud VIM and (Network Function Virtualization Infrastructure) NFVI enabler. OpenStack is currently the prevailing open-source cloud controller with a wide ecosystem of services and plug-ins. It is also the most widely used controller for NFV platforms, and also a part of the OPNFV (Open Platform for NFV) suite. The services can be deployed on VMs, and as VNFs. Figure 3-2 depicts the dashboard of the OpenStack VIM, giving an

overview of the cloud infrastructure that implements the IHIPGW. The OpenStack VIM is responsible for controlling and managing the NFVI compute, storage and network resources.

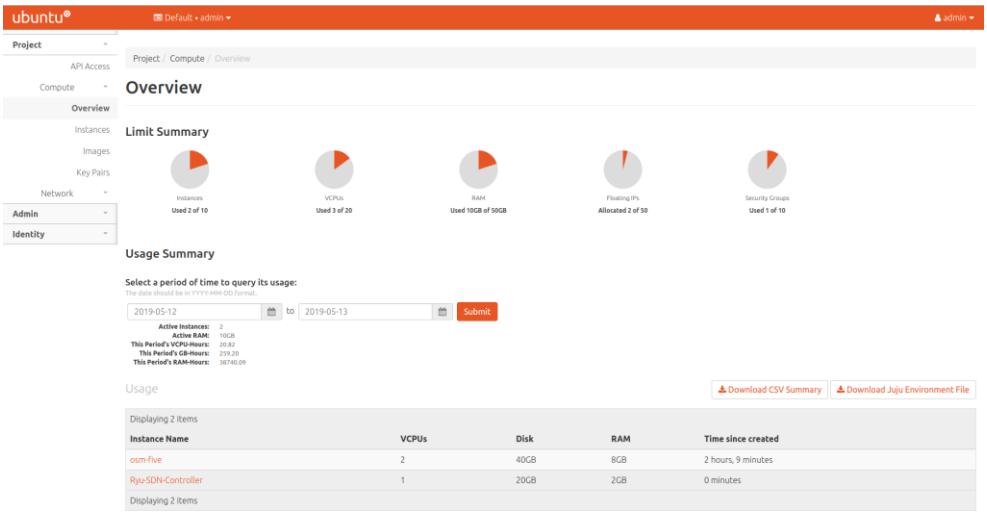

Figure 3-2 snapshot of the OpenStack Dashboard

3.4.1 Network design architecture

The IHIPGW platform is technically connected to three external networks namely, Provider, WiFi and VLC/mmWave networks. The network architecture is depicted in Figure 3-3. The figure highlights the external networks driven by the IHIPGW as well as the provider network. There are more networks configured internally, one for platform management and others for attaching the deployed VNFs to as will be shown in the VNF subsection.

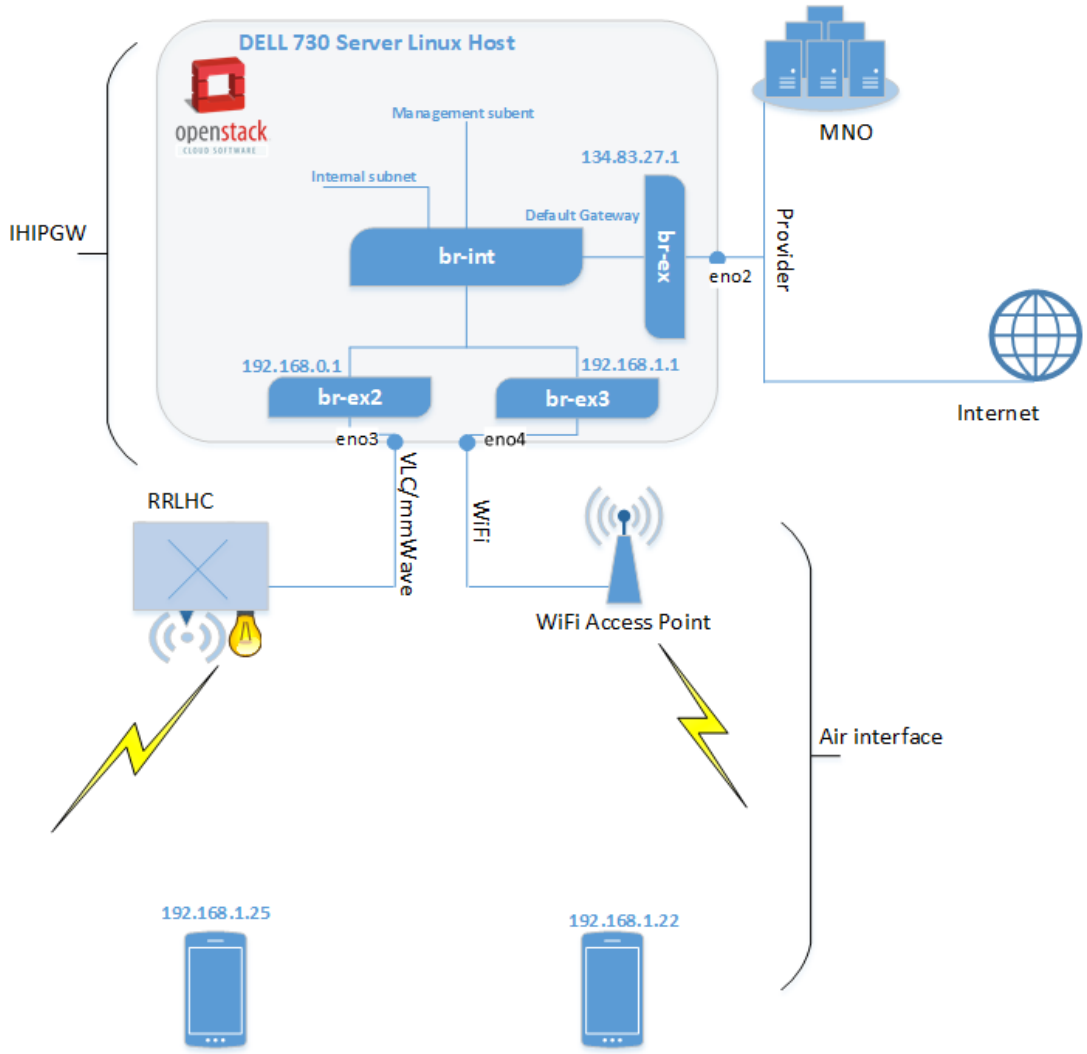

Figure 3-3 illustration of the network architecture

The network topology depicted in Figure 3-4. The terminologies refer to the actual purpose of the subnets e.g. the Uplink network refers to the wifi link, which is used for routing uplink traffic during the development phase, while the L2 subnet refers to the link to the layer 2 processing server and eventually Layer 1 processing and VLC/mmWave transmission links.

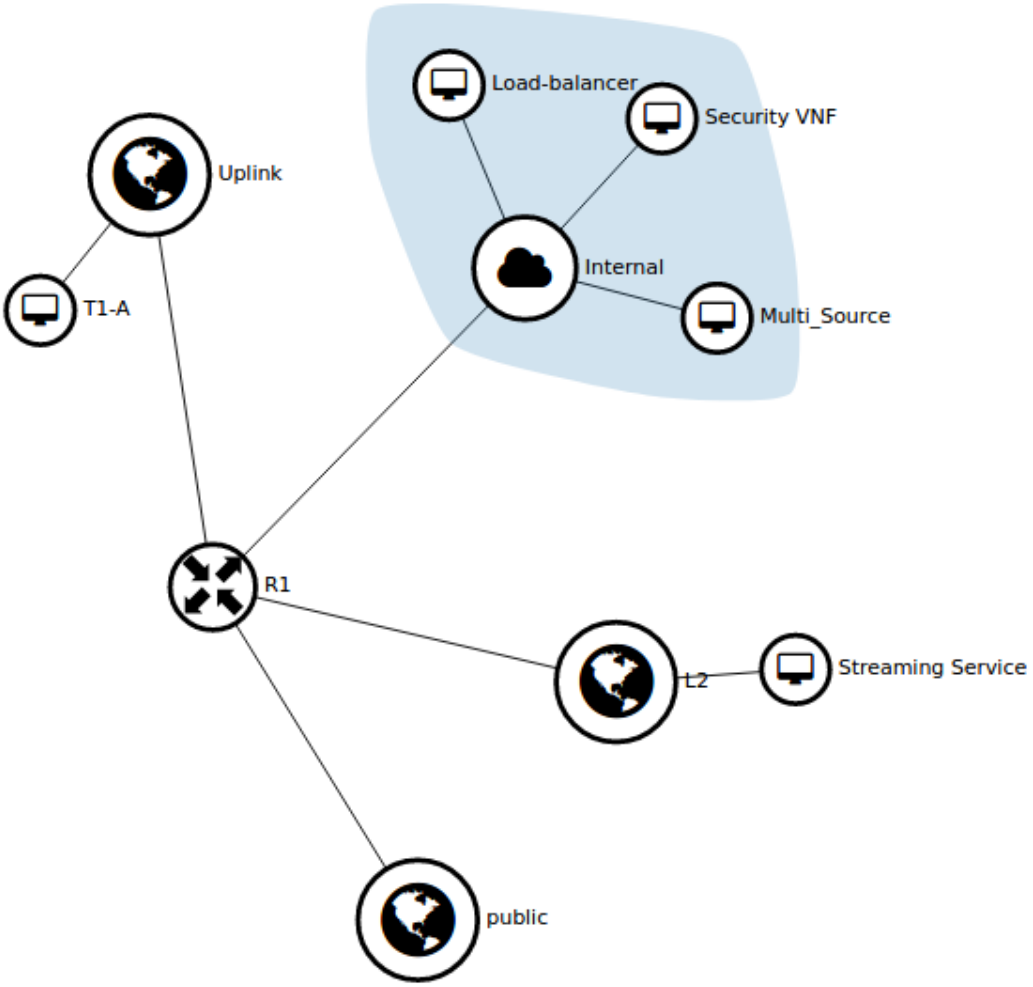

Figure 3-4 snapshot of the network topology at the OpenStack dashboard

Figure 3-5 provides another realization of the deployed networks over the OpenStack platform. The virtual router is utilised to route the traffic amongst the various connected subnets, internal and external. R1 is OvS based router that is configured with external gateway towards the Internet and three logical networks namely, Uplink, L2 and internal. Uplink is mapped to physical port eno4 via the br-ex3 external bridge.

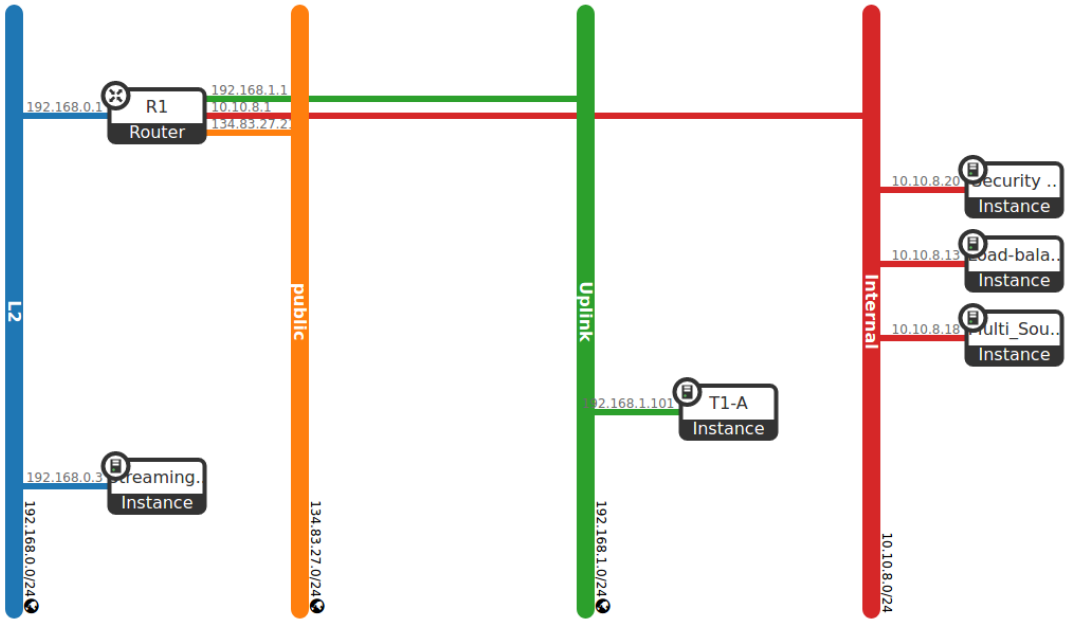

Figure 3-5 snapshot of the deployed networks (OpenStack Dashboard)

3.4.2 Network functions virtualization

Network functions virtualization (NFV) is a network architecture concept, which refers to the process of implementing and deploying various network services by using virtual network appliances. There are unlimited range of services and applications that could be implemented in this paradigm e.g. security solutions, caching services, WAN accelerators, firewalls, and more. The use of orchestrators made this solution more appealing for network administrators, by facilitating the deployment of the solutions including all the required resources with few clicks by using orchestration templates. The only requirement is to configure the templates with the required specifications for service deployment and the rest is taken care of internally. Later subsection refers to some of the developed services for IoRL solution, while the next subsection provides an example of the traffic route through the OpenStack platform.

3.4.3 Traffic routing within OpenStack

In order to enable traffic routing traversing properly from the VNFs to the external networks, OpenStack utilizes OvS to interconnect the networks together. Figure 3-3, depicts the network architecture including the integration and the external OvS bridges within OpenStack, namely br-int, br-ex, br-ex2 and br-ex3. br-ex2 is the external bridge that is mapped to the eno3 physical port on the physical host. The command line interface enables the cloud manager to see the details of the

deployed networks by the use of the appropriate commands. Figure 3-6, provides the details of br-ex2, including all the available ports of this bridge.

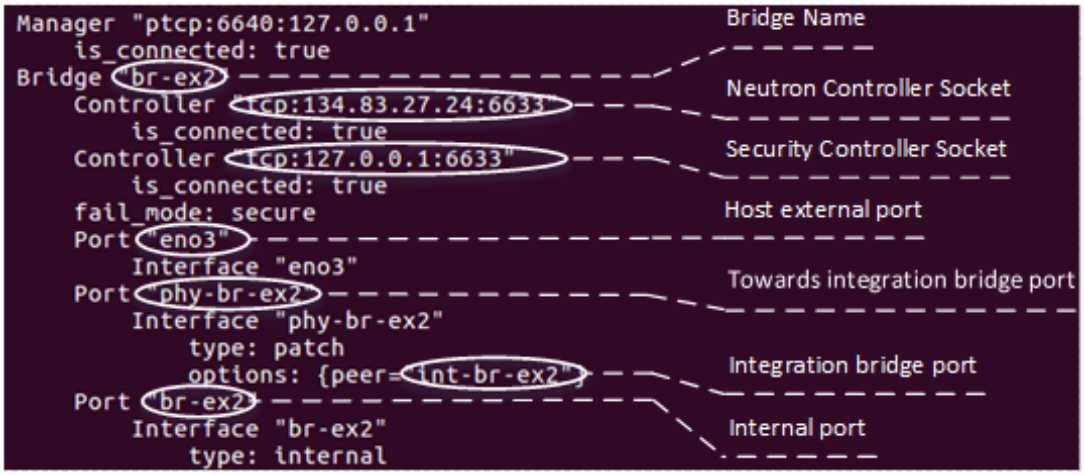

Figure 3-6 details of one of the external bridges within OpenStack platform

The traffic coming from the external network, reaches to the external physical port eno3. Being one of the ports of “br-ex2”, “eno3” enables the traffic to get to the external bridge “br-ex2”. Another port of “br-ex2” bridge is “phy-br-ex2”, which is an internal port that has a peer port within “br-int” bridge.

Figure 3-7 shows a snapshot of the “br-int” bridge, where one of the attached ports is the highlighted “int-br-ex2”, which is the peer port of the “phy-br-ex2” port from the “br-ex2” bridge. At this point the traffic gets to the integration bridge and virtual router. According to the SDN routing, either there is flow rule to deal with the incoming traffic or the OvS asks the SDN controller for the instructions about such traffic, which will install flow rule accordingly.

```
Bridge br-int
        Controller "tcp:127.0.0.1:6633"
            is_connected: true
        fail_mode: secure
        Port "qr-e45c4a68-2b"
            tag: 1
            Interface "qr-e45c4a68-2b"
                type: internal
        Port "int-br-ex4"
            Interface "int-br-ex4"
                type: patch
                options: {peer="phy-br-ex4"}
        Port "qvo5838fbde-56"
            tag: 2
            Interface "qvo5838fbde-56"
        Port int-br-ex
            Interface int-br-ex
                type: patch
                options: {peer=phy-br-ex}
        Port "qr-064e7fe4-e4"
            tag: 2
            Interface "qr-064e7fe4-e4"
                type: internal
        Port "int-br-ex3"
            Interface "int-br-ex3"
                type: patch
                options: {peer="phy-br-ex3"}
        Port "gre0"
            Interface "gre0"
                type: gre
                options: {remote_ip="192.168.0.25"}
        Port "tape570919b-b5"
            tag: 3
            Interface "tape570919b-b5"
                type: internal
        Port "int-br-ex2"
            Interface "int-br-ex2"
                type: patch
                options: {peer="phy-br-ex2"}
        Port "qvocf59f9ad-6f"
            tag: 1
            Interface "qvocf59f9ad-6f"
```

Figure 3-7 snapshot for br-int Bridge

3.5 OpenStack platform deployment: command-line and web-based instructions

The deployment of new VNF requires multiple steps in the process of constructing the required services. This subsection follows the steps towards establishing a hypothetical service in the form of VNF manually. It is worth mentioning that no orchestrator was involved in the process, otherwise, the steps would be different and automated. Each step is achievable via either Command Line Interface (CLI) or

Web-based Interface (Horizon), therefore some of the steps are presented with Horizon interface and others with CLI.

3.5.1 Creating logical networks

Assuming the OpenStack platform has just been created, so there are no networks available for usage; therefore, creating proper networking infrastructure is a vital step. Figure 3-8 shows the OpenStack dashboard in the process of creating logical networks. Most of the tabs are self-explanatory. Although Horizon dashboard is available for both, cloud admin and tenant admin, there are some differences in the available options for each, for instance the “Shared” and “External network” tick boxes are only available for cloud admin, as it is his responsibility to configure the external network properly and map it to the physical network of the physical host. Figure 3-8 also shows the next step in the process of configuring the subnet of the created network.

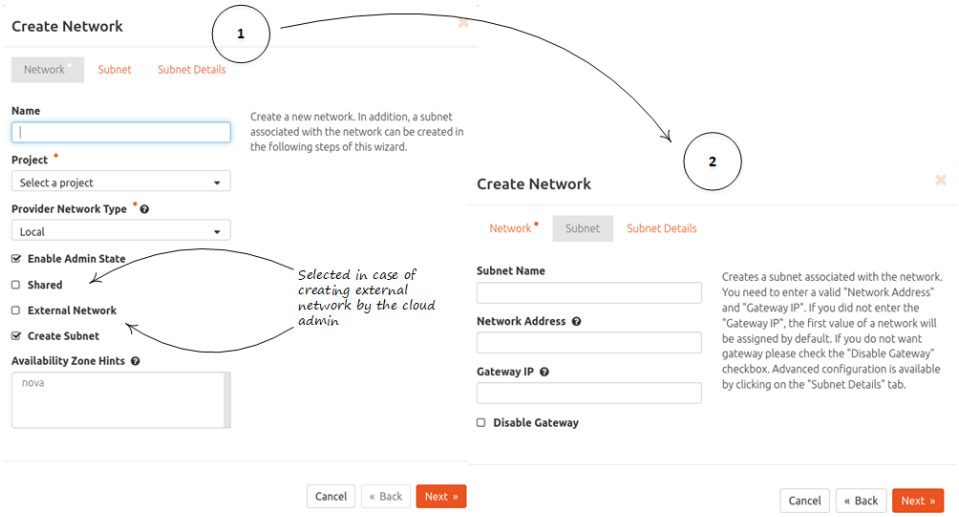

Figure 3-8 snapshots of the OpenStack dashboard: creating network

After finishing the configuration of the networks, Figure 3-9 shows the output of the available network information using Horizon dashboard and CLI interface.

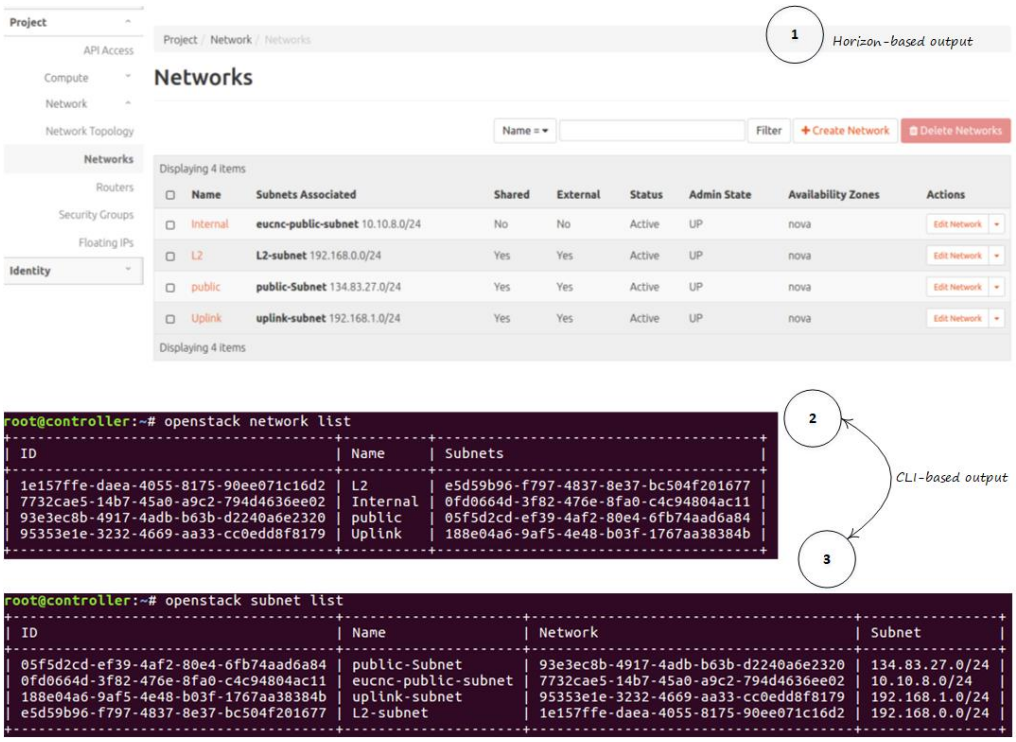

Figure 3-9 snapshots of the available network: 1 web-based, 2 and 3 LCI based view

As shown in the Figure 3-9, web-based view is more polished and convenient, while CLI-based is more detailed and less attractive to ordinary users, since it is dedicated for network administrators to provide them with useful details in the network's troubleshooting procedures.

3.5.2 Deploying instances

The deployment of virtual instances requires some preliminary preparation and configuration for the virtualized platform e.g. creating proper networking infrastructure, configuring suitable flavours (buckets of resource parameters), preparing the required images and so on. Flavour creation is based on two aspects, first the required values of parameters in the instance to perform specific task, and second, the available resources in the pool of resources. There are two types of Images, either plain OS image that is downloadable directly from an image repository with no modifications, or preconfigured image, which is ready to be deployed with no further configuration to perform the required tasks. Instances deployed with the former type of images might require software installation after deployment to customize its performance as required, while instances deployed with latter type of images becomes ready to perform the required tasks as soon as they become live. Figure 3-10 and Figure 3-11 depict the results of querying the list of available flavours and images respectively. There is an equivalent method in the Horizon dashboard.

```
root@controller:~# openstack flavor list
+--------------------------------------+------------+-------+------+-----------+-------+-----------+
| ID                                   | Name       |   RAM | Disk | Ephemeral | VCPUs | Is Public |
+--------------------------------------+------------+-------+------+-----------+-------+-----------+
| 15ad92fd-f36e-4fc6-b05d-9f136070c05a | Large      |  1024 |   65 |         0 |     4 | True      |
| 6                                    | web-flavor |   128 |    1 |         0 |     1 | True      |
| 927a00f4-3333-4f0e-9988-62a9b17065af | medium     |  1024 |   50 |         0 |     4 | True      |
| 95bc2ff4-0892-48f5-af20-f26136afb50e | Large+     |  2065 |   80 |         0 |     6 | True      |
| aa012cf0-96f3-4c28-a496-c655b5011dbe | moderate   |  2400 |   24 |         0 |     2 | True      |
| cdcd833d-485b-48e6-8208-ad880f7bc378 | tiny       |   512 |    4 |         0 |     1 | True      |
| ef06bd15-83c2-4e42-95e3-158f1e85899f | Large++    | 16000 |   80 |         0 |     6 | True      |
+--------------------------------------+------------+-------+------+-----------+-------+-----------+
```

Figure 3-10 snapshot of the available flavours

As shown in the figure, various flavours were preconfigured based on the required resources for each service.

```
root@controller:~# openstack image list
+--------------------------------------+------------------+--------+
| ID                                   | Name             | Status |
+--------------------------------------+------------------+--------+
| 60c18ba1-5925-4543-9b68-21ca57de918e | DB Final         | active |
| 0972f5cb-141e-4689-9bc5-4c5e0cc72dd0 | Database         | active |
| dd7bacb3-d716-41d0-9d2d-2963722d331a | Database Final   | active |
| a2ed34cf-c18d-4d4f-9d0e-34e8d1ff2393 | Database UB      | active |
| 374c0851-921f-455c-932b-219556921239 | Database Ubuntu  | active |
| 1e07a7a6-682e-481e-9b41-5a01b4ae48d4 | ISA              | active |
| 1f43ed92-719d-4223-b48b-ba150ab0da0b | LB               | active |
| c0d515d0-9c99-4cdf-9357-0d4496d3e26c | Load-Balancer    | active |
| 9472edb6-7b56-4414-baf6-189bc73c6969 | MSS              | active |
| 2a079efc-070b-4939-9ef7-1075a57ff94a | Security VNF     | active |
| ede4ee21-8e12-4618-8cd0-c1bcd21698b2 | StreamTV         | active |
| bc993867-bc1a-4a47-8f6b-02f09760d4b7 | cirros           | active |
+--------------------------------------+------------------+--------+
```

Figure 3-11 snapshot of the available images

Figure 3-11 shows the available images in the current OpenStack platform, some of which were downloaded directly from the image repository and others where downloaded after being created and tested in different platforms to ensure readiness state of the service. One of the images was the “cirros” image, which is a lightweight image with basic functionality that is normally used for testing purposes.

After the preparation of the infrastructure, a cloud administrator is capable of deploying new VMs by following the prompted steps. Figure 3-12 shows these steps, some are mandatory, while other are optional. The upper right corner of the figure shows the maximum number of deployable instances by the tenant, this number is configured by the cloud admin based on the “service provider- tenant” agreement.

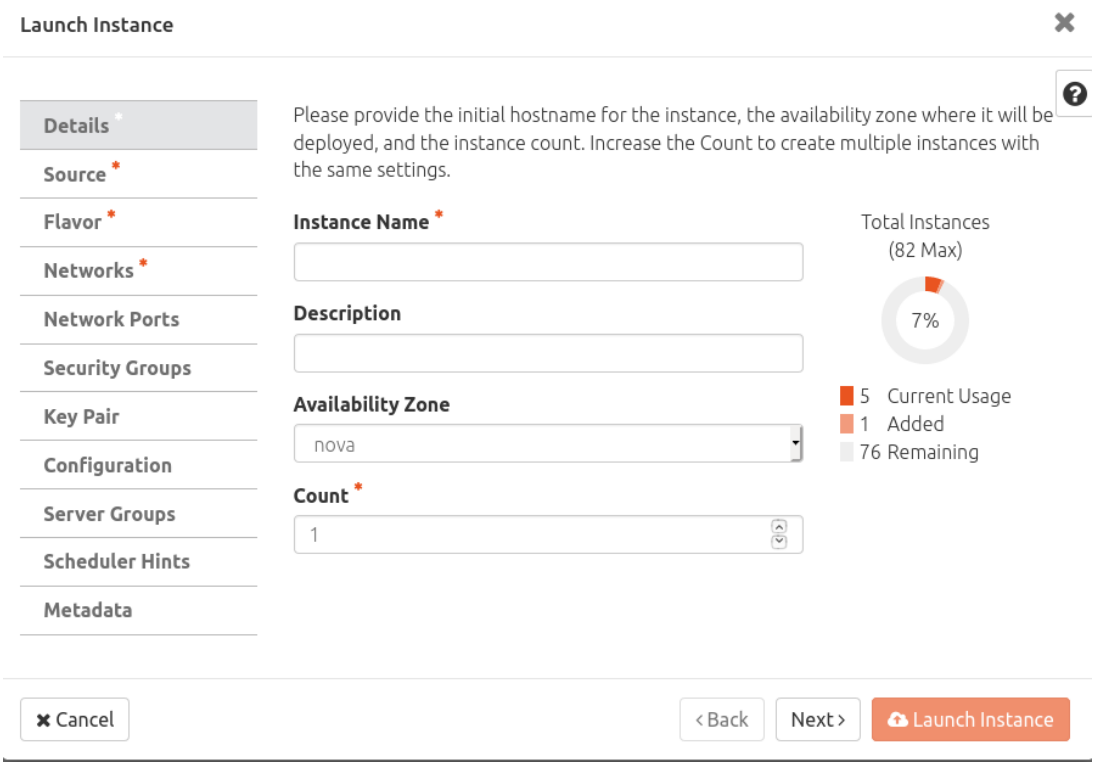

Figure 3-12 snapshot for the Horizon dashboard (instance launching)

After successful deployment of VMs, they appear in the Horizon dashboard or can be queried using CLI. Figure 3-13 shows the available instance using both methods.

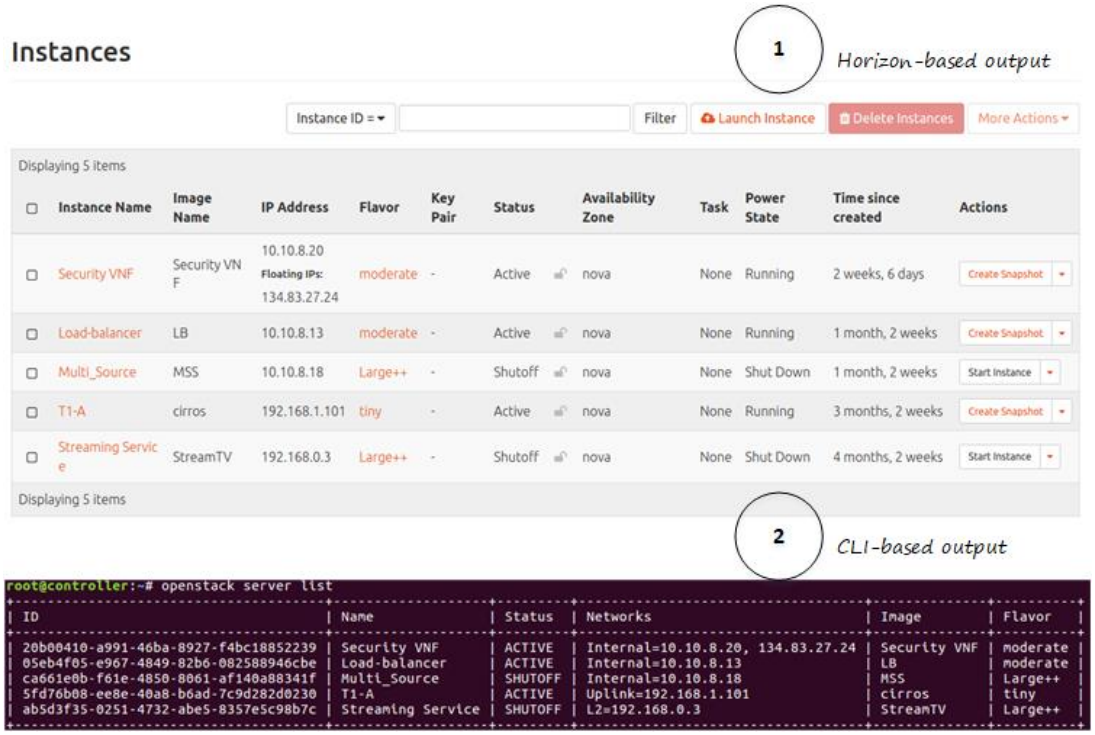

Figure 3-13 available instances: 1 Horizon-based view, 2 CLI-based view

3.5.3 Platform troubleshooting

The OpenStack platform provides a complete log register for the all the services, thereby enables the cloud admin to troubleshoot the system efficiently. Based on the type of error, cloud admin can go into the service directory and check the logs for each component within that service. Figure 3-14 shows an example of the available log files for the neutron service.

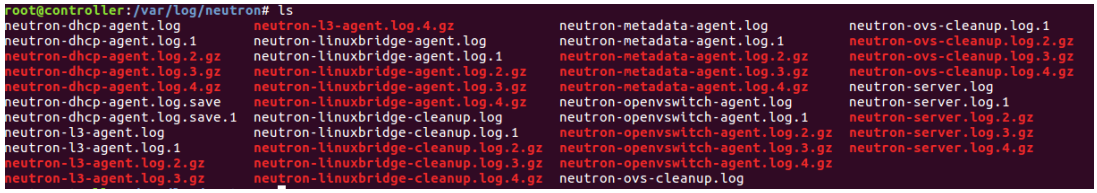

Figure 3-14 snapshot of the log files within neutron service

There are similar folders for the other services, and within each service the system logs all the information related to system performance including, instance creation, spawning, deletion, network creation, network deletion, and so on. Successful cloud admin keeps track of the log files and is capable of finding out the solutions after determining the exact error based on the log files information.

3.6 Virtualised Network Functions

This section provides an overview about few of the IHIPGW deployed services, which are in the form of VNFs, providing brief description about each service along with available features and integration procedures.

3.6.1 Multi-Source Streaming

Multi-Source streaming (MSS) VNF is a virtualised service exploited to add reliability at the application level for the system, by streaming sub-flows of video data from different sources or through different paths. Those sub-flows can be read independently, giving a lower video quality, or can be merged, giving a higher video quality. The main use case for this service is when VLC link gets interrupted by blockage or one of the three available links in usage becomes unavailable or quality degrades.

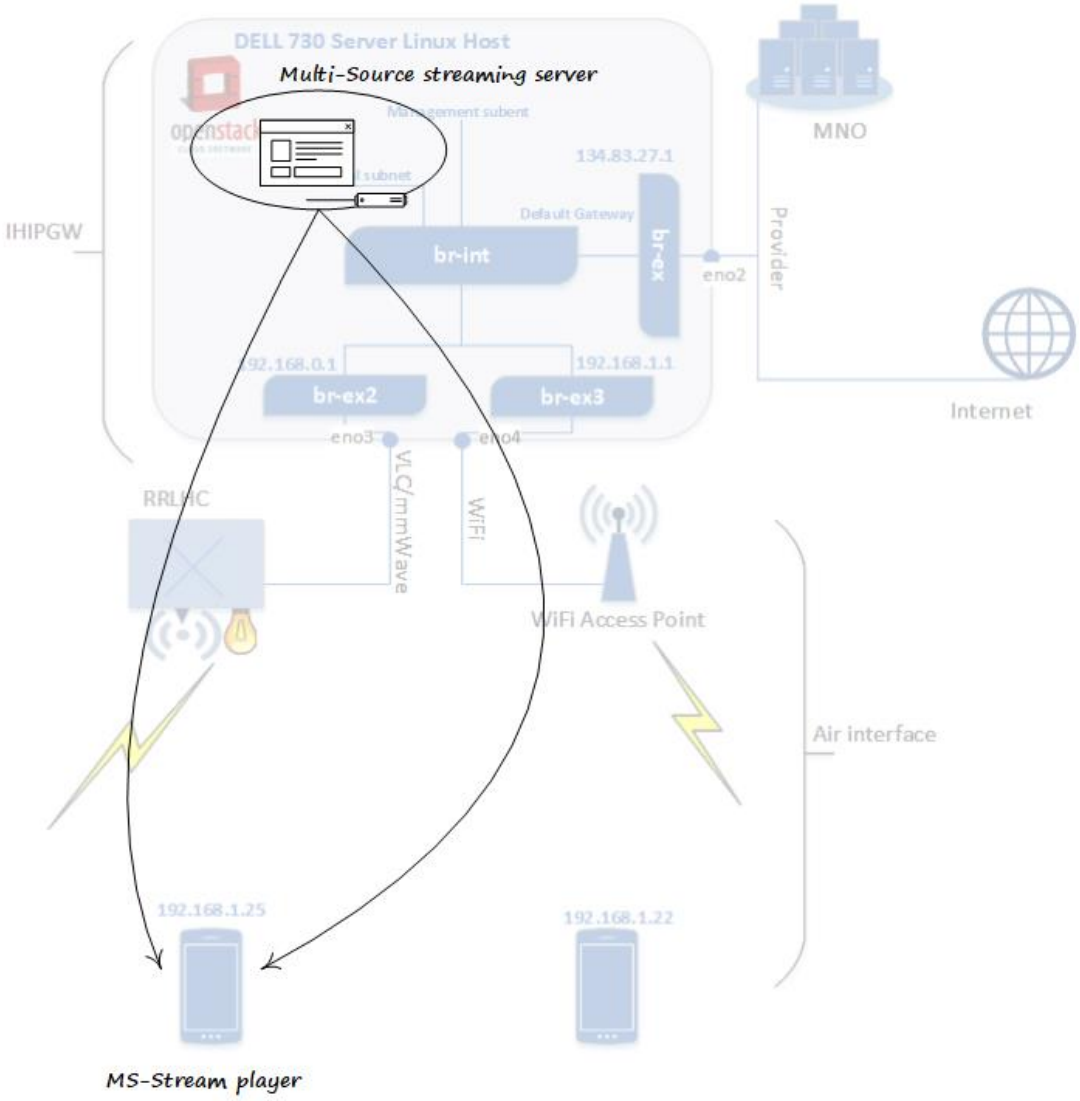

Figure 3-15 multi-Source streaming service

Figure 3-15 depicts the MSS Service within the IHIPGW platform. Technically MSS comprises three components namely, MS-streaming server, MS-stream transcoder and MS-stream player. Both of the MS-streaming server and MS-stream transcoder are at the IHIPGW side, while the MS-stream player is at the end user side.

On the client side, the MS-Stream Player is a video player that can be integrated in a web page and accessed through a web browser. This client may also be running in a native application for specific UE. As the algorithm of MS-Stream is defined as client-centric, the MS-Stream Client is responsible for the creation of the HTTP requests sent to the MS-Stream Server for Multiple-Source Streaming sessions. The adaptation algorithm is implemented in the MS-Stream client, which is capable of requesting the proper quality depending on the previously downloaded speed of each link.

On the server side, the MS-Stream Server is an application that can answer client requests for specific video contents. This module is responsible for the creation of video segments adapted to Multiple-Source Streaming. The video segments are created from video data transcoded in numerous different qualities.
In other use cases, like live streaming from a camera scenario, the video data are streamed from the cameras through the SDN of the Home IP Gateway and can be redirected to the Transcoder. At this point, the stream is transcoded into several levels of lower qualities and sent to the MS-Stream Server. Then, the live client would be able to get the Multiple-Source-adapted live video data through available links towards the client.

3.6.1.1 MSS features

MSS offers several features for its clients as follows:

- Transcode video files and/or live streams from IP camera via Real Time Streaming Protocol (RTSP).
- Create streaming sessions for Video on Demand (VoD) and live streaming.
- Serve 360° videos to a browser.
- Serve MS-Stream video segments from multiple-networks

3.6.1.2 MSS optimal resource requirement

In order to integrate MSS service with IHIPGW platform properly, there is an optimal resource requirement.

- vCPU: (3-4) virtual CPUs due to transcoding high processing consumption.
- vMemory: 3 GB or higher.
- vStorage: (20-30) GB internal or external based on the deployment options.
- vNetwork: IP address reachable by MS stream client and other video sources, with connectivity to external storage.

3.6.2 Load Balancer

Load Balancer is instantiated as a VNF on IHIPGW platform. End users are connected to the IHIPGW via both links namely, VLC/mmWave and Wi-Fi Access Networks simultaneously. Traffic traverse via both links using as default-gateway of the Virtual L3 Router inside OpenStack, as shown in Figure 3-16 . The default option is to route traffic through the VLC/mmWave Access Network. The Load Balancer uses the RYU SDN controller in order to monitor the incoming/outgoing traffic at the interface connected to the VLC/mmWave Access Network. Depending on the traffic and the defined upper and lower bitrate limits, the Load Balancer can take one of the following decisions:

Route the traffic of a user to the Wi-Fi Access Network, if the monitored traffic is above the upper bitrate limit.

Reroute the traffic of a user that was connected to the Wi-Fi back to the VLC/mmWave Access Network, if the monitored traffic is below the lower bitrate limit.

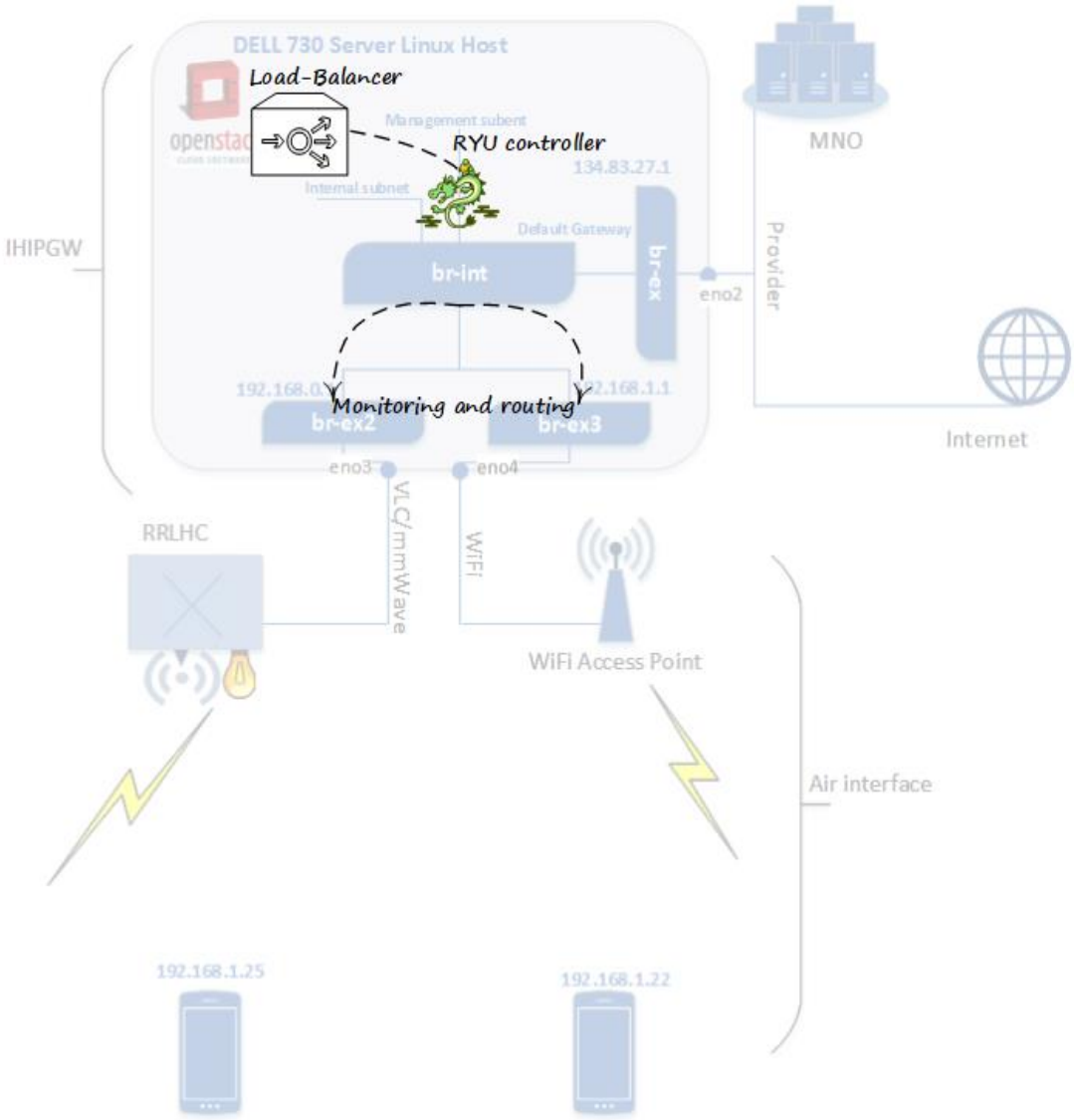

Figure 3-16 load balancer service within IHIPGW

The resources requirement for the proper deployment is two virtual CPUs, two Giga Bytes of virtual RAM and twenty Giga Bytes of root disk.

The load balancer VNF is tested in Brunel to provide proof of concept by deploying the load balancer VNF over the OpenStack platform and enable it to manage traffic going over the two external bridges “br-ex2” and “br-ex3”. “br-ex2” was linked to an ethernet link and attached to the end user machine, as shown in Figure 3-17. End user was able to communicate via WLAN link via “br-ex3” bridge. The test started by initiating a downlink traffic stream from the network side (OpenStack

platform) towards the end-user. Traffic was using the higher priority link "Ethernet" resembling a VLC link, then after some point the link were physically disabled to emulate traffic congestion or link failure, therefore, load balancing VNF routed the traffic via the WLAN link to the end user to enable service continuity.

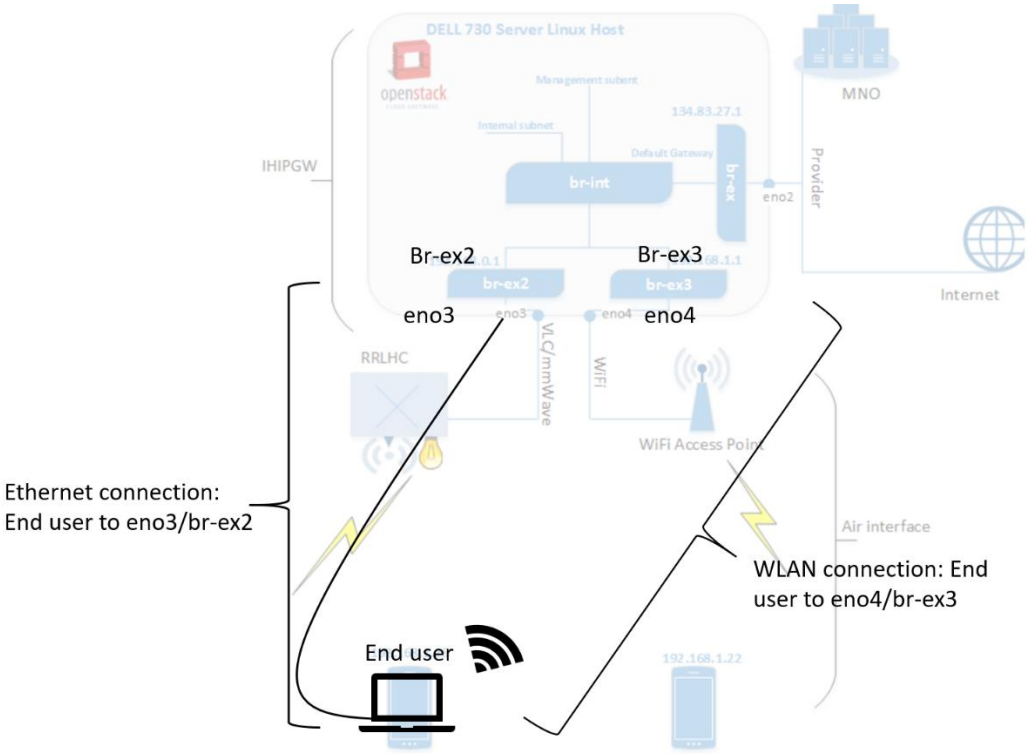

Figure 3-17 Description of deployment at Brunel University

3.6.3 Security Suite

The VNF of the security suite is developed to detect and stop various types of attacks on the IHIPGW platform. The VNF exploits Ryu controller to monitor the network and enables the controller to intercept and block malicious traffic. The security suite is equipped with GUI dashboard to enable cloud admin to view security issues, enable/disable security modules and modify configurations.

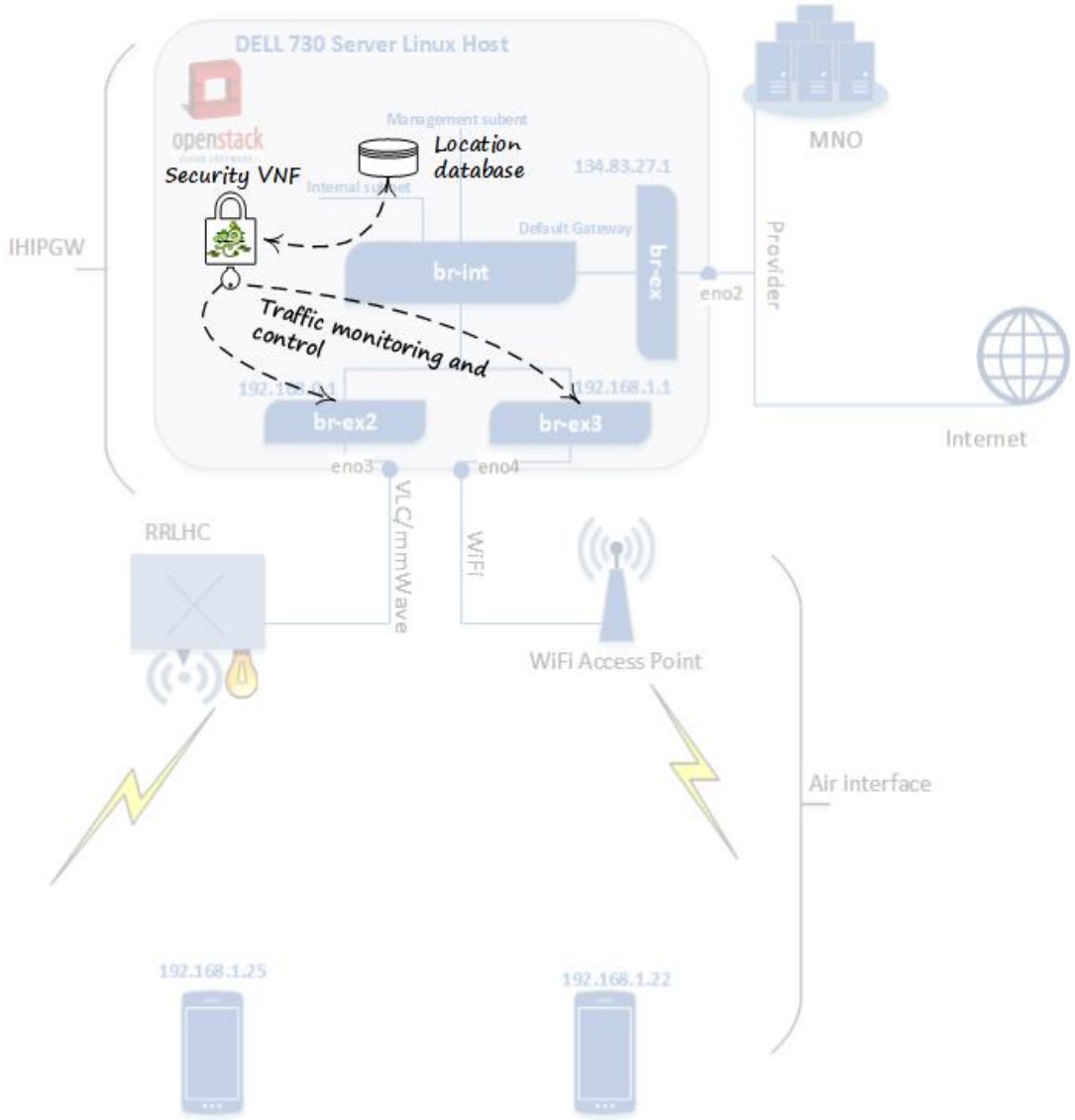

Figure 3-18 security suite VNF with IHIPGW platform

Figure 3-18 depicts the security suite VNF, which includes a Ryu controller that is connected to the both of the external bridges towards the access network, for traffic monitoring and control. In addition, it interacts with local location database for accurately locating the attacker within the coverage area.

3.7 Summary

This chapter briefly described OpenStack platform as the virtual infrastructure enabler and manager, with its available services that enable its performance and features. Then an actual deployment of the IoRL mobile edge cloud was presented,q with an overview about the steps followed to deploy the developed services to the existing virtualized platform. Finally, quick view was given about some of the

deployed services namely, Multi-Source streaming, Load Balancing and Security suite.

4 Geo-location Multimedia services for 5G indoor coverage network

4.1 Introduction

Next generation mobile networks are being designed to offer new features for their clients (UEs) and to host more services to improve UEs' QoE. In this regard, the aim of this chapter, is to present concept, implementation and the components of two of the next generation services so-called: Follow Me Service (FMS) and Multicast Sharing Service (MSS). FMS is a service offered by 5G gNB to UEs in indoor environments (e.g. home), which enables its clients to use their smart phones to select media content from content servers, then cast it on the nearest TV set (e.g. living room) and continue watching on the next TV set (e.g. kitchen) while moving around the indoor coverage area. Similarly, MSS is another service offered to UEs in indoor environments (e.g. Museum); it enables its clients to use their smart phones in a server-client mode. Server (or host) selects media content from content servers, then casts it to a group of registered clients based on predefined criteria (subscription-based or relative proximity-based). FMS and MSS have their own front-end applications that utilise signaling protocols such as Session Initiation Protocol (SIP) to enable successful session initiations and management.
MSS does not rely on UEs smartphone capabilities, rather on network capabilities. Both services can be provisioned by utilising UEs geolocation information, robust switching mechanism between multiple 5G Radio Access Technology (RAT) and relying on the intelligence of the SDN/NFV Intelligent Home IP Gateway of the Internet of Radio Light (IoRL) project paradigm. By the time of proposing the services, the system was not fully implemented and ready for services testing, therefore, I stepped forward to use Mininet platform to provide performance monitoring with measurements for the service. Simulation results show the effectiveness of our proposal under various use case scenarios by means of eliminating the packet loss rate and improving QoE of the museum users.

4.2 Overview and related work

Smart Televisions (TVs) are internet-connected appliances that offer numerous online features, such as an on-demand video content, internet-based services, etc.

However, using traditional remote controllers for browsing the available media contents on smart TVs is not as convenient as using smartphones, since the latter rely on contemporary screen mirroring via casting mechanisms, which enhances the remote controlling experience. Yet, an issue arises when using smartphones as remote controllers, which is the requirement for fine tuning between the smartphone and the TV set prior to sharing media content. Most phone manufacturers have developed methods for screen mirroring between smartphones and smart TVs such as Mirror Share, File Share, HTC Connect, Miracast, etc. [55][56]. Unfortunately, these remote mirroring developments are numerous and do not share a common approach, which makes it difficult to evolve towards a holistic approach that can serve a wide range of different mirroring and sharing services. The proposed solution provides higher data rates and offers new features to enable direct sharing capabilities to multiple smart TVs, using generic smartphones as smart remote controllers without the need for compatibility and a priori pairing processes.

As explained in previous sections, the Internet of Radio Light (IoRL) project is an indoor solution that represents an evolutionary system architecture that utilizes new access technologies namely, millimeter Wave (mmWave), and VLC with IHIPGW built based on SDN/ Network Function Virtualisation (NFV) network paradigm. Furthermore it is compatible with existing system/standards, e.g., IEEE 802.11, 802.15, 3GPP [9].

Our intention is to exploit the intelligence offered by the synergy between SDN and NFV technologies, to design and develop new smart TV solution with improved QoE. This solution enables a user to search content on-demand via smartphone, select media content to be viewed on a TV and leave the rest to be taken care of by the intelligence of the IoRL network. The required task is to enable users to continue viewing media content while moving among multiple TVs within the home. This is accomplished by using a users' location information to allow the SDN to perform smart switching of packet flows to multiple nearby TV devices seamlessly. FMS introduces two main features. Firstly, it enables the registered UEs to enjoy watching video content on the nearest smart TVs, whilst having the flexibility to continue watching the content while moving around the coverage premises. This process is managed without going through pairing and restarting the videos each time they change their location around the house, which improves the QoE of the home users. Secondly, it enables the UE to perform Video Multicasting (MC) within the home area. Selecting the MC feature within the FMS,

configures the network to forward the video stream to all the available TV sets within the home environment, thereby performing video multicasting. Note that FMS features do not require any other action from the UEs, except to keep carrying their smartphones while moving around the home area. On the other hand, MSS enables UEs to share media contents with other UEs within the same IoRL coverage area, on one of two bases: location proximity or subscription based. Similar to FMS, MSS also exploits the network infrastructure rather UEs smart phone capabilities to deliver and share media intelligently.

Moreover, the use of virtualization technology in the network infrastructure facilitates incorporating proxy servers, location servers, location database, etc. in form of VNFs, which contributes to flexible and agile system deployment, as well as reduces traffic delays and latency.

4.3 Technical overview

This section briefly describes the main technologies that are considered the enablers and the main building blocks of the proposed services over the IoRL network platform namely, proxies, SDN, and NFV.

4.3.1 Proxies

The concept of the Proxy in mobile networks is to intercept and split the connection between the UE and the external content server, therefore seamlessly becoming the content server to the UE and a client to the content server. One type of proxy server is the TCP split connection proxy, where this type of server creates two separate asynchronous connections on both sides of the server namely, (upstream) from the proxy server towards content server, and (downstream) from proxy server downwards the UE. This separation provides the network operator with flexibility in the traffic management to achieve an enhancement to the system performance.

Depending on the network deployment, and user requirements, the use of proxies becomes beneficial by reducing the Round-Trip Time (RTT) for the data request by creating two separate asynchronous sessions, enabling data buffering at the proxy. This kind of content management promotes higher downstream speed supported by the high bandwidth available for the UEs. The use of such proxies provides complete control of the whole transport segment between the client devices and the proxy [57], which enables fine-tuning of the connection link. For example, connections between the TV set and the proxy can use advanced protocol features, which cannot be easily deployed in a public network due to potential third-party

interference. Therefore, specially customized data routing patterns and protocol sessions can be established for providing the FMS for the IoRL clients.

4.3.2 Software Defined Networking (SDN)

SDN is an emerging networking technology, designed to eliminate network complexity as well as the static nature of traditional distributed network architectures. SDN enables the separation between control plane and data plane, thereby centralizing the network intelligence. In addition, SDN offers network abstraction layer from the underlying network infrastructure layers, as shown in Figure 4-1.

SDN architecture enables network automation by the aid of the orchestration mechanism. Network management layer along with network applications are interoperated within the orchestration level to add autonomous capability to the network, by which network administrator can configure the system to spawn ready to deploy network solutions, in response to actual load, with no intermediate action by humans. [58] The SDN controller is a logically centralized network entity, it offers performance and fault management, which involves configuration and management of the SDN complaint devices. SDN controller is capable of configuring the network entities based on the desired requirements, such as QoS level and performing link management between devices. SDN data plane layer refers to the traffic-forwarding layer. While control plane entails the network intelligence. The control plane is centralized, which promotes faster application deployment and delivery, in addition to reducing the cost of network operations via policy-enabled workflow automation. Many industrial bodies have embraced SDN technology e.g. Google, Microsoft, Yahoo and Facebook because of the characteristics offered by SDN. Here are some of the main advantages of SDN:

- Establishing a central control unit within multi-vendor environment: contrary to the traditional networking systems, SDN enables network administrators to use unified protocol to configure all compliant networking devices from various vendors. This approach was unachievable with traditional networking due to the existence of vender specific interfaces and management protocols, which made the process of network configuration more expensive and complex [58].
- Network automation and reduced complexity: the use of SDN controllers enables network admins to manage the network flexibly, by exploiting the available management tools, thereby eliminating the need for manual

management and reduce the margin of human-induced errors. This approach leads to reduced network OPEX. Especially with cloud-based solutions, SDN enables self-provisioning and intelligent orchestration that helps to reduce the operation overhead and increase the business agility.

- Improved network innovation: the centralized control platform enables the network admins to program and reprogram the entire network as required in relatively short period to achieve faster business innovation, thereby meeting user expectations. Network admins are enabled to exploit the SDN and virtualization technologies to deploy new service within few hours' time [59].
- Higher network security: network admins can exploit the SDN networking to achieve faster and more secure network configuration. In traditional networking, changing network configuration requires the network admins to (re) perform the tasks on the networking devices individually, which increase the risk of network instability, while SDN platform eliminates this risk by centralized procedures that maintains higher network security and stability [59].
- Granular level of network control: SDN paradigm utilizes open source protocols e.g. OpenFlow (OF). OF provides granular level of policy control, e.g. session, user, device, and application levels. With this granularity along with the network abstraction and automation, cloud operators are able to coexist multiple tenants on a shred resources platform.
- Enhanced user experience: The Centralized control plane offers a unified interface to manage the various network components efficiently, which eliminates the interoperability issue due to various vendor interfaces. This advantage is reflected on the network performance, thereby enabling the new services that enhances end user QoE.

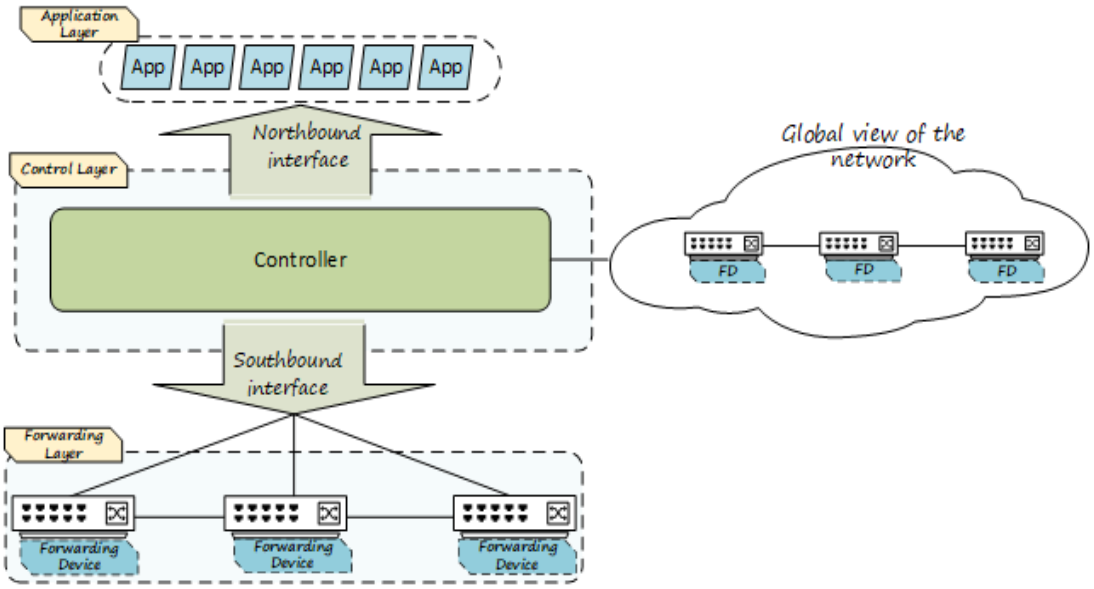

Figure 4-1 software Defined Networking Architecture

4.4 System architecture

IoRL is a 5G gNB proposed as an indoor solution for improving the mobile networks indoor coverage as well as UE's QoE. IoRL architecture comprises of a radio access network, UE and IHIPGW. These are structured in three layers which are, the service/application layer, NFV layer/SDN layer, and the access layer. The system architecture is shown in Figure 4-2.

The IoRL approach includes designing and controlling a radio-light communication system that combines WLAN, mmWave and VLC access points, to offer access network for the users. At the same time, IoRL enables deploying new services on top of its architecture. These services aim to improve UE's QoE by exploiting the SDN intelligence and NFV flexibility. OpenStack platform is used to develop the virtualized SDN/NFV network architecture [60][61], which enables flexible deployment of cloud computing for controlling large pools of compute, storage, network resources.

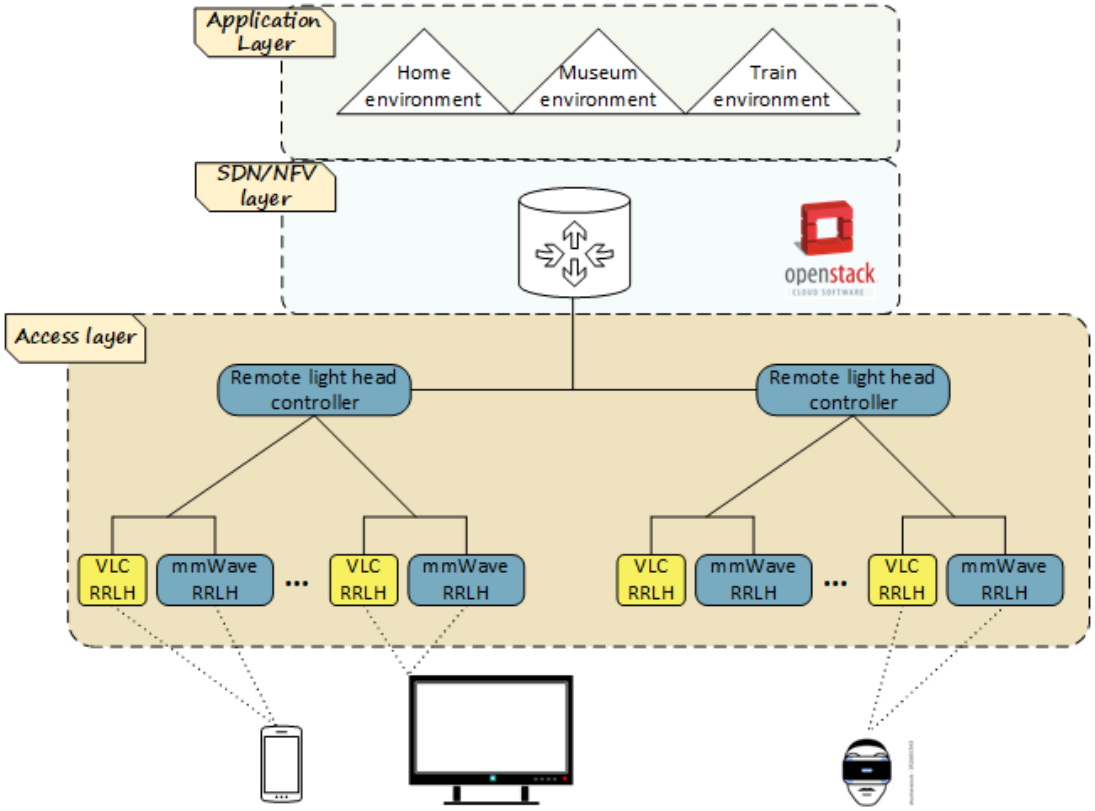

Figure 4-2 IoRL Layered Architecture

4.4.1 IoRL Radio Access Network

The access network represents the access layer of the IoRL architecture. It comprises of Remote Radio Light Head (RRLH) controllers installed throughout the indoor environment premises. RRLH controllers contain mmWave and VLC modules together, and they are connected via 10Gbps Ethernet ring.

Each RRLH Controller drives up to eight VLC and mmWave modules with the same transmission block sub-frame, in order to provide full coverage of the building premises.

4.4.2 IoRL User Equipment UE

UE integrates mmWave and VLC antennas within its design. mmWave antennas enable data duplex transmission, while VLC antennas enable data reception only.

For FMS, the UE's smartphone acts as a remote controller, as well as a client location indicator as will be explained in FMS operation section later.

4.4.3 Intelligent Home IP Gateway IHIPGW

The IHIPGW hosts the intelligence of the IoRL system, since it is a contemporary SDN/NFV environment, which enables flexible and cost-efficient deployment

scheme. SDN offers a centralized control plane by logically separating the control plane from the underlying forwarding plane, which enables the IHIPGW to get a global view of the network, and manage the routing of the data and enables services based on this data, as depicted in Figure 4-3. The controller configures the forwarding device to route the traffic amongst multiple destinations that includes, IoRL Radio Access Network (RAN), MNO Evolved Packet Core (EPC), local internet breakout, or local applications running on multicore Cloud Home Data Centre Server (CHDCS).

On the other hand, NFV facilitates the implementation of network functions as a software applications running on a non-dedicated hardware in the form of VNF. This design increases system flexibility while reducing CAPEX and OPEX. Meanwhile, the virtualized environment facilitates network service developers to create VNFs for location sensing, multiple-source streaming and security monitoring.

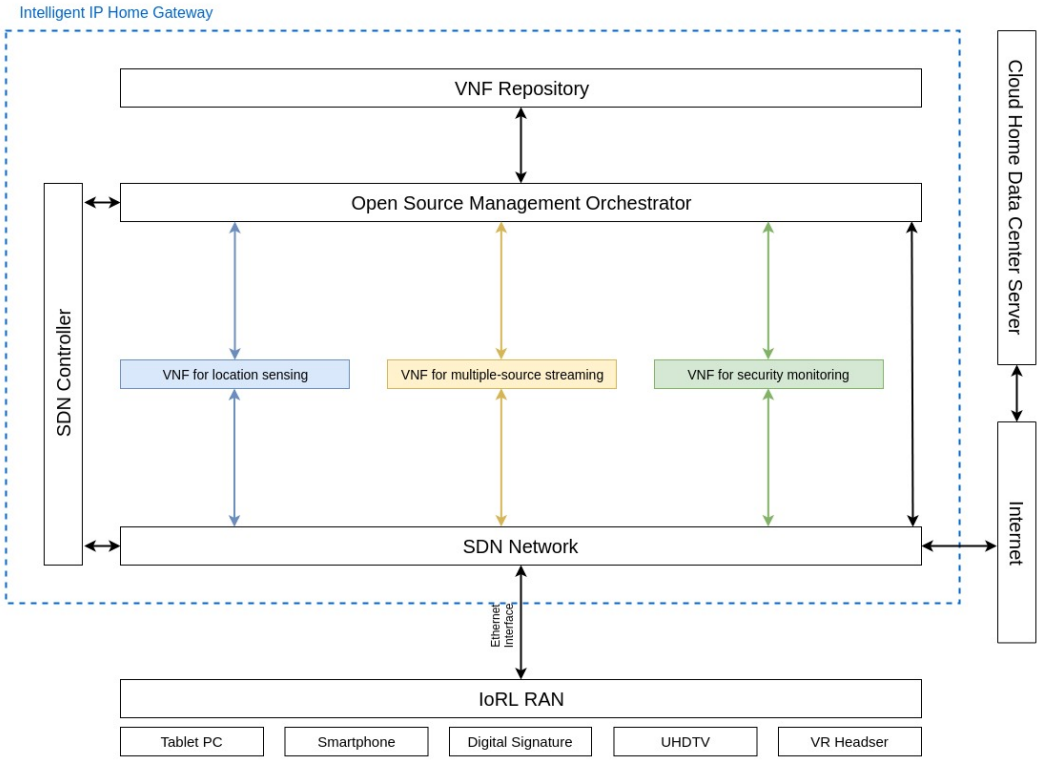

Figure 4-3 illustration of the IoRL Home Network with IHIPGW deployment of the identified VNF modelling

4.4.4 Location service

In IoRL project, the system relies on VLC and mmWave positioning. It targets a high positioning accuracy of sub 10cm.

This can be provided by combining both VLC and mmWave techniques. The main goal of position sensing in IoRL system is to support location-based data access, monitoring, guiding and interactive applications. The main components of the location sensing architecture are location server, location database, location service client and RRLH controller.

The mmWave positioning system uses electromagnetic waves to determine the location of UEs. The RRLH controller is in charge of measuring time differences of arrival in the uplink. The mmWave measurements are reported by the RRLH controller to the radio resource controller, which communicates them to SDN in the form of packets using Packet Data Convergence Protocol (PDCP) protocol. SDN transfers the mmWave measurements to the location database.

The positioning system based on VLC, uses visible light signals for determining the positioning of UEs. The signals are transmitted by RRLH lamps and received by light sensors (e.g. photodiode or camera) on UEs. The VLC received signal strength data is measured in the downlink and transported to the location database by the UE's application software.

The location server estimates location coordinates of all connected UEs. The location estimation is based on the measured parameters as well as location assistance data. The location assistance data is represented by coordinates of all mmWave antennas and LEDs that are stored in the location database. The location server and the location database are VNFs implemented at IHIPGW. The location server is represented by location estimation and data fusion algorithms. The location estimation algorithms estimate UE position based on the Time Difference of Arrival (TDoA).

The distance d_i between the a RRLH and the UE is shown below as a product of c, the speed of light and T_i, the measured time delay of the signal between the UE and the RRLH.

$$d_i = cT_i \qquad 4.1$$

This distance d_i is equal to the Euclidean distance between the priory known cartesian coordinates of the RRLH (x_i, y_i, z_i) and those to be found of the UE (x, y, z).

$$d_i = cT_i = \sqrt{(x - x_i)^2 + (y - y_i)^2 + (z - z_i)^2} \qquad 4.2$$

The range difference d_{i1} or TDOA, between the ith receiver and the RRLH where the signal arrives first, presents the following nonlinear hyperbolic equation[60][61].

$$d_{i1} = d_i - d_1 = \mathrm{c}(T_i - T_1) = \sqrt{(x - x_i)^2 + (y - y_i)^2 + (z - z_i)^2} - \sqrt{(x - x_1)^2 + (y - y_1)^2 + (z - z_1)^2} \quad 4.3$$

where *x, y* and *z* are the unknown coordinates of the UE, x_i, y_i and z_i are the coordinates of the i-th RRLH while *x1, y1, z1* are the coordinates of the reference RRLH and (T_i-*T1*) represents the time offset between the it-th and the reference RRLHs. To solve for *x, y* and *z,* four RRLHs are required to provide a set of three nonlinear hyperbolic equations, the solution of which gives the 3-D coordinates of the UE. Solving the nonlinear equations is non-trivial. Consequently, linearizing this set of equations is commonly performed using linearization by Taylor series [61][62]. The data fusion is represented by Kalman filter tracking algorithm which fuses the mmWave and the VLC localization.

The location database is implemented in MySQL, and it decouples UEs, RRLH controllers, location service clients and the location server. When RRLH controllers and UEs collect a new set of mmWave and VLC measurements they transfer them to the location database, then the location server obtains the measured parameters from the database together with the assistance data and does not need to interact with RRLH controllers and UEs. Estimated coordinates are saved in the corresponding table within the database.

The parameter measurements as well as the computation of UE coordinates are continuously performed by UEs, RRLH controllers and by the location server. The measured parameters as well as the location estimates are periodically updated in the location database. The location service clients access the location estimates only on demand according to the needs of their application software. The mmWave based localization benefits from the large absolute bandwidth of license free 60GHz ISM frequency band and frequency bands that were deregulated for 5G communication systems such as 37-38.6 GHz, 38.6-40 or 64- 71 GHz. The absolute bandwidth of mmWave ISM and 5G bands (with carrier aggregation) covers a couple of GHz.

Such frequency bands can provide excellent time resolution which may result in sub-decimeter localization accuracy. This accuracy corresponds to the vision suggested for 5G networks [62], [63] and clearly outperforms accuracy of the commercial global navigation satellite system with its outdoor accuracy of about 5 meters, or the accuracy of indoor wireless local area network fingerprinting techniques which is about 34 meters.

The VLC based localization benefits from inexpensive installation by utilizing already existing illumination systems with few modifications. Its further advantage

is its improved resistance to multipath in comparison to mmWave systems [64]. The accuracy corresponds to the lighting environment and LED device quality [65]. Based on three LED lights, the estimated UE position accuracy can be between (4-20) cm with different positioning algorithms [65]. Lab simulation were performed to test our positioning system, where the lab dimensions were 5m (W)*5m (L)*3m (H), and where it was equipped with 4 LED lights on the roof, which transmit visible light signals. The test subject moved in curve on the ground, as can be seen in Figure 4-4 three different position points, the blue dots in the first five boxes represent the actual position of the object, while the green "x" marks represent the estimated location for the object, the last box in the figure represents a global view of the object, with dots with different colors that represent the trajectory of the object during the test. The VLC power received by the UE device can obtain an estimated distance between the UE and the LED. With the location of the LED, the estimated location of the UE can be calculated. 95% of the performed simulation tests showed position error within 10cm.

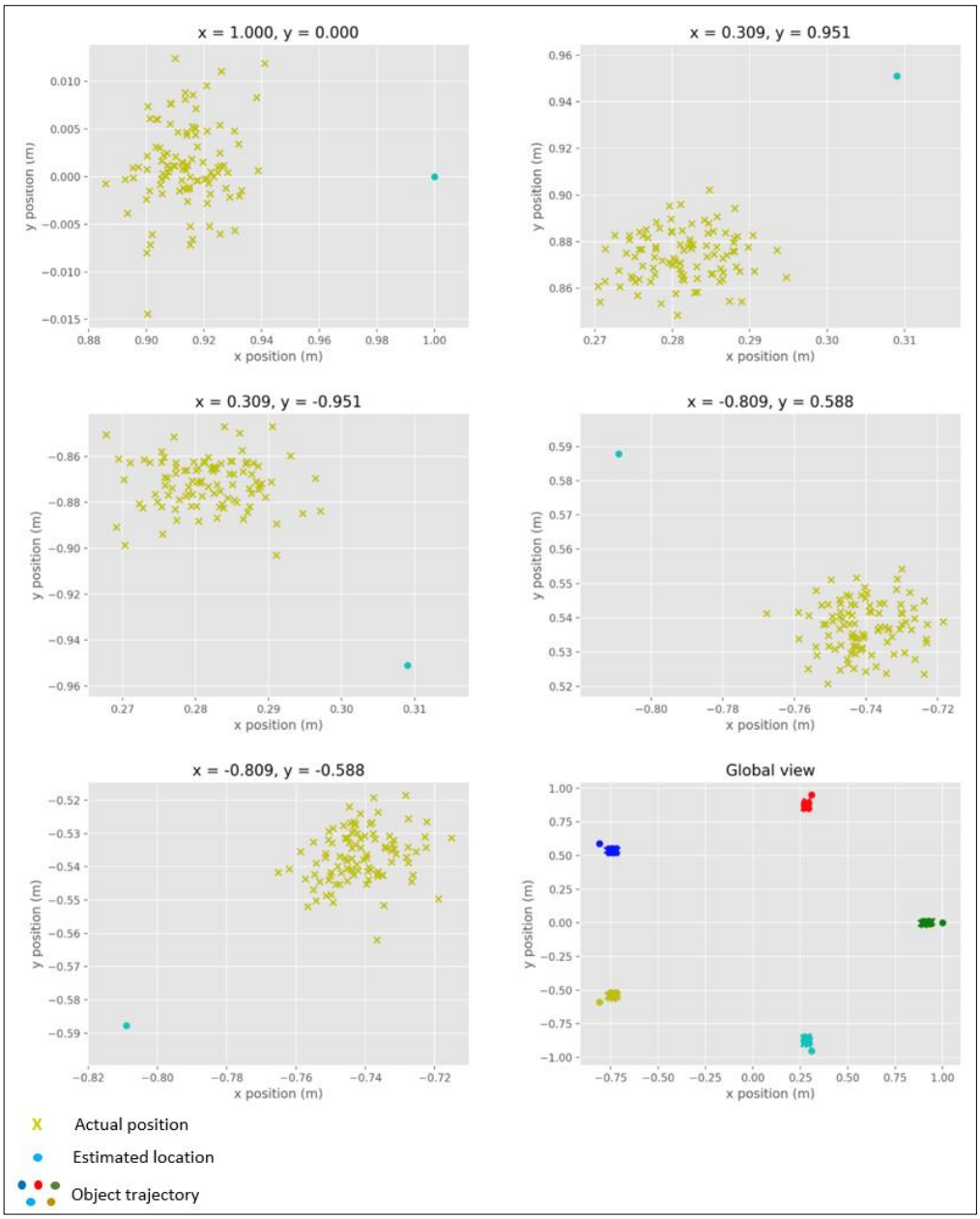

Figure 4-4 VLC location estimation on different positions

4.5 Services' architecture and operation procedure

This section presents the architecture of the services in addition to the operation procedures.

4.5.1 FMS and MSS architecture

FMS and MSS are deployed on the same platform, therefore share a common architecture. The main technical difference between the services, is the service application, where each service configures the system according to the service application, thereby achieving the required service performance. Form this view

point, this section presents only FMS architecture, to reflect the architecture of both services.

System architecture of the FMS is shown in Figure 4-5. It consists of an end user application installed on a UE smart phone, SDN controller, Follow Me Application (FMA) written in python running on top of the SDN controller, and proxy/cache server. The main focus of FMA is to enable the SDN controller to update the OvS forwarding tables with the correct TV destination. FMA utilizes UE location update information, to ensure real time traffic switching instructions to be sent via SDN controller.

FMS seamlessly interacts with network entities to realize the service. The service procedure as the following:

- UE registering for this service and requesting video content.
- UE chooses to either display the content on a specific TV set, or multicasts it to all TV sets.
- FMA configures the OvS to forward the downlink stream to the correct TV device.
- The proxy server intercepts the TCP connection request originated by the UE, and fetches the video content from the content server, then passes the content to the end destination using UDP sessions.
- TCP Proxy maximizes throughput by establishing asynchronous connection to the external server, downloads and buffers the downstream data at the proxy and generates ACKs towards the server. Then it uses connectionless UDP for sending the downstream to enable fast stream redirection. FMS exploits very wide bandwidth (20-100 MHz) offered by IoRL, which highly reduces the probability of dropping packets due to congested network.

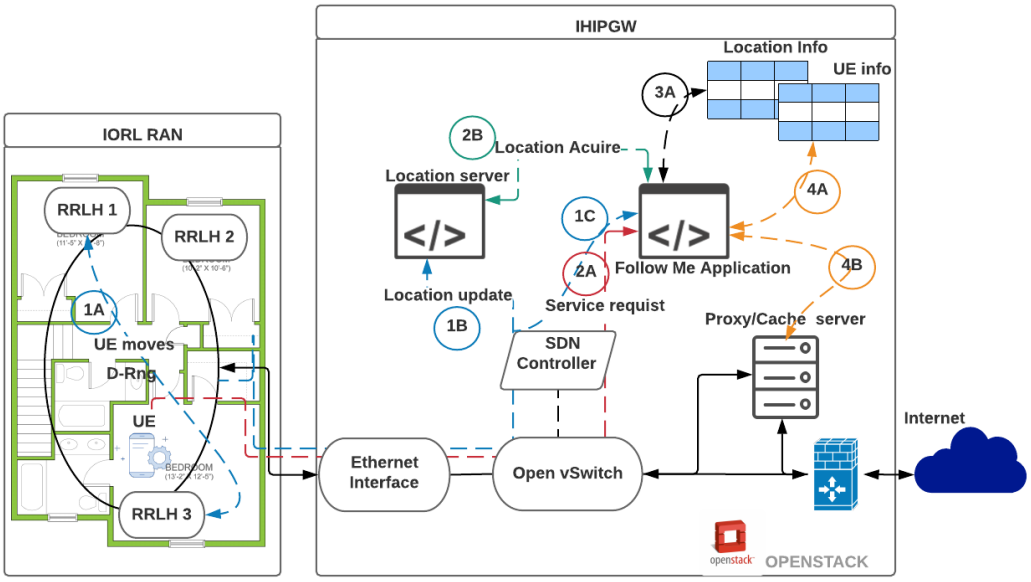

Figure 4-5 FMS Architecture

4.5.2 FMS Signaling sequence

The connection establishment for the FMS and down streaming of the video is based on location updates of the UE, as shown in Figure 4-6.

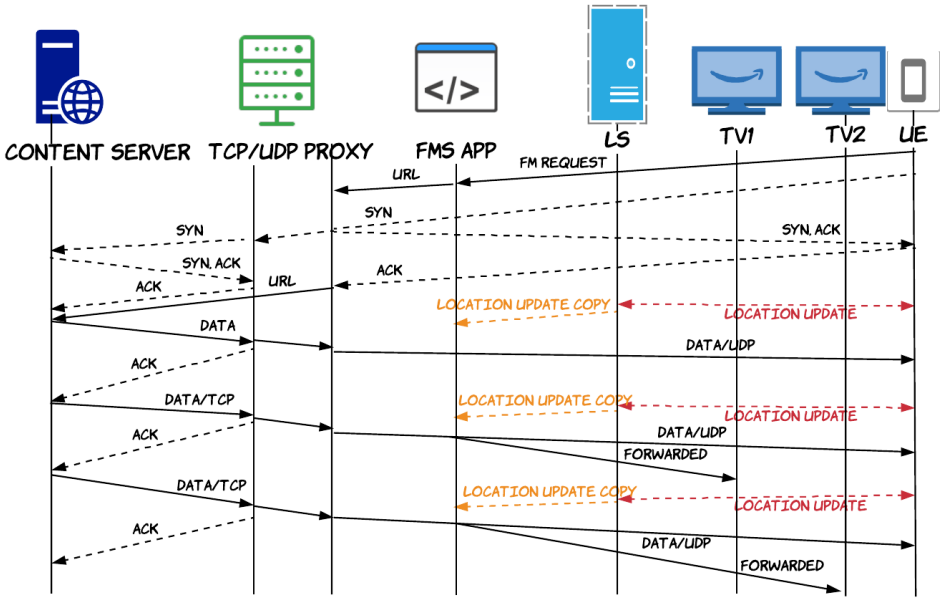

Figure 4-6 FMS signalling sequence

It requires a client to request video content, consequently requested by the proxy from a content server (if it has not already been cached earlier). The proxy splits the session into two TCP sessions: i) UE-Proxy; ii) Proxy-Content Server.

FMA monitors FMS clients' location updates, and modifies the OvS flow rules when necessary, ensuring traffic delivery to the correct TV destination. SDN intelligence plays an important role in improving FMS performance, where Ryu controller [66] configures the OvS to select the downstream path either via VLC/mmWave, or WiFi link, based on the acquired network metrics feedback.

4.5.3 FMS OPERATION

Follow Me Service uses three different components to provide its services to its clients within IoRL, namely: proxy server, FMA, and end user application. The service also utilizes the connection-less nature of UDP protocol to send the video to multiple destinations without the risk of losing datagrams due to handshakes and session establishments. The operational procedure of the FMA and the proxy server are introduced in this section.

4.5.3.1 FMA operation procedure

As shown in Figure 4-7, FMA is registered as the listener for two types of incoming messages: follow me request messages, and location estimation messages.

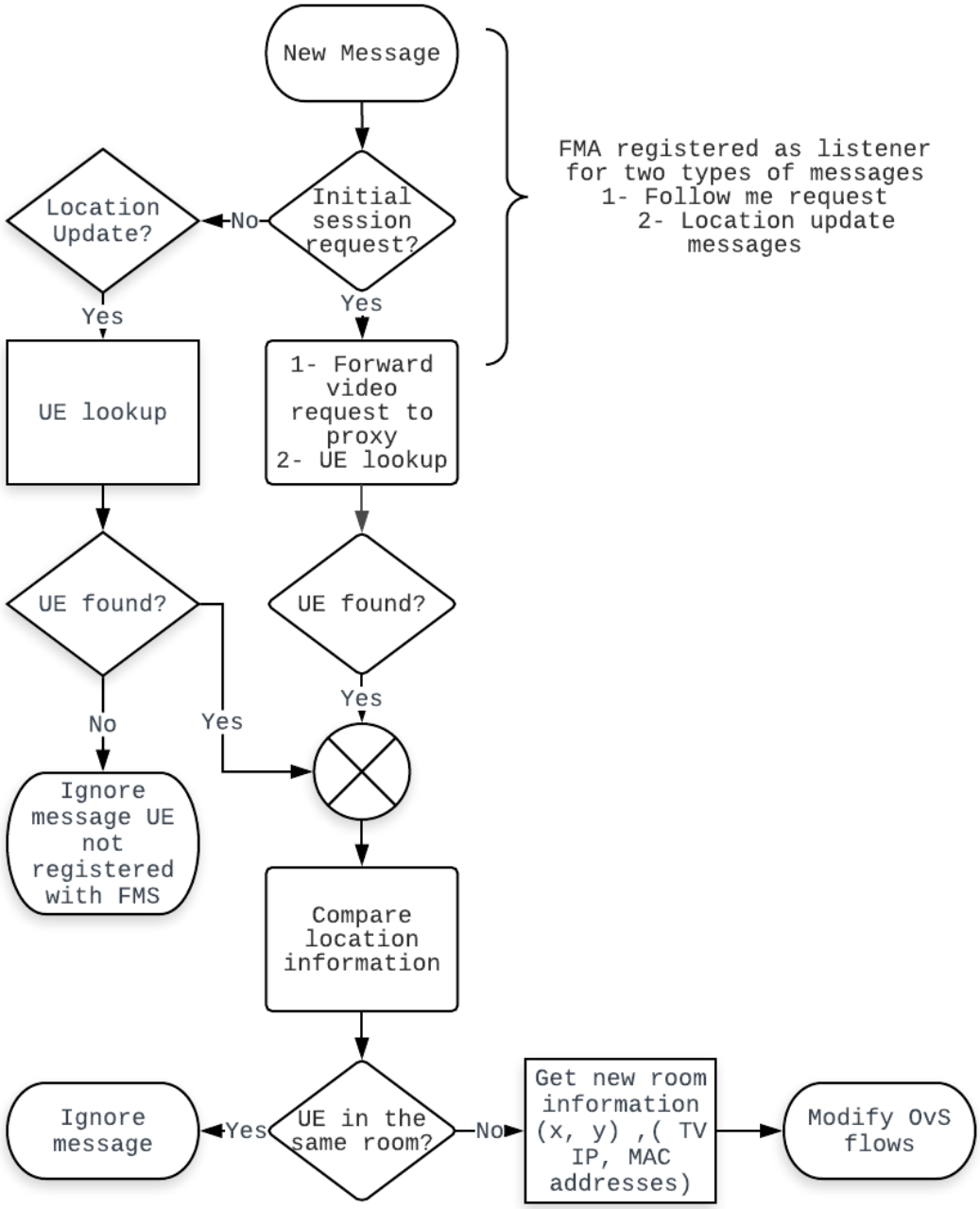

Figure 4-7 FMA operation procedure

1. When a location server sends a location estimation message to the location database, the FMA receives a copy of that message as shown in Figure 4-5. FMA checks if the UE is registered for the FMS by inspecting its local table of the registered UEs, and if a match is found, then it goes to step 3 (compare location information), otherwise, the message is ignored (UE is not registered with FMS).
2. If the received message is an initial request message from the UE, then FMA provides the controller with a forwarding policy to configure the OvS to

forward the URL request to the proxy server. At the same time, the FMA looks up its local table to check if the UE is a new client or if it is an existing client.

   a. In case of a new client, FMA sends location acquire message to location database, and stores the UE location information from the returned response message.
   b. Otherwise, it is an existing client, forwarding to step3 (compare location information).

3. FMA compares the UE current location information against the previously stored location information. Note that:
   a. Client's location information is used by an algorithm to workout TV information (IP and MAC addresses) of the area where the client exists.
   b. New client has no previous TV information; therefore, the current information is used by the OvS to forward the video to his first location.
4. If the client's new location is in the same room, then nothing needs to be done because the movement of the UEs does not require traffic redirection, while if the new location is in another room then the follow me procedure is triggered, which includes:
   a. Storing the current TV information that corresponds to the client's new location.
   b. Controller modifies the OvS flows to forward the flow to a new location. If the UE is not in the FMS enabled area, then the controller sends a pause request message to the proxy server.

4.5.3.2 Proxy server operation procedure

Proxy operation sequence is shown in Figure 4-8. When a new message is received by the proxy, it is either a client initial session request for a video or an update message from FMA.

1. In case of the received message is an initial session request
   a. The proxy checks if the video is already available and cached on the local caching server, or it needs to request it from the external server.
2. In case of the received message is an update request message
   a. If the update message includes a request to pause the video, then the proxy will pause and buffer the video until further request.

While if the update message includes a stop request, then the proxy will stop the video.

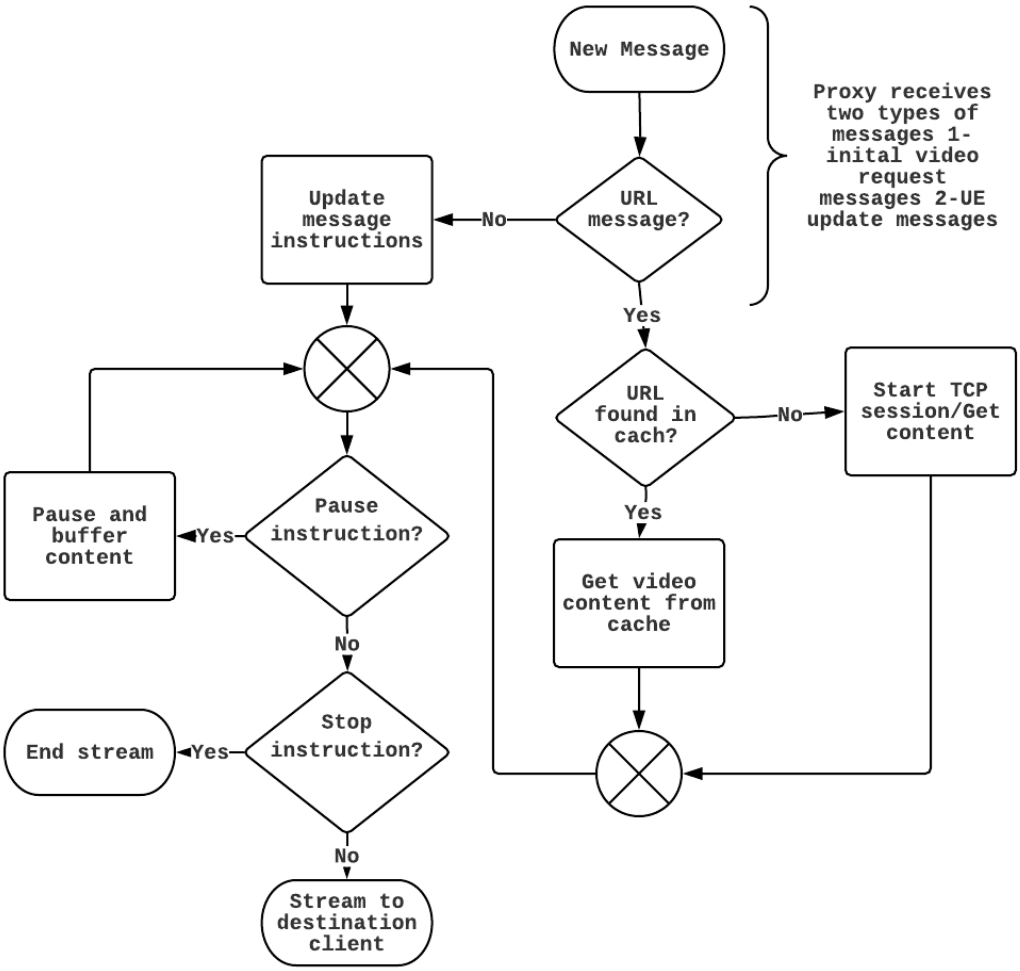

Figure 4-8 Proxy Server Flowchart

4.5.4 MSS operation procedure

To clarify the operation procedure of the MSS, assuming a hypothetical deployment scenario as shown in Figure 4-9. A lecturer (denoted UE1) is making a scientific museum tour with a group of students (UE2, UE2, UE3 and UE4). UE1 would like to share scientific media content within the group by utilising the MSS Service, which is offered by the IoRL network covering the museum. The operational procedure is as the following:

1. UE1 initiates MSS group, and registers as a host.
2. UE1 selects whether to share the contents based on location proximity (for subscribed listeners) or based on preselected group of listeners.
    a. In case of UE1 selects sharing based on location proximity, then he would be prompted to select multicasting diameter.

b. MSS Service maintains list of active listeners based on continuous location information update for host and listeners, in order to ensure proper diameter of multicasting.

3. Other UEs search for the available groups and register themselves as listeners for their group.
4. MSS Service prevents any media collisions by allowing only one UE host per subscribed group.

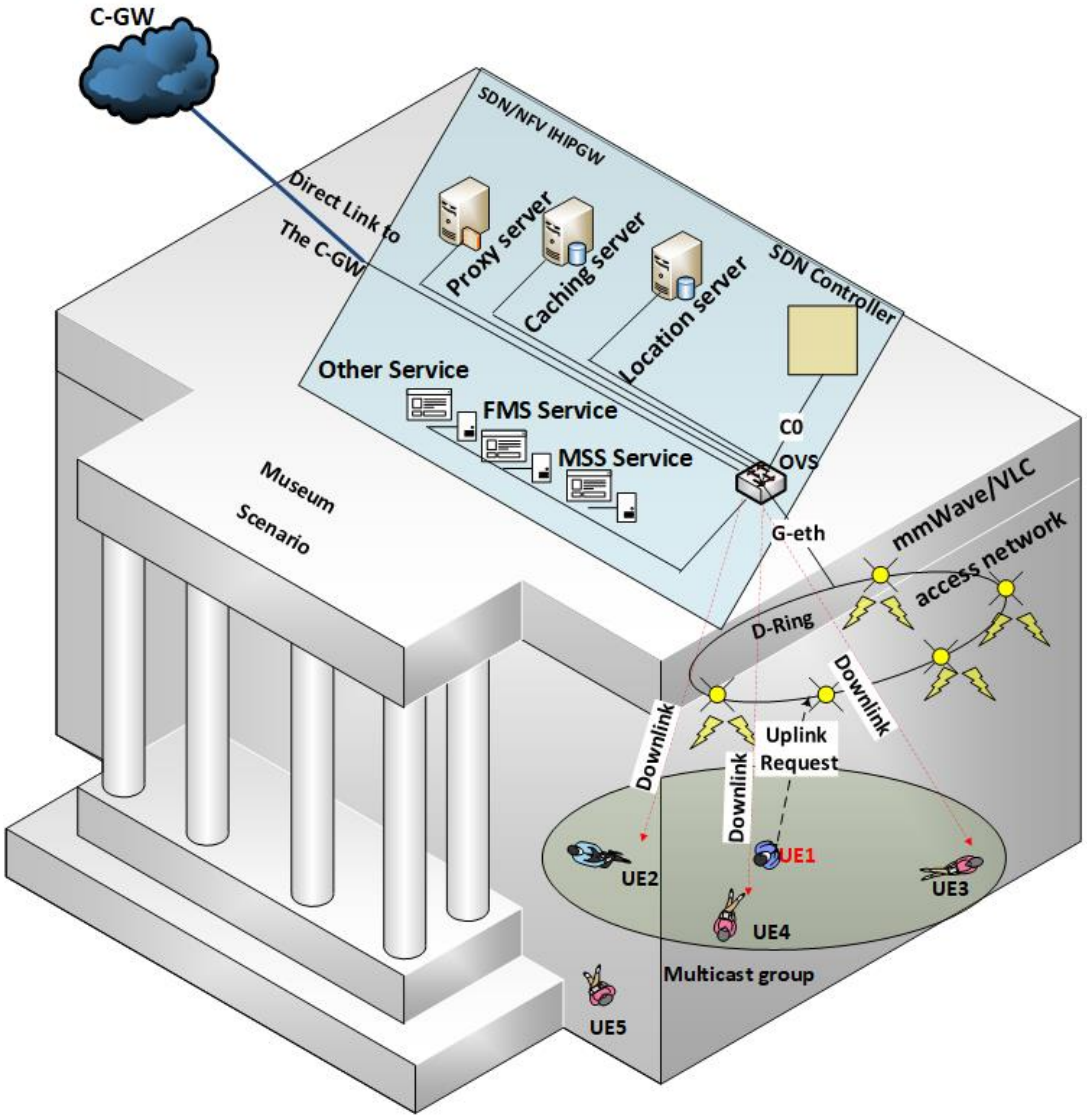

Figure 4-9 IoRL Multimedia Services/ Museum Deployment

The MSS Service's simplified operation procedure is shown in Figure 4-10. The request and response procedure shown in Figure 4-11 starts by the host client UE1 requesting video content to share it with a group of UEs. The request is forwarded by the OvS to the proxy/cache server. Cache server search for the requested video is performed locally, and if it is not found, then the video is requested from an

external content server. The MSS application monitors the clients' locations and sends API call to the SDN controller in order to send flow mod signaling to update the OvS when necessary to make the required change to ensure traffic delivery to the correct destination.

SDN intelligence plays an important role in improving MSS performance, for instance, SDN controller is utilizing the network metrics feedback to configure the OvS for selecting the downstream path either via VLC/mmWave, or WiFi link.

Also based on clients' location, MSS application modifies OvS group table to add/delete listener clients based on their relative location information between the listener UEs and the host UE.

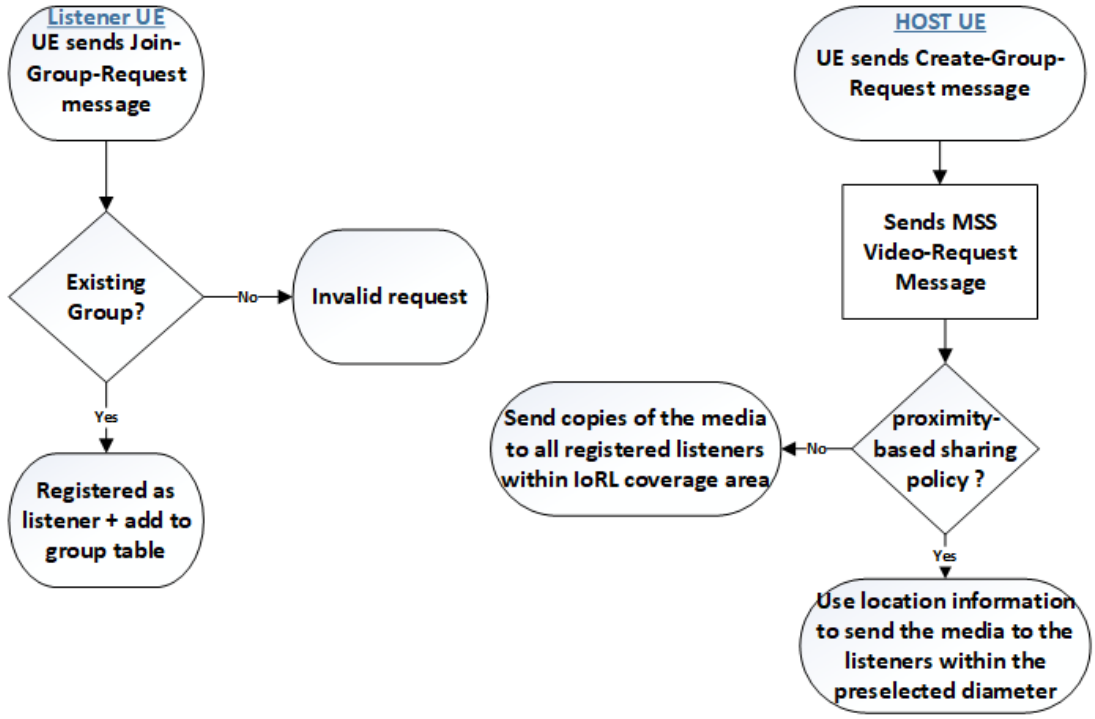

Figure 4-10 MSS Operation Procedure

FMS and MSS services are utilising the users' location information to offer intelligent services for IoRLs UEs. During the operation of the aforementioned multicasting services, they encounter modifying (deleting/ adding) new flows, triggered by the UEs location update. The deletion/ addition of the flows is enabled by the OF protocol. However, it creates new technical challenges. For instance, the transport layer at the clients could reject the received traffic due to unexpected sequence number of the transmitted segments. A proxy server that performs TCP-UDP mapping mechanism were proposed, to overcome such challange.

FMS and MSS services exploit a very wide bandwidth (20-100 MHz) offered by IoRL, which highly reduces the probability of dropping packets due to network congestion. In IoRL architecture there is a location estimation mechanism, which

uses dedicated algorithms to calculate user location depending on mmWave TDoA, and VLC Received Signal Strength (RSS) to locate the user. Location database maintains an updated location information for all UEs within its network coverage. The location information is made available for a third-party service developer, to offer intelligent location-aware services.

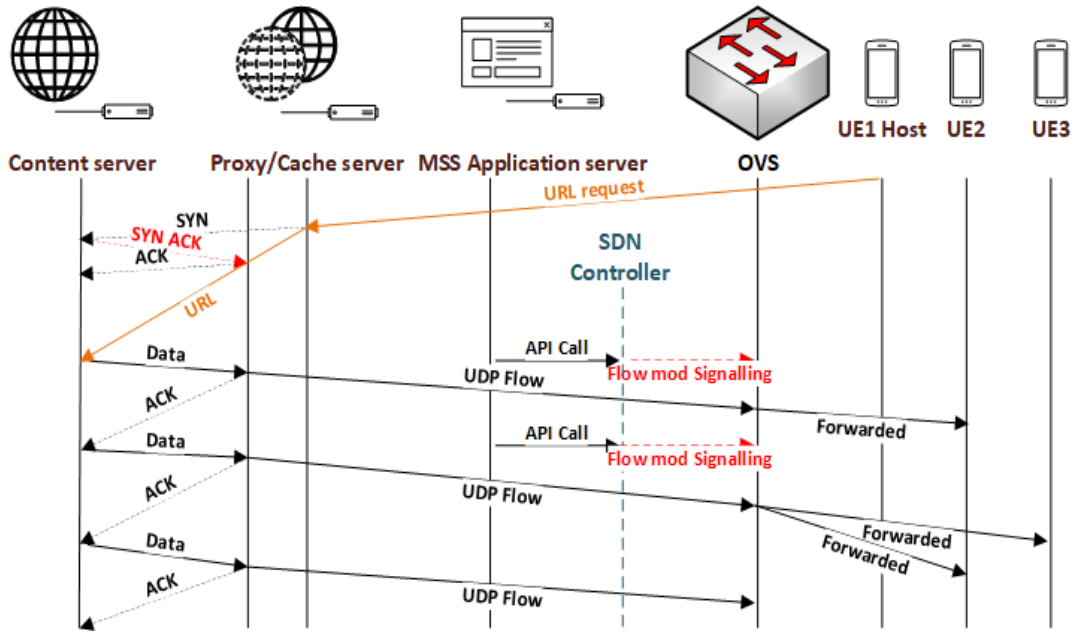

Figure 4-11 MSS Signalling Sequence

4.6 Performance evaluation

This section aims to quantify the potential of the proposed services, by providing multiple hypothetical scenarios in order to examine and evaluate the services' performance. First, the testbed is described, then the various performed scenarios are presented.

4.6.1 Testbed description

The testbed comprises of four layers, namely: service layer, control layer, access layer, and transport layer. Scenarios were implemented using Mininet emulation tool; Mininet was chosen since it is a well-known tool and has been widely used for emulating software defined networks. Our platform emulates two parts of the network: the IHIPGW, and the Radio Access Network (RAN). Since the actual RAN is under development, the implementation was simplified temporarily by assuming that layer2 of the RAN is capable of routing the packets to the correct RRLH controller. Then the RRLH controller forwards to the correct mmWave/ VLC modules, to be delivered to the correct TV set. These simplifications were made to show proof of concept of our developed service. The layers' details as the following:

- Service layer: this layer exposes the interaction between several network entities to provide the FMS. This service consumes UE location information, to perform video redirection to the current UE location.
- Control layer: it is the central part of the system architecture, and it reflects the SDN intelligence by utilizing the UE location information to provide seamless traffic redirection by configuring the OvS to perform UDP traffic forwarding. This layer comprises of the Ryu SDN controller, as well as the SDN applications.
- Access layer: as previously mentioned, the RAN has not been completely implemented therefore the Mininet tool was used to emulate this layer.
- Transport layer: this layer allows interconnection between the access layer and service layer. It utilizes virtualized network entities to perform its task, such as, OvS, location database, and proxy server.

4.7 Results and analysis

This section includes the services' deployed scenarios and performed tests, presenting the obtained results and critical analysis.

4.7.1 FMS deployment scenario

The deployed network topology in the hypothetical, multiple-room home scenarios is shown in Figure 4-12.

The network consists of one OvS, one Ryu controller, and eight hosts to emulate different network entities. Four of the eight hosts emulating TV sets are distributed within multiple rooms of the home. One host emulating the UE is registered as an FMS client, two hosts running in server-client mode are emulating the location server and location database. These servers are used to perform clients' location estimation and a storage mechanism. The last host is running Iperf traffic generator to emulate proxy server that is streaming cached video stream to a client. A Python script running as SDN application within the Ryu controller is used to configure the OvS by installing flows to route the video traffic to user's location. The controller is also configured to listen to the clients' location estimation update messages thereby, modifying the flows to reroute the video stream to the TV set within the new location of the client.

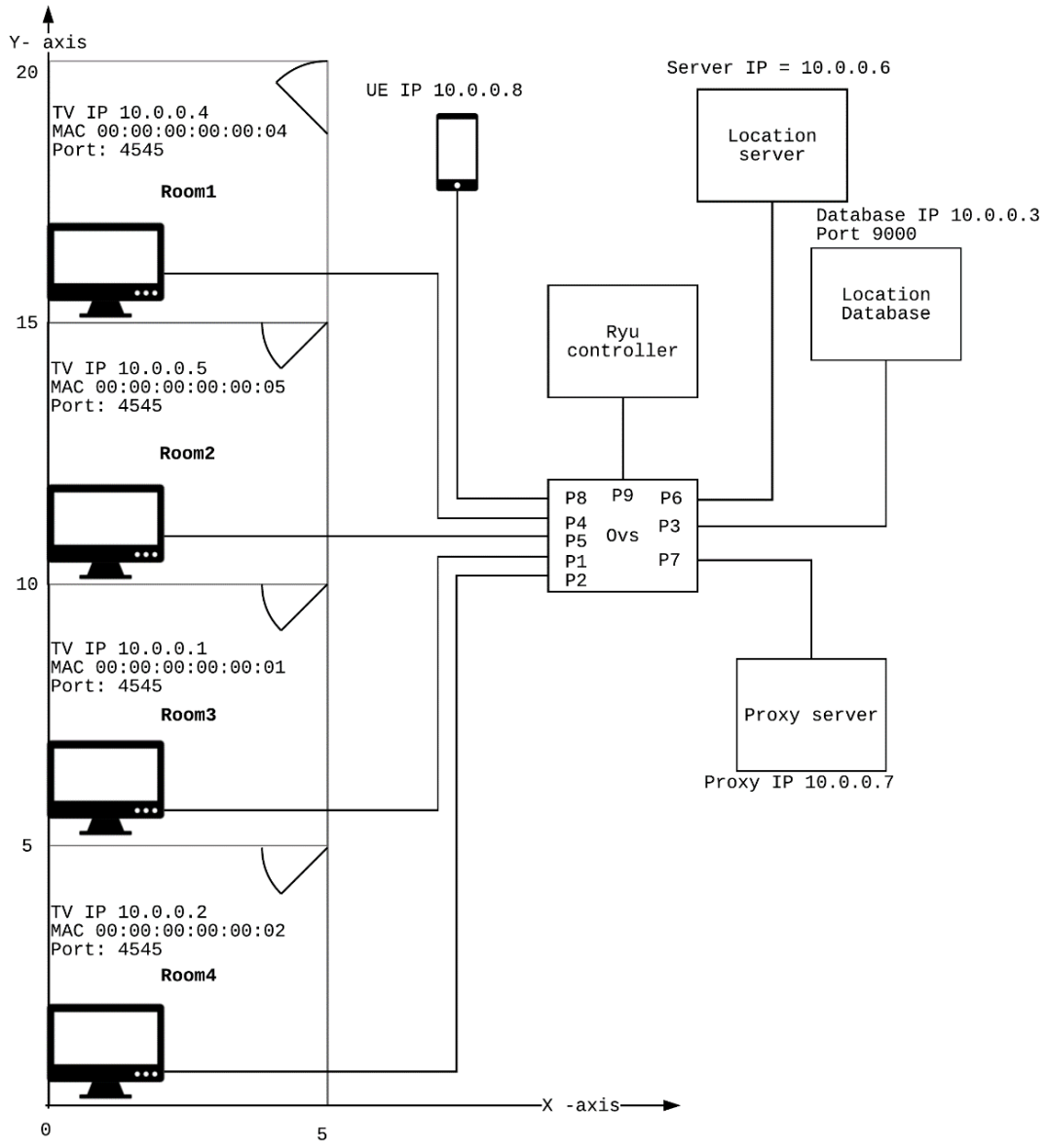

Figure 4-12 deployment scenario topology

Several scenarios were used to demonstrate FMS reactions to client's movement within the home premises. The calculated quality of service (QoS) parameters in all scenarios were compared with each other to evaluate the system performance. The tested QoS parameters are packet jitter and throughput, I have chosen these specific parameters as they can have a significant impact on the offered service, especially for multimedia services and having issues with packet jitter in live stream can cause a detrimental effect.

Each test lasted 100(s), with Iperf generating 80 (Mbps) UDP stream, while UE location estimation data being randomly generated within home coordinates range and stored in the database.

4.7.1.1 Test one: No-hop vs One-hop Room change scenarios

In the first scenario run, a client is viewing a video stream on the nearest TV set. His location information values for the entire test period are confined to one room coordinates (Room1). Therefore, the video stream header contains the source IP address (10.0.0.7) of the proxy server, and destination socket address (10.0.0.4/4545), which remains unchanged. In the second scenario run, the client starts at one room receiving on socket address (10.0.0.4/ 4545) until he changes location to another room to continue viewing the video stream on the new socket address (10.0.0.5/4545). Some QoS parameters were measured (i.e. Throughput, packet loss, and Jitter) in each scenario and compared to each other to evaluate the system performance when the client's location changes results in one-hop destination change.

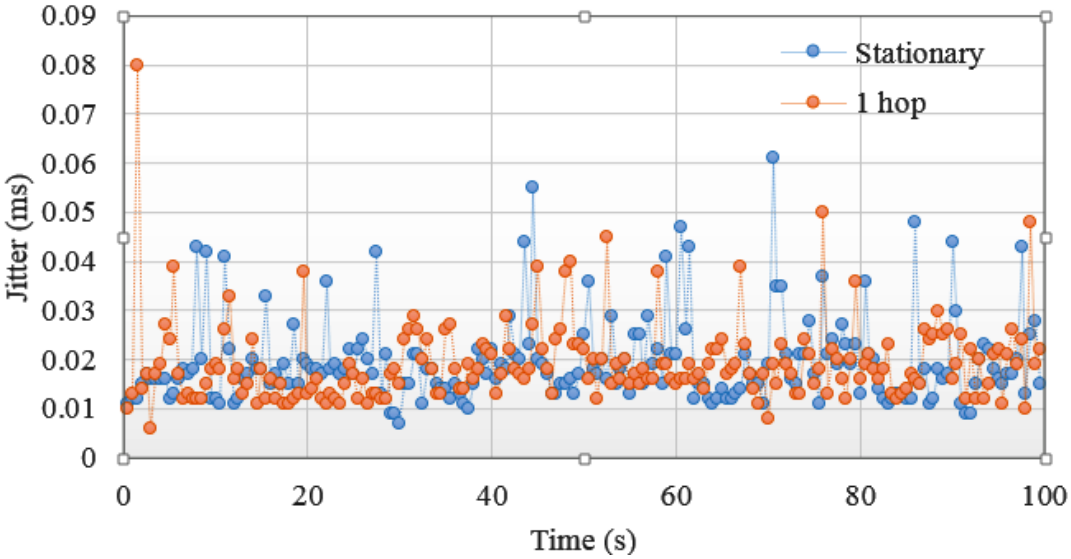

Figure 4-13 jitter No-hop vs one-hop scenarios

As shown in Figure 4-13, the stream jitter for the test scenarios were compared, where during the first scenario run, the SDN controller configured the OvS to route the stream to one destination address, while during the second scenario run the OvS was configured to reroute the stream twice during the test run time, therefore, the comparison considering a stationary user verses a mobile user with only one hop mobility. Jitter ranged around 0.03 (ms) with some spikes in random times. Although the worst jitter experienced was below 0.1 (ms), this jitter was due to the nature of the Iperf tool.

It is worth mentioning that, according to Cisco [65], the acceptable jitter for video is below 30 (ms), this acceptable jitter should also have delay conditions, i.e. the one-way transmit delay should not exceed 150 ms, while round trip delay should

not exceed 300 ms whenever possible. Therefore, this proves that our system performance is excellent.

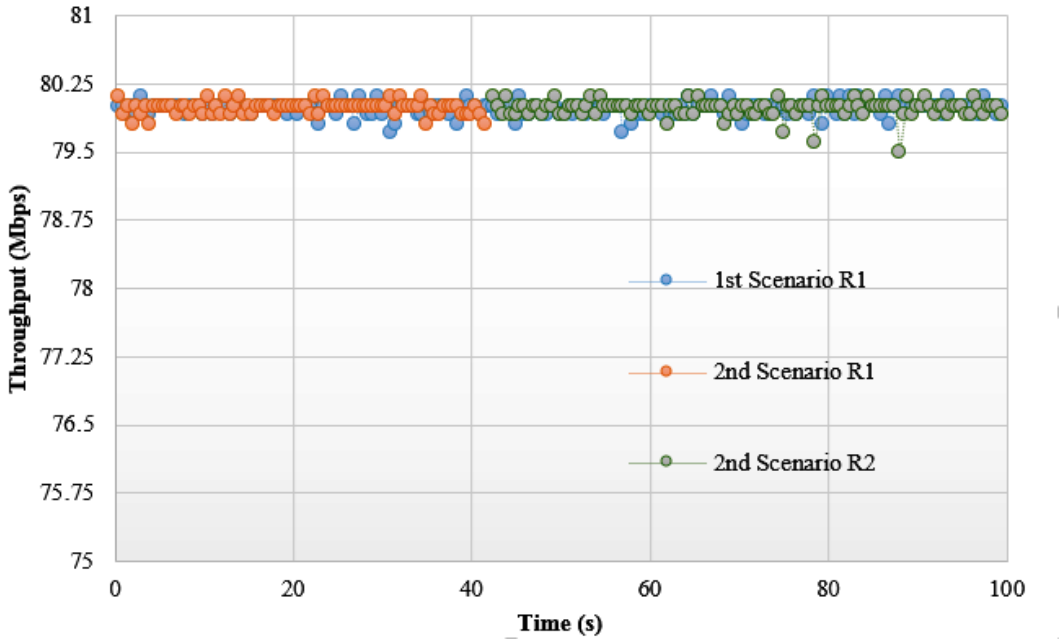

Figure 4-14 throughput No-hop vs one-hop scenario

System throughput for both scenarios of the first test were compared, as shown inFigure 4-14, where the throughput is ranging around 79.75(Mbps), with lowest throughput at 79.5 (Mbps); this result shows very good system performance during the first test scenarios.

4.7.1.2 Test Two: two-hop, three-hops, and four-hop client room change scenarios.

In the first scenario run of this test, a client is requesting a video stream to be viewed on the TV set of Room1 (socket address10.0.0.4/4545), for (35 s), then changes his location to enter another room (Room 3) and continues watching the video stream for (19 s) on the new room's TV set (socket address 10.0.0.1/4545). The client finally renters the first room (Room1) and continues watching the rest of the video stream (46 s) on the room's TV set (socket address 10.0.0.1/4545). Similarly, in the second scenario of this test the client starts in one room viewing the video stream from the proxy server on the nearest TV set for (27.5 s). He then changes his location between rooms resulting in rerouting the video stream three times to the socket addresses (10.0.05/4545, 10.0.0.1/4545, and 10.0.0.2/4545) at times (27.5, 41.5, and 59 s) respectively, to accommodate continuous viewing of the video stream in all locations.

Finally, the client in the third scenario has started viewing in one room on socket address (10.0.0.4/4545) and then changed rooms for three more times, resulting in changing destination socket addresses in the following pattern (10.0.05/4545, 10.0.0.1/4545, 10.0.0.2/4545 and 10.0.0.1/4545) at times (15, 28.5,38.5, and 48 s) respectively.

These frequent location changes enabled us to examine the system performance during and after these changing times to evaluate how it may or may not affect user experience.

The system throughput among all three scenarios of the second test was compared, as can be seen in Figure 4-15, it was observed that the throughput worst drop was below 78.75 Mbps at one point during the first 5 seconds of the two-hop scenario, and it did not drop any further during the rest of the test run time. Furthermore, according to the test scenario, the first hop time of the two-hop scenario was at (35 s) of the run time, which proves that this drop was not related to service rerouting performance, rather to the Iperf traffic nature.

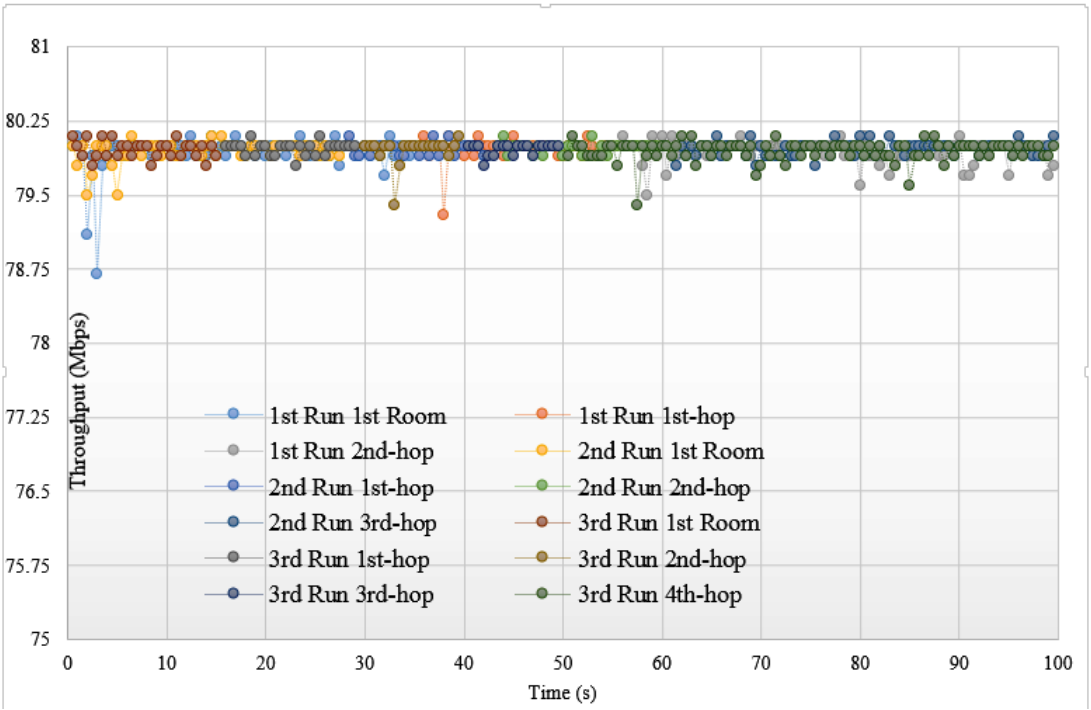

Figure 4-15 throughput two, three, and four hops

Although 1% packet loss is considered acceptable according to [66], the systems performance showed resiliency to packet loss in the network, and no packet loss is encountered in the obtained results of different scenario runs. Figure 4-16 shows how the OvS distributes the received packets from the proxy server via different output ports based on the location of the UE.

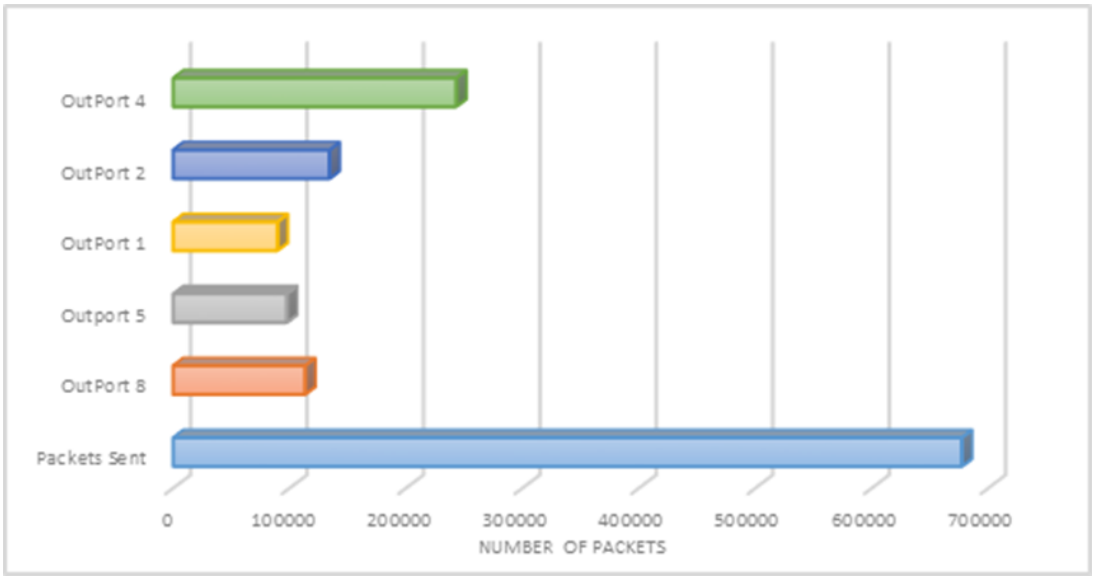

Figure 4-16 packets distribution

This proves that there was no packet loss within the network during testing. To stitch all results together, Figure 4-17 and Figure 4-18 show the client's random location with respect to the hypothetical home, during test run time, time spent in each room for each scenario, and client's hop times.

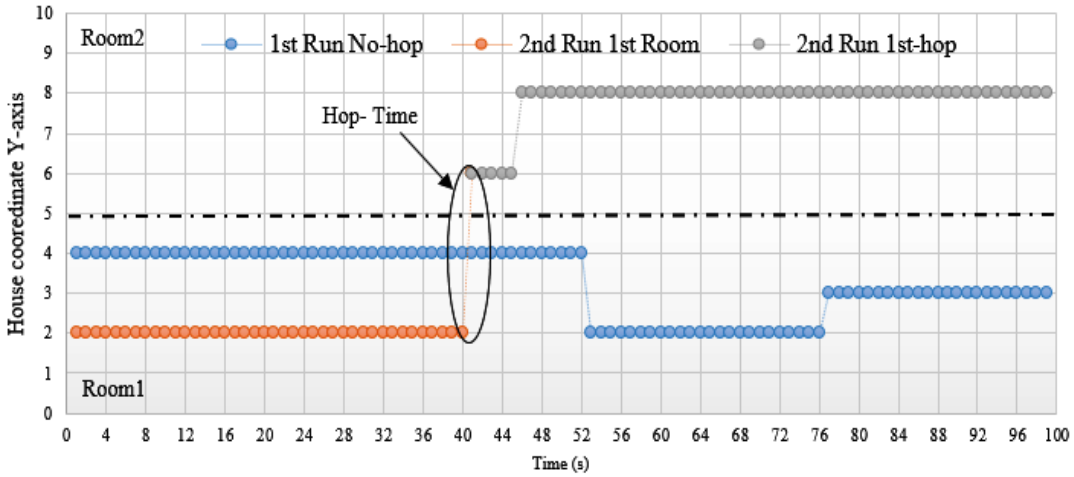

Figure 4-17 hop-time No-hop vs One-hop scenarios

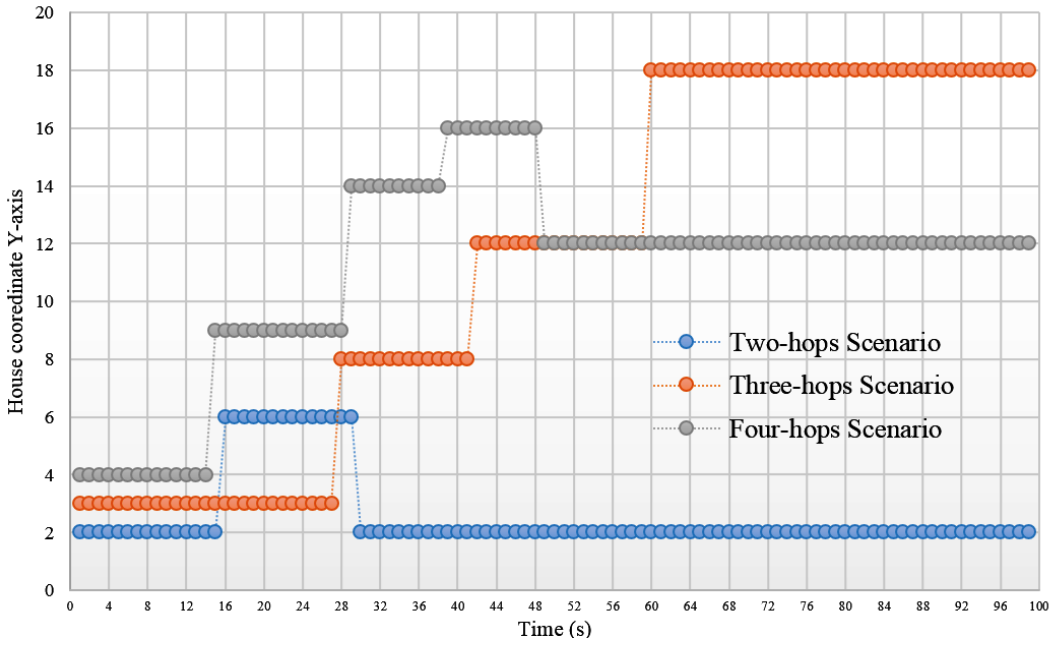

Figure 4-18 hop-time two, three, and four -hops scenarios

When comparing hop times for each run in each scenario as shown in Figure 4-17 and Figure 4-18 with system performance, it can be seen clearly that there is no performance degradation during handover times whatsoever. For instance, at 41.5 (s), the client hands over from room1 to room 2 at the one-hop scenario, and as can be seen in Figure 4-13 the jitter did not jump at this time or shortly after, furthermore the throughput of the system did not degrade at this time or shortly after. It is worth mentioning that similar performance has been observed between stationary, single and multi-hop scenarios. Thus, service achieved good performance during the stationary, single-hop and multi-hop scenarios as reflected in the results.

4.7.2 MSS deployed scenario

The adopted scenario is shown in Figure 4-19, consists of an IoRL small cell deployed in a museum building, where there are 5 UEs, one host and four listeners. Each UE uses a dedicated tour tablet with MSS front-end application installed. Group creation and client registration is performed based on the procedure mentioned in the FMS service section. MSS-create-group message is sent to the MSS back-end-server to create the group, then each client sends Join-group-request message.

MSS Service application listens to the location updates of the host and clients and configures the network components accordingly, to ensure the delivery of the content to valid UEs only. After the creation of the group, UE2, UE3 and UE4 Join

the group, while UE5 does not join since the distance between the host and UE5 greater than 10 meters. After UEs join the group, they start to receive the multicast stream. After some time UE2 leaves the multicasting coverage area for a while and then renters the coverage area again. The MSS performance details based on this scenario are explained in the results section for evaluating the service reaction to the location updates.

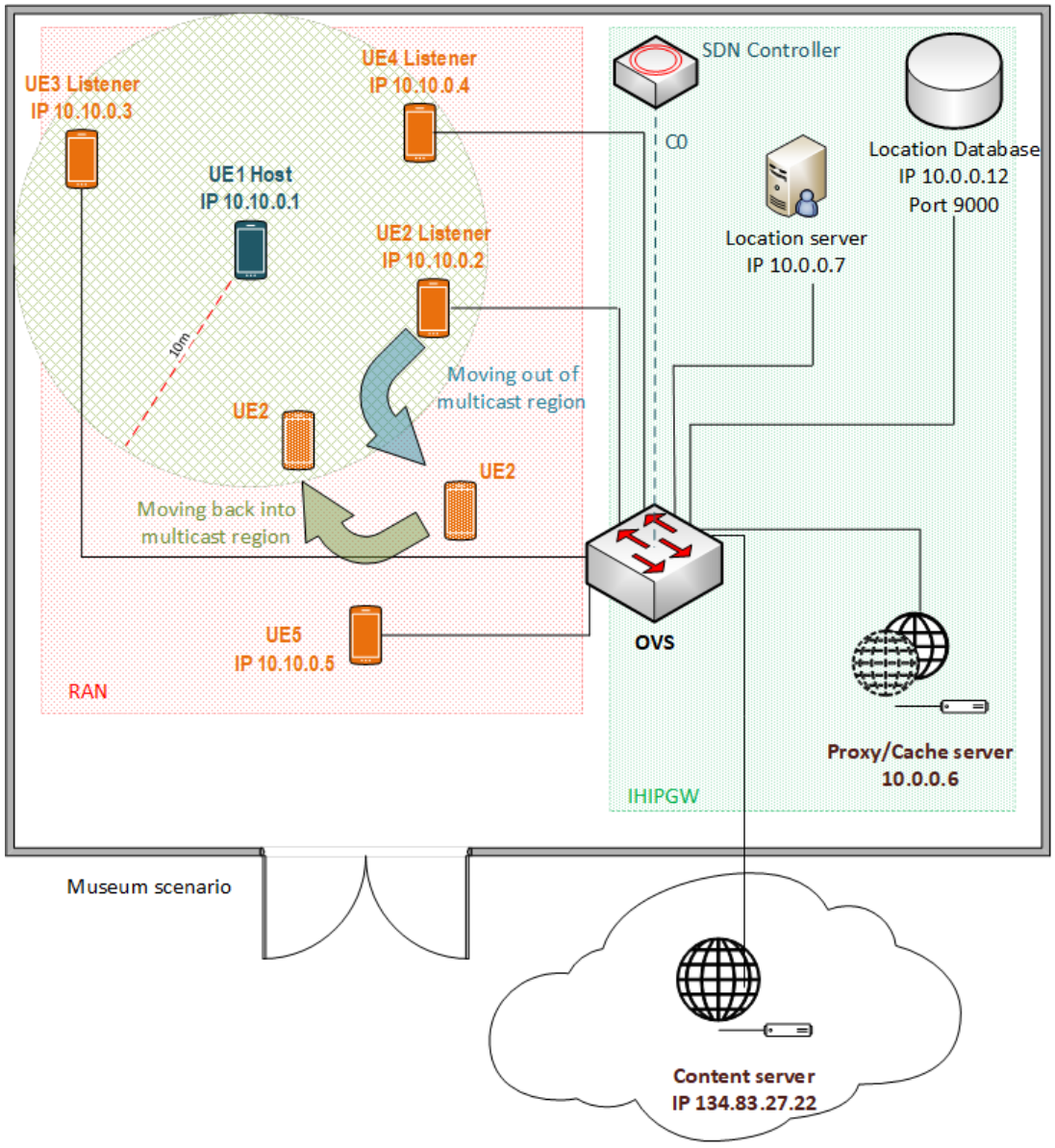

Figure 4-19 museum Deployment Scenario

4.7.2.1 MSS service test scenario

UE1 (host) requests to create MSS group and specifies the multicast range diameter to 10 meters. UE2, UE3 and UE4 join the multicast group, where the host requests

video content and selects to multicast it to the registered clients within the group coverage area. The video Bandwidth is 10 Mbps, for 100 seconds long.

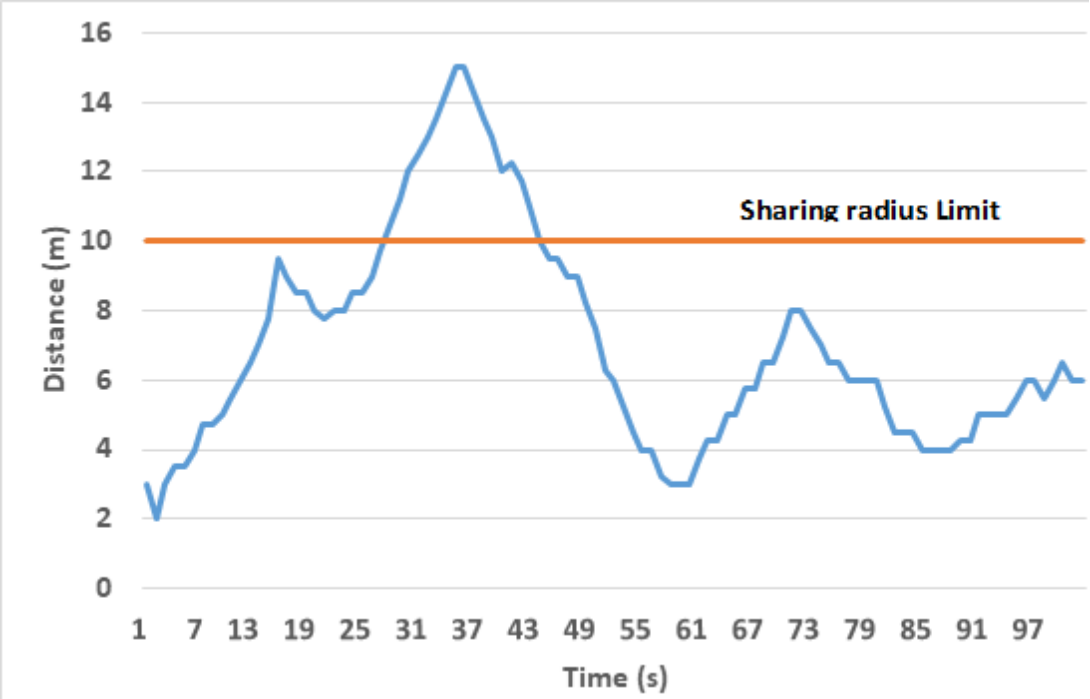

Figure 4-20 separation Distance between UE2 and UE1

The relative distance between UE2 (listener) and UE1 (host) is based on the randomly generated trajectories of the UE2 and UE1, and as can be seen in Figure 4-20 and Figure 4-21, UE2 separation distance is >10 meter from the time (26 second) to the time (42 second).

As shown in Figure 4-21, the stream jitter for UE2 were compared to the other UEs including the host jitter. Jitter ranged around 0.3 (ms) and below, with spikes at random times.

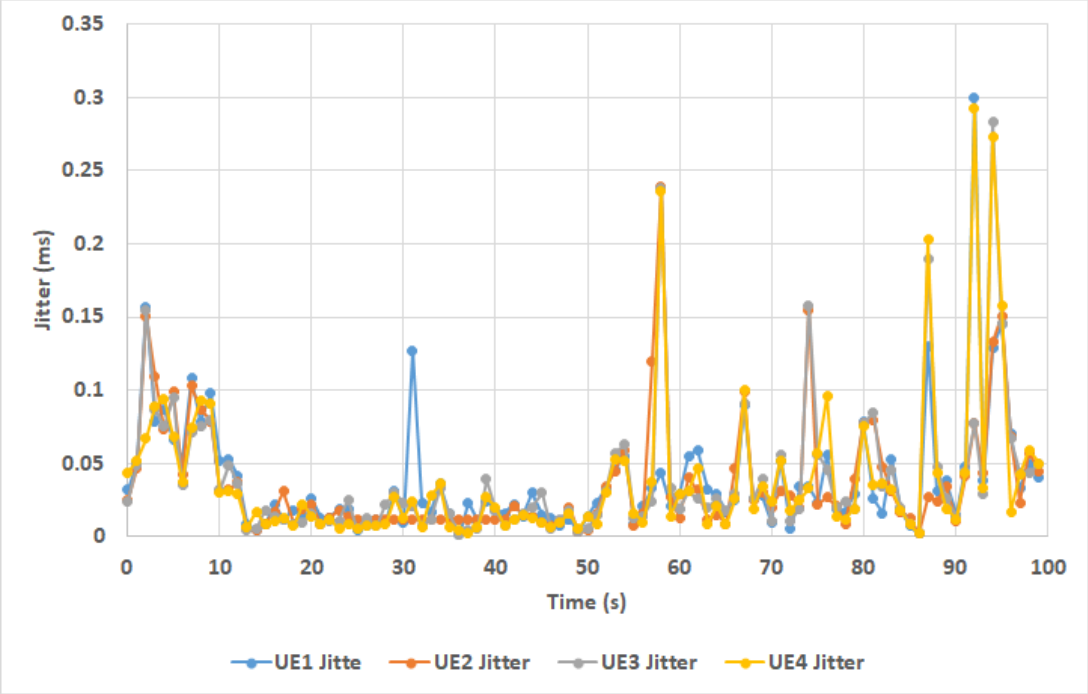

Figure 4-21 traffic jitter

Although the worst jitter experienced was below 0.35 (ms), this jitter was related to the nature of the Iperf tool. According to Cisco [65], video jitter below 30 (ms) is acceptable, which means excellent system performance under MSS Service.

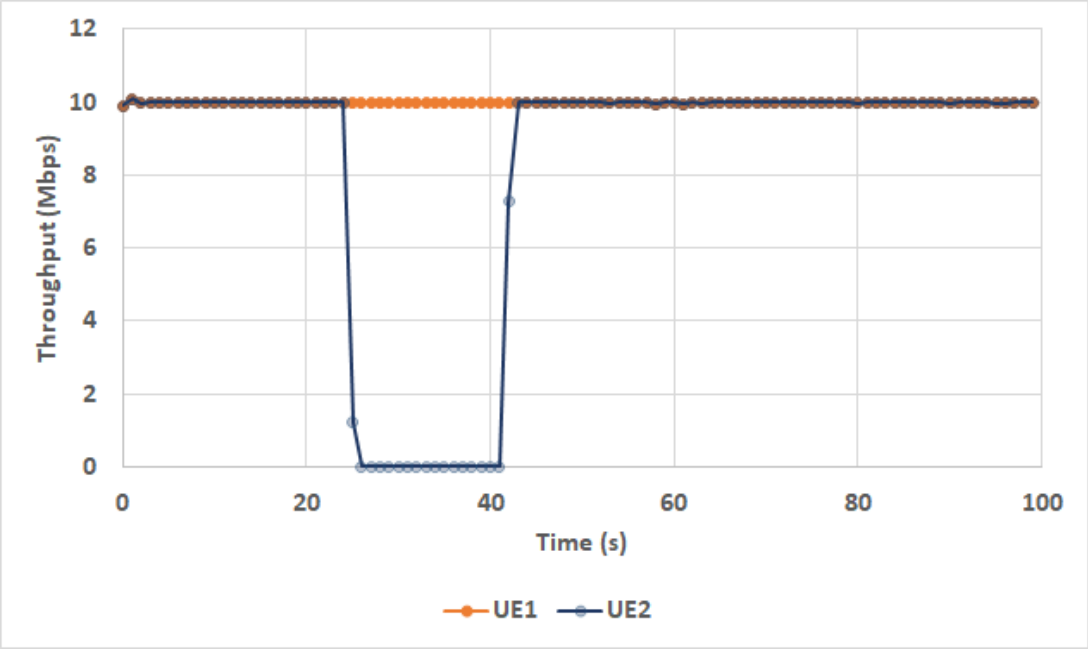

Figure 4-22 UE1 and UE2 throughput

Throughput for UE1 and UE2 is shown in Figure 4-22, where throughput is ranging around 10 (Mbps) for both UEs without fluctuations over the time, which reveals very good system performance. Throughput of UE2 drops to zero at time 26 (s), stays zero for 13 seconds and comes back up to 10 (Mbps) again at time 42 (s) until

the end of the scenario run. This result is coherent with separation distance shown in Figure 4-20, where UE2 throughput drops to zero when the separation distance between him and the host is greater than the preconfigured service coverage radius (10 m), and this fulfils the required service performance.

4.8 Conclusion

Next generation mobile networks are evolving constantly by exploiting the latest technologies to improve the services that are offered to end users. Tremendous efforts are spent to reduce the network CAPEX and OPEX, to enable mobile providers to enhance their network performance efficiently. This chapter provided an insight on two example multimedia services designed for the next generation indoor gNBs namely: FMS and MSS. Both services utilize the virtualization technology as well as cloud-computing concept to offer high performance, cost-efficiency and flexibly deployed services. The services are designed to boost UEs' QoE by facilitating media contents acquisition, distribution and control in a very simple form. The services are tested in an emulated environment to provide proof of concept. The obtained results reflect the effectiveness of SDN networking in the adaptive routing configuration by maintaining zero packet loss due to route switching. In addition, the services highlight the practicality of the IoRL platform in hosting such services along with a high accuracy location tracking system.

5 Small cell caching deployment for IoRL gNB

5.1 Introduction

Fifth Generation (5G) mobile networks are expected to perform according to the stringent performance targets assigned by standardization committees. Therefore, significant changes are proposed to the network infrastructure to achieve the expected performance levels. Network Function Virtualization, cloud computing and Software Defined Networks are some of the main technologies being utilised to ensure network design, with optimum performance and efficient resource utilization. The aforementioned technologies are shifting the network architecture into service-based rather device-based architecture. In this regard, IoRL-Cache service (IoRL-C), is introduced, which is proposed as a solution for improving IoRL small cells caching efficiency. As mentioned earlier, IoRL is an emerging 5G small cell for indoor environment, which utilises mmWave and VLC as access technologies, while exploiting SDN and Network Function Virtualisation (NFV) technologies to offer flexible and intelligent services to its clients. The caching solution is proposed for IoRL small cells specifically, not because there is a specific challenge with IoRL but instead it is mainly making efficient use of the available resources in the proposed IoRL architecture.

The aims of this chapter are to introduce IoRL-C service and performing simulation work for testing the IoRL-C as a proof of concept that supports future deployment of IoRLs efficiently. The network was simulated using OMNeT++ simulation tool and the services efficiency validated by comparing their performance with traditional deployments.

The cache service performance at different link lengths is also examined and found out that IoRL-C is able to support IoRL small cells with more than 50+Km separation distances. It is worth mentioning that the work presented in this chapter is published in IEEE transaction journal on multimedia [67].

5.2 Technical overview

This section briefly describes some of the enabling technologies and concepts, which along with other emerging technologies play an essential role in the development of the proposed solution as a part of the mobile networks' architecture.

5.2.1 Cloud computing

Cloud Computing is the concept of providing pools of managed resources to end users' reliably, enabling the users to utilize, release and maintain the resources as required with minimum management overhead. Cloud computing has made a big impact on application service provisioning. Cloud computing is defined as a specialized form of distributed computing that introduces utilization models for remotely provisioning scalable and measured resources [68].

The cost of storing data has significantly decreased, enabling application developers to maintain and utilize large amounts of data efficiently. The resources can be combined together to form a resource cluster. Figure 5-1 depicts a general topology of cloud computing. A general rule of resource clustering is that it requires high level of hardware and operating systems similarity, to provide similar performance levels. Resource similarity requirement limits cluster creation to homogeneous hardware pool only. Virtualization technology is the enabler for creating the pool of logical resources from underlying heterogeneous resources. This concept has simplified the process of resource utilization and the creation of the virtual resources [69].

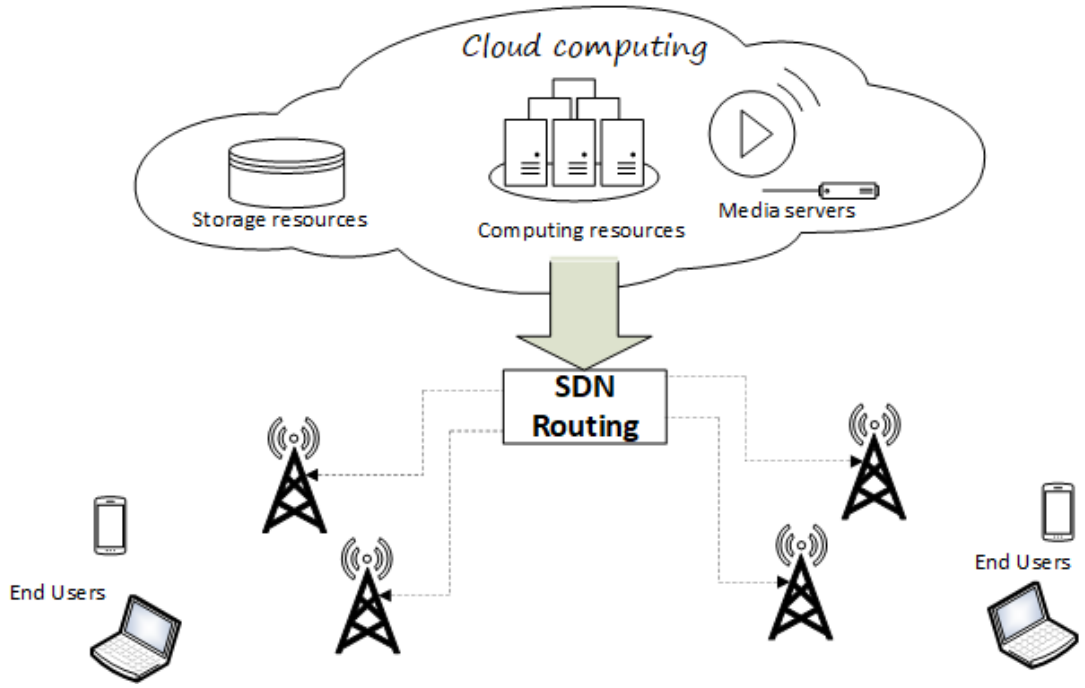

Figure 5-1 general topology of Mobile Cloud Computing

5.2.2 Mobile Cloud Computing

Mobile Cloud Computing (MCC) concept is driven from cloud computing and mobile computing concepts. MCC exploits the RAN to offer rich computational resources to mobile users. The purpose of MCC is to enhance the users' QoE, by

enabling end users to enjoy execution of rich mobile and network services, without relying on the computational power of their devices. MCC provides business opportunities for mobile network operators as well as cloud and media providers. In the traditional service models, normally lightweight clients request a specific service from a powerful server in client-server service model. In the cloud computing service model, the resource provisioning approach is based on cloud computing technologies. In other words, a new resource allocation technology is utilized, based on virtual- physical interactive model, where the virtualized system and the physical system are interactively providing services to each other, to provide the required service to end users. The virtual system can be seen as a set of software applications built on top of a physical system composed of computing, networking, and storage pools, interacting via well-defined interfaces and APIs [70].

5.2.3 Edge cloud

Edge cloud concept refers to relocating services and resource management, from centralized datacentres to logically peripheral network entities as shown in Figure 5-2, thereby maximizing user-service proximity, to enhance the overall network performance and user experience. The main advantage of edge computing can be summarized as follows:

- Reducing significantly the need for traversing large amounts of data between end users and logically centralised data centres, thereby reducing the OPEX via eliminating the need for maintaining high bandwidth backhaul links. At the same time, processing the data locally contributes to reducing the latency and improving QoS.
- Edge processing promotes distributed platforms, which in turn helps to remove network bottlenecks and single point of failures.
- Enhances the security of the networks, by reducing the amount of traffic that reaches the core network, as well as implementing extra security points at the network edge for extra security examination by extra set of firewalls.

The ability of resource virtualization supports faster scalability of the network, which is considered as complementary technology for the edge cloud concept, as resource virtualisation, cloud edge and SDN technologies enable network programmability and automation [70].

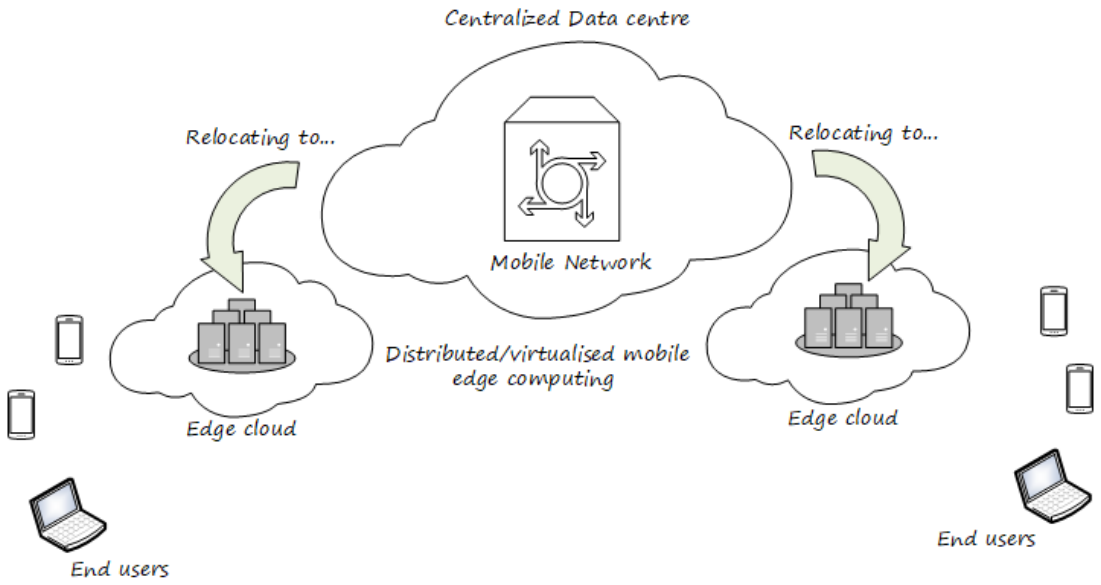

Figure 5-2 Edge Cloud in Mobile Networks

5.3 Related work

The mobile users consuming media contents at incremented rate, are influenced by the continuous emergence of rich multimedia services, which leads to redundant media content transmission causing links' congestions over the network.

On the other hand, researchers from academia and industry are always looking for solutions to enhance the network performance for improving the user's QoE ultimately. Therefore, the network architecture is constantly evolving to include new entities (e.g. small cells) and to exploit new technologies (e.g. SDN, NFV, cache servers …etc.). The deployment of small cells improves the bandwidth and coverage area. However, it encounters some difficulties. For instance, the authors of [71] proposed cooperative communication, for ultra-dense networks, to tackle the interference of small cells and achieve maximum net-profit. While authors of [72] have studied some of the caching techniques in 3G, 4G and 5G networks along with deployment locations within the network, and have proposed edge-caching scheme based on the concept of content-centric networking, highlighting the role of NFV in enhancing the benefits brought by the caching techniques in mobile networks. Authors of [73] have explored the cooperative edge caching in large-scale user-centric mobile networks, proposing greedy content placement algorithm based on the optimal bandwidth allocation, to reduce the average file-transmission rates. They have concluded that the proposed algorithm can reduce the average file-transmission delay by up to 45 percent compared with the non-cooperative caching and hit-ratio-maximal schemes. In [74], researchers have

proposed collaborative caching scheme for 5G network, where they investigated the impact of different constraints on their cache scheme, focusing on the placement and retrieval of highly popular contents. They claimed that, their proposed scheme could improve the performance of the 5G wireless network, via minimizing cumulative transmission delay, and increasing local caches hit ratio within the hotspot. Authors of [75] proposed a novel caching approach to achieve lower traffic compared to the traditional caching schemes, which considered using a single multicast transmission for multiple users whom are requesting the same files over short period, rather than sending the same files separately. They found that their caching policy reduces the cost by up to 52%. Meanwhile, other research works were looking into leveraging SDN and NFV technologies in mobile networks. Authors of [72] have performed an architectural survey of mobile networks that have utilised SDN/NFV technologies. They summarised some of the benefits by utilising the SDN/NFV technologies such as, reducing the network Capital Expenditure (CAPEX), OPEX costs and control signal load, automating scaling of resources and flexing services agility. While authors of [76] presented the architecture for edge caching, which utilises service function chaining and NFV technology. They have proposed caching framework in the mobile edge that enables the UEs to retrieve the requested content by dynamic collaboration between the virtualised RAN functions and the edge services.

5.4 Caching placement

There are many proposed locations for deploying content caches within the mobile networks such as, local caching, device caching, small cell caching, Macro-cell caching and core caching [72].

5.4.1 Traditional deployment

Deploying cache servers anywhere within the mobile network reduces the latency relatively for the end user. The closer the content to the users, the lower is the latency. However, there are other factors that are involved in the decision of selecting the most efficient deployment location, such as, number of served UEs, available cache size etc. The traffic latency is proportional to the time spent by each packet through the links of the mobile network. Therefore, the total transmission time for the downlink propagation delay is a result of adding the times spent in each link along the path from the external content server to the end user, as can be expressed in Equation 5.1 . [77]

$$T = T_{transport} + T_{core} + T_{backhaul} + T_{radio} \quad 5.1$$

Where $T_{transport}$ is the time spent by the packets in the links from the content server on the cloud/internet to the core network, T_{core} represents the processing time for the packets within the core network, which is contributed by different network entities (e.g. SGW, PGW). $T_{backhaul}$ is the transmission time of the packets over the connection links between the core network and eNBs/gNBs, whose value varies depending on the link type (e.g. copper wires, optical fiber, microwave links etc.). T_{radio} represents the packets processing time at UE/eNB, propagation delay and transmission/retransmission between the eNB/gNB and UEs. The aforementioned cache locations have been proposed to reduce the latency and the cost of the links. MNOs' OPEX include the cost of links' rental (backhaul links, transport links). Deploying cache servers at the core network minimises the amount of traffic over the transport links, which reduces the need for renting large bandwidth links and eventually reduces the cost. Similarly, deploying cache servers at eNBs or small cells reduces the rental cost of backhaul and transport links. In summary, the optimal solution for the cache deployment location should consider reducing the delay as much as possible, while maintaining high server utilization.

5.4.2 Caching within IoRL gNB

Mobile networks are leveraging content caching techniques, as a means for reducing latency, optimising network performance and enhancing user experience. Content caching at eNBs/gNBs could potentially reduce the backhaul capacity requirements by up to (35%) [78]. Deploying cache servers within small cells are predicted to become an important research topic due to its potential for performance enhancement [72]. Deploying a caching system at the small cell achieves the highest content-user proximity, i.e. users receiving cached contents within small cell, experience the lowest latency compared to cached contents elsewhere in the mobile network. On the other hand, small cells serve relatively small number of UEs, which means, small cell caching deployment is not so efficient in terms of the number of served UEs. Therefore, IoRL-C system is proposed to address the limitation in the number of served UEs by a small cell caching system. IoRL-C is an edge caching system designed for IoRL small cells; it is a form of Mobile Edge Computing (MEC), where it merges the Information Technology (IT) with the telecommunication networking to enhance users' QoE. IoRL-C utilizes the smart-switching feature of the SDN networking, to serve more UEs under multiple IoRL small cells by cloud-based centralized caching service. In summary, in traditional caching solutions, caching servers could be either deployed locally, thereby offer the lowest latency, while suffer lower servers' utilization, or deployed centrally,

thereby improving server utilization with the expense of degrading latency parameter.

The proposed IoRL-C provides an optimal caching solution for IoRL gNBs, by enabling the deployed MEC-caching servers to increase the amount of served UEs significantly with an efficient trade-off between latency performance and resource utilization.

5.5 Network Architecture

This section provides an overview about the IoRL small cell and IoRL-Cloud architectures, as well as providing the details of the operational procedures concept of the proposed service, IoRL-C that is.

5.5.1 IoRL-small cell architecture

In this subsection, the IoRL architecture is re-introduced in order to present the proposed solution with consistency. IoRL is an architecture for smart indoor 5G gNB, which utilizes VLC along with mmWave technologies to communicate with UEs under its coverage. Indoor gNBs designed based-on this paradigm are referred to as IoRL small cells or VLC-gNBs [79]. IoRL gNB architecture is depicted in Figure 5-3.

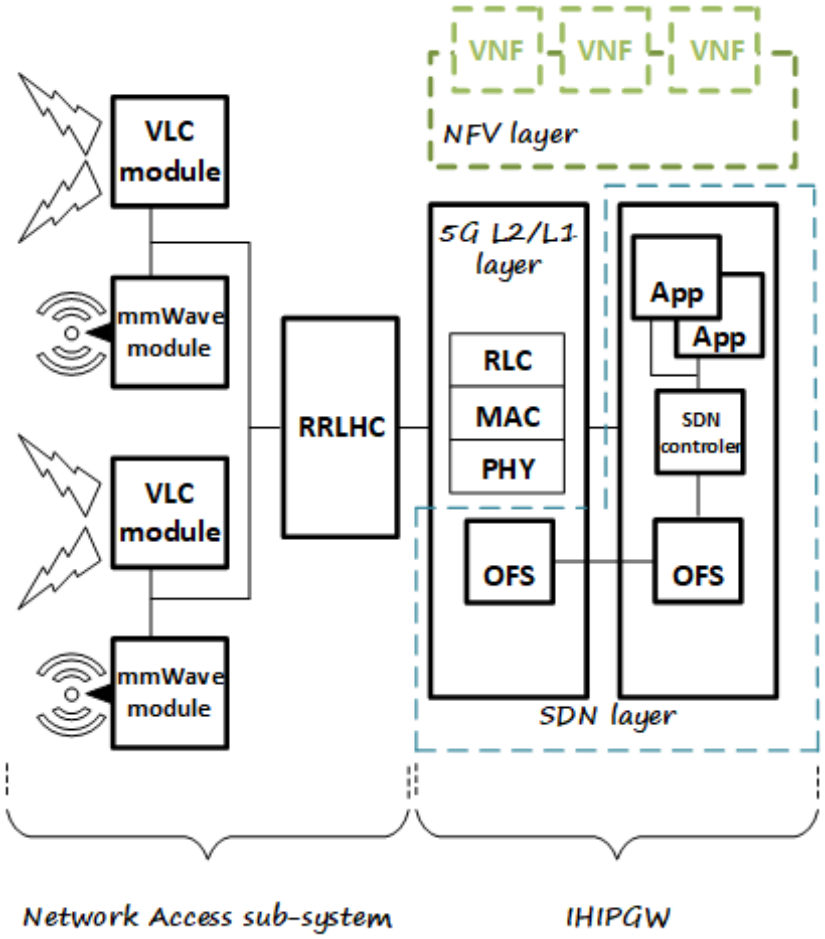

Figure 5-3 IoRL gNB architecture

IoRL architecture comprises of two sub-systems, Radio- Access (RA) sub-system and IHIPGW sub-system. RA comprises of Distributed Radio Access Network (D-RAN) connecting multiple Remote Radio Light Head Controllers (RRLHCs), where each RRLHC drives up to eight mmWave and VLC modules. The 5G L1/L2 processing is performed within the IHIPGW unit. The IHIPGW is built by utilizing COTS general purpose hardware [80]. The virtualization technology has enabled creating multiple services in the form of VNFs that serve different purposes. UEs that are served by IoRL small cells can enjoy some of the latest emerging location-based multimedia services, such as Follow Me Service (FMS) [81]. Although traditional small cell deployments provide network coverage with improved QoS parameters for users, they are incapable of offering other services for their clients, due to lack of intelligent switching ability and application layer entities (e.g. servers), while IoRL small cells overcome the aforementioned limitations by exploiting the intelligence of the SDN networking (smart switching) and virtualization technology (e.g. Virtualized-Servers).

5.5.2 IoRL-Cloud architecture

The IoRL-gNB and the IoRL-Cloud system design is depicted in Figure 5-4. IoRL small cell uses SDN network to connect the lower layers (L1 and L2) to the IoRL-Cloud service via an ideal backhaul.

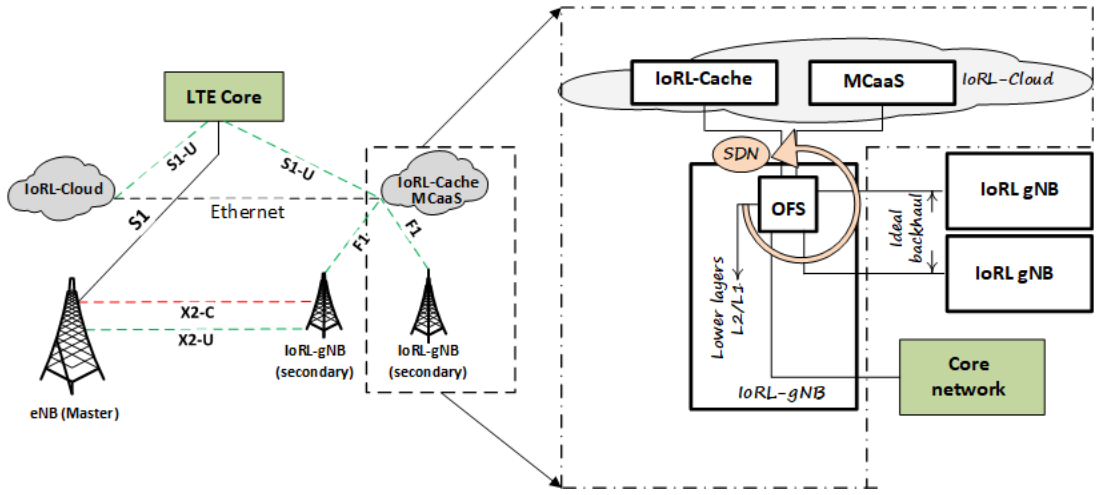

Figure 5-4 IoRL-Cloud system design

Figure 5-5 depicts the IoRL-C service design, which comprises of Open Flow Switch (OFS), Virtual-Cache servers (V-Cache), SDN controller and IoRL-C application.

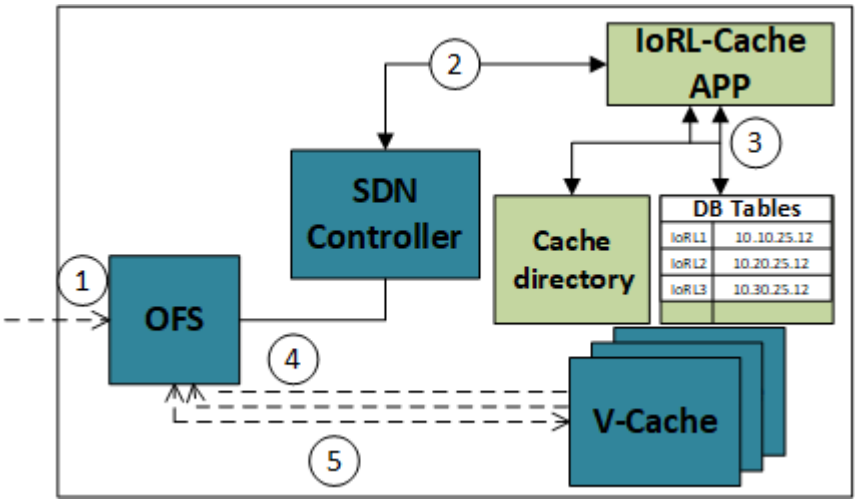

Figure 5-5 IoRL-Cache service design

When a user sends a request, the OFS forwards the first packet of this new flow to the SDN controller as a packet-in event. The IoRL-C application unwraps the packet and performs packet classification to identify the type of the traffic. If the traffic is classified as non-cacheable then, the controller configures the OFS to forward similar traffic towards the LTE-Core. Otherwise, the application looks in the cache for the location of the requested content. If the requested content is found locally then, the SDN controller configures the OFS to forward the request for this content to the address of the correct local V-Cache server. Alternatively, it checks if the

cache directory has got the address of the next IoRL-C servers that caches the requested content. If so, the SDN controller configures the OFS to forward the flow to the correct IoRL-C. The operation flowchart is depicted in Figure 5-6.

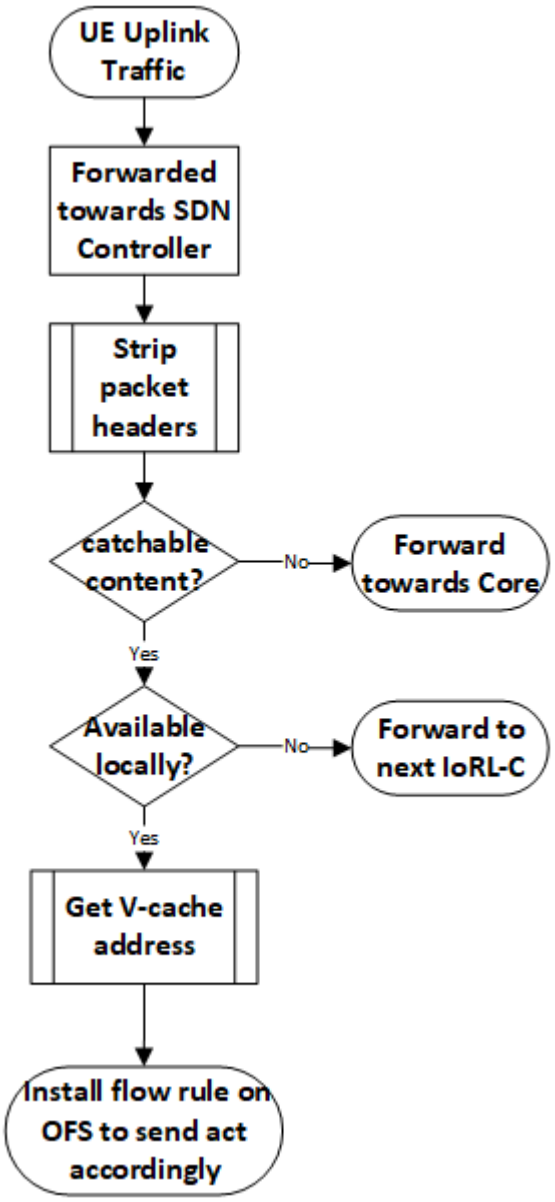

Figure 5-6 IoRL-cache operation flowchart

5.6 Performance evaluation

A hypothetical mobile network has been simulated by using OMNeT++ simulation tool, to obtain performance indicators. This section provides an overview about the network setup and configuration.

5.6.1 Network setup

The network setup is depicted in Figure 5-7, where the simulated network consists of five eNBs, four IoRL gNBs, IoRL-C unit, LTE core and external media

server/broadcaster. Two eNBs (eNB1 and eNB2) include local cache servers, UEs under the coverage of eNB3 and eNB4 receive contents cached at the core network, while UEs covered by eNB5 are assumed to be receiving content from the external media server without the involvement of caching servers, each eNB serves 400 UEs. IoRL small cell-1 contains a local cache to serve its UEs, while IoRL small cells 2, 3 and 4 are connected to a centralised IoRL-C platform at the mobile edge cloud, where each IoRL small cell serves 100 UEs. During the simulation,
UEs are configured to request video traffic and are being served from the closest available cache server according to the aforementioned network setup. It is worth mentioning that the requested contents assumed to be cached to avoid simulating miss-cache events. The link between the IoRLCache and IoRL small cells is optical fibre, the delay induced
by the optical fibre link is related to the length of the link, and it is calculated by:

$$Link\ Delay\ (s) = \frac{Distance}{C/_{RI}} \qquad 5.2$$

- Distance = link length (in meters), which refers to the separation distance between the IoRL-C and IoRL gNB.
- $C = 3 \times 10^8 (\mathrm{m/s})$, which is the propagation speed inside the optical fibre link.
- RI is the refraction index, which is equal to 1.5 for the optical fibre.

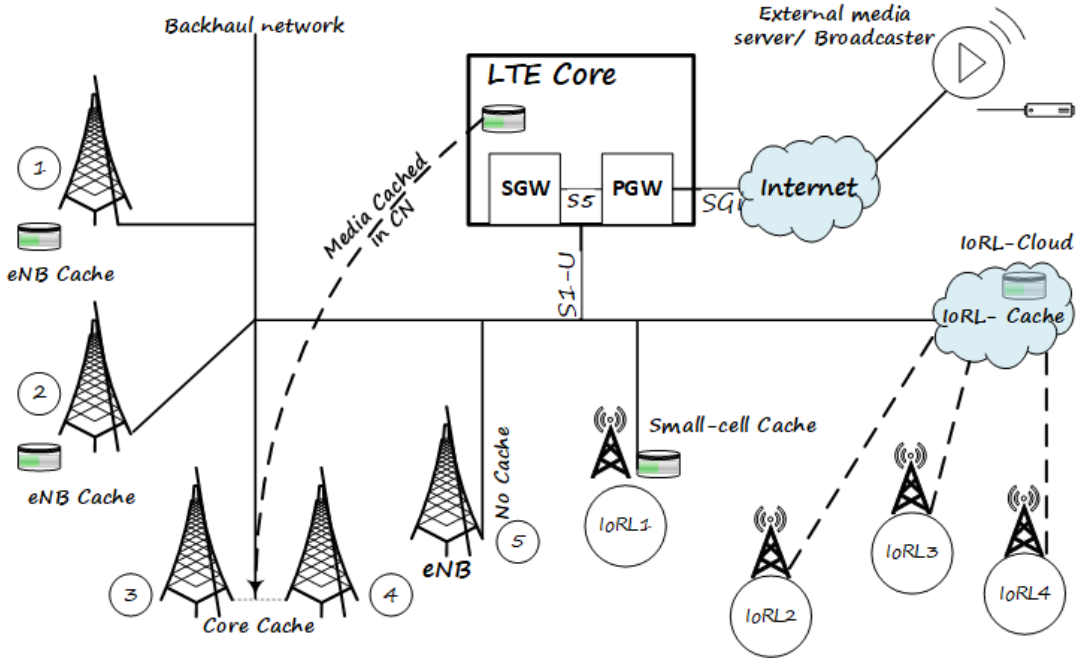

Figure 5-7 Network setup

5.6.2 Network configuration

The network scenario setup of the performed simulation is depicted in Figure 5-8. The LTE core network comprises of data plane components namely (SGW and PGW), set of (eNBs), set of indoor small cells, denoted by (IoRL), IoRL-Cache servers denoted by (SDN-cloud), set of users denoted by (UEs) and external media server (Server) connected to the core network via ISP denoted by (router).

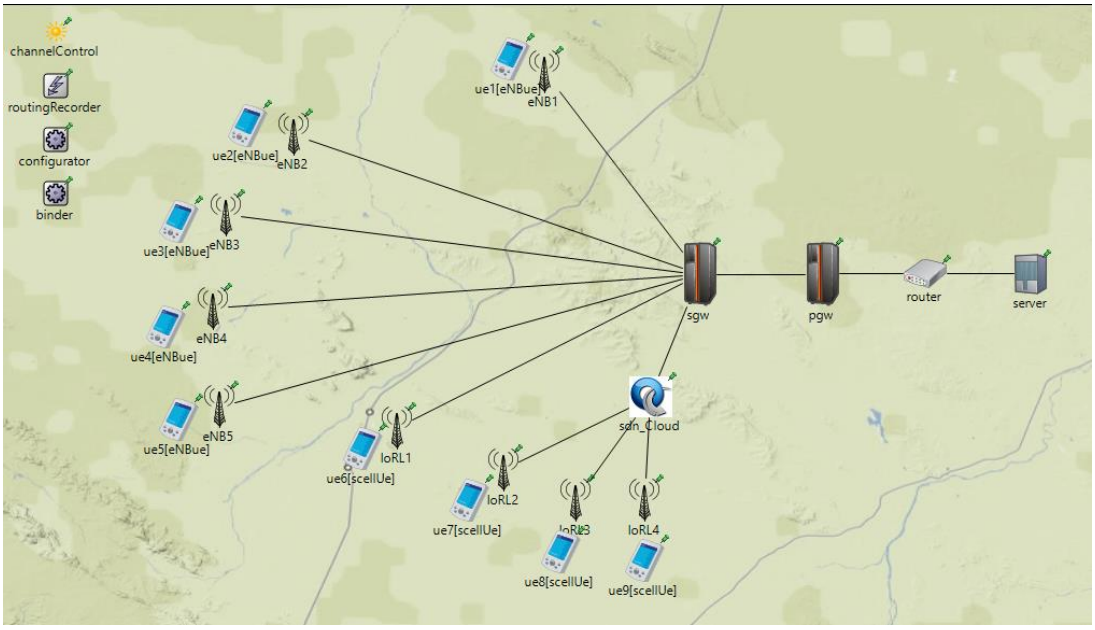

Figure 5-8 scenario setup

The simulated EPC consists mainly of SGW and PGW. The core cache is assumed to be deployed within the PGW. The network entities and links add delays for the traffic to reflect a realistic flow delays, specifically, 4ms for the S1-U interface, 1ms for the S5 interface and 4ms for the SGi interface. The access network consists of five eNodeBs, four IoRL small cells and approximately 2400 UEs. The eNodeBs are equipped with omnidirectional antennas of 40 dBm transmission power each, 2 dB cable loss, 5 dB noise figure and 18 dB antenna gain. The RLC layer at the eNodeB is configured with unacknowledged Mode, with fixed PDU size of 40 bytes. A realistic channel model that considers both path loss and fading is being used in the simulation. The Urban Macro Path Loss Model and the Jakes model for Rayleigh fading is used in all scenarios. The UEs are configured to use 26 dBm transmission power in stationary position throughout the test. The network was implemented using OMNeT++ version 5.0 utilising its independently developed open source INET 3.4 library together with simuLTE. OMNeT++ runs on linux operating system hosted by a DELL PRECISION Tower equipped with Intel Core i9-7900X/3.30 GHz CPU with 128 GB of RAM. The parameters included in table 5.1 below. It is worth mentioning that the performed simulation does not require such powerful machine

to perform the simulations, but this machine already exists as part of the equipment set for the IoRL EU project, therefore was used in the simulations. INET library and the aforementioned extensions offered flexible tools for creating the simulation platform, which in turn was used to evaluate and quantify the efficiency and performance of the proposed cache architecture.

Parameter	value
C	3x10^8 m/s
RI	1.5
No. of small cells	4
No. of eNBs	5
No. of UEs covered by small cell	100
No. of UEs covered by eNB	400
Total number of UEs	2400
S1-U delay	4ms
S5 delay	1ms
SGi delay	4ms
PDU size	40 bytes
OMNET++ version	5
Host specifications	Intel Core i9-7900X/3.30 GHz CPU with 128 GB of RAM

Table 5.1 simulation scenario parameters

5.7 Results and analysis

This section presents the various test scenarios, as well as analysing the obtained results critically in order to evaluate the efficiency of the proposed service and its impact on the mobile network architecture of the next generation network.

5.7.1 End-to-end delay scenario: a comparison of the experienced delay (no-cache, VS cache deployments)

Multiple runs were performed for the hypothetical network scenarios, with cache servers deployed at multiple locations. Figure 5-9 illustrates the obtained average delay experienced by UEs receiving contents cached at different locations. Based

on the obtained results, deploying cache at the small cell provides the lowest latency (below 2 ms).

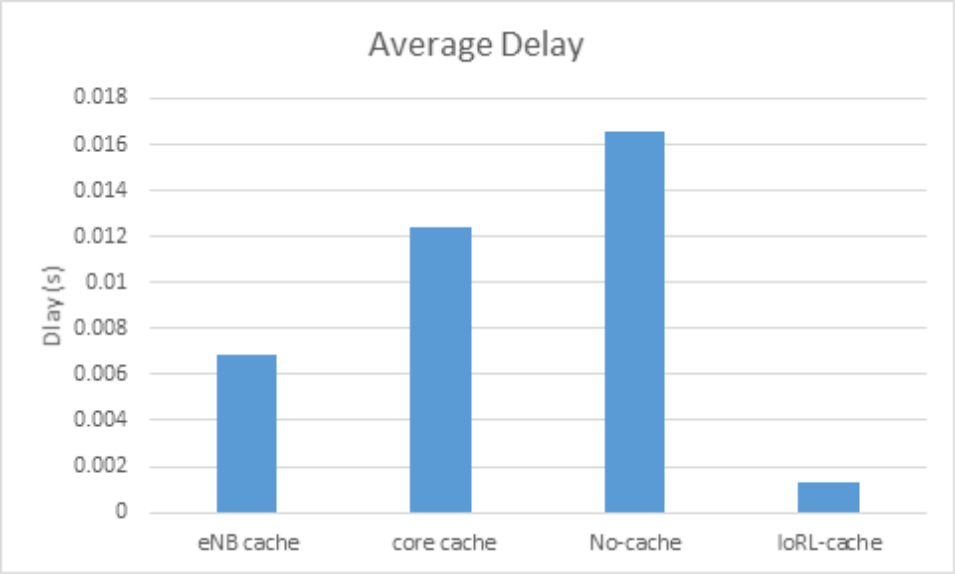

Figure 5-9 average content delay cached at different locations

Deploying a cache at the core network reduced the latency by 25% compared to non-cache scenario, similarly, deploying cache servers at the eNBs reduced 59% of the latency by eliminating the delays caused by backhaul and transport links. Figure 5-10 depicts the average delay experienced by UEs at different eNBs. UEs under the coverage of eNB1 and eNB2, receive locally cached contents with delays below (8ms), while UEs under the coverage of eNB3 and eNB4 receive contents from proxy-cache servers at core network, experience more latencies (above 12ms). UEs connected to eNB5 receiving non-cached contents have experienced the highest delay amongst the other UEs.

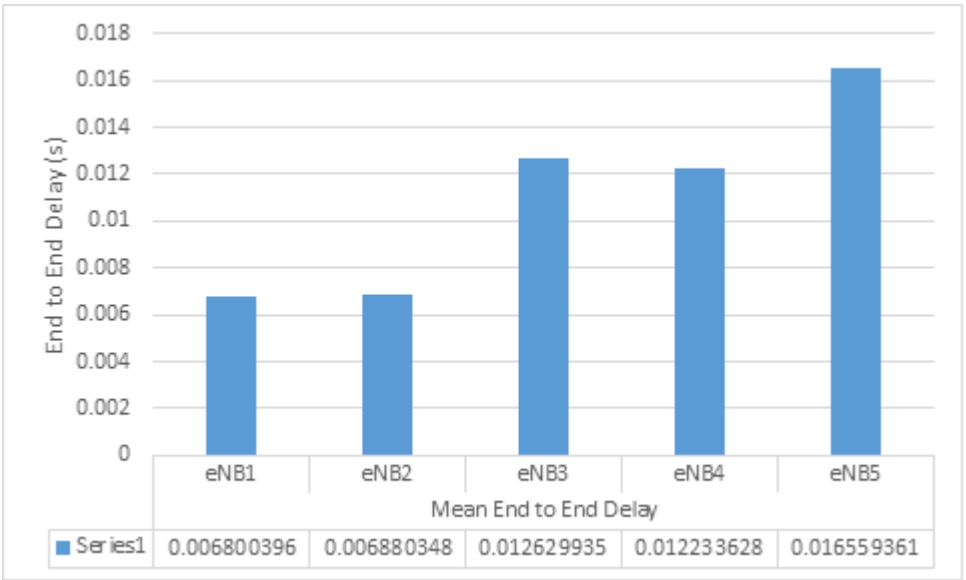

Figure 5-10 mean delay for UEs connected to eNB1-5

5.7.2 Cloud-cache scenario: a comparison between small cell cache and IoRL-C

The overall delay at IoRL UEs was measured during the simulation runs. IoRL small cells with cloud-cache deployment were compared to IoRL small cells with local-cache deployment. Figure 5-11 depicts the average delays experienced by UEs under the coverage of small cells. The delays were all below (1.5 ms), with little differences. The delays for UEs receiving from IoRL-C servers were about (0.1 ms) more. This is due to the delay of the link between the IoRL-C servers and the individual IoRL small cells.

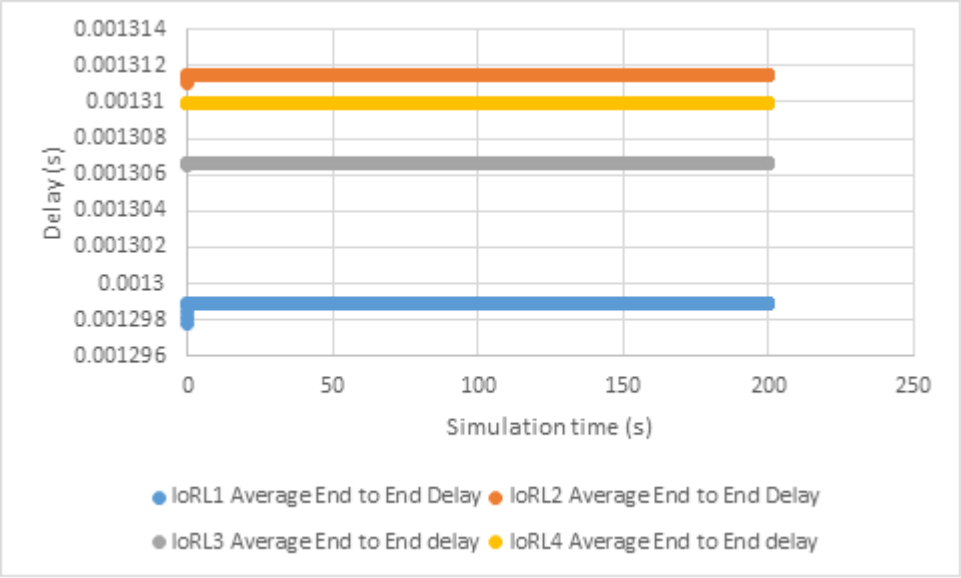

Figure 5-11 average UEs delay for small cell cached contents

The measurements were based on 50Km link length. The other runs were performed with different link lengths between the IoRL-C and the IoRL small cell. Figure 5-12 depicts the different delays obtained, where the link delay was measured in ONMeT++ based on equation 5.2.

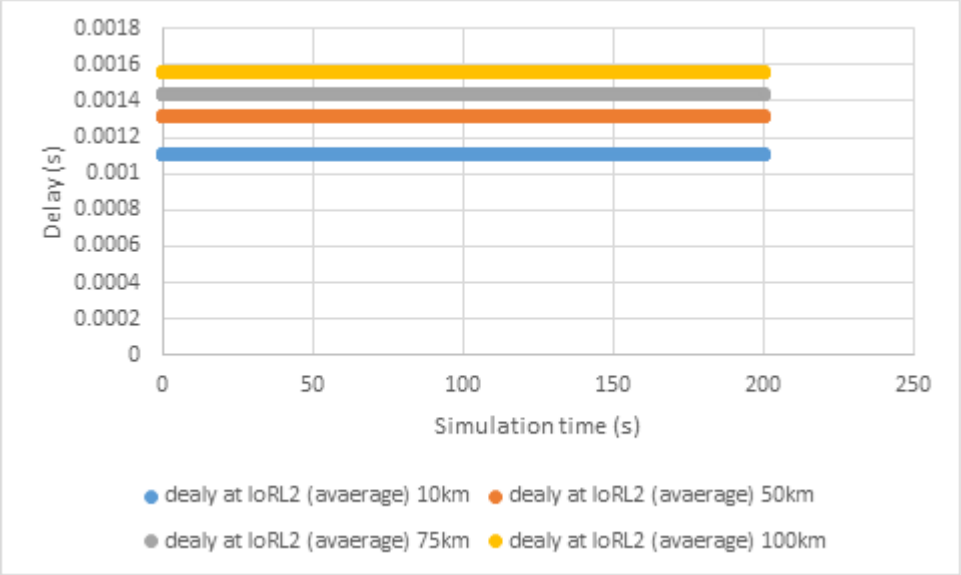

Figure 5-12 average delay for IoRL small cell at various distances

Based on the measured delays, the biggest delay difference between small cell cache and IoRL-C is (0.5ms). However, for the current deployment there is 300% improvement by server utilization. Figure 5-13 compares the number of served UEs by both cache deployments. IoRL-C can serve multiple IoRL small cells located relatively remotely from the IoRL-C. The obtained results, highlight the possibility for deploying IoRL small cells at lower costs, by relocating the application layer services to a logically-centralized location (i.e. IoRL-cloud).

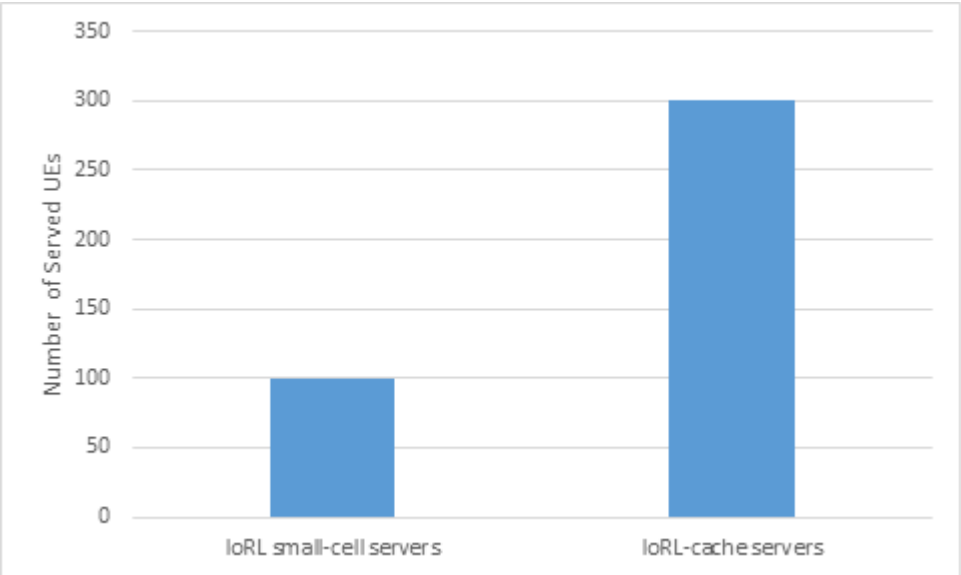

Figure 5-13 number of served UEs by small cell and IoRL-cache servers

5.7.3 Link utilization scenario: a comparison between link loads at different cache deployments

Three scenarios were implemented, with content/cache servers deployed at three locations namely a. content server (outside the mobile network), b. proxy-cache server (at core network) and c. local cache server (at eNB/gNB). Figure 5-14 depicts the links utilization at different cache deployments.

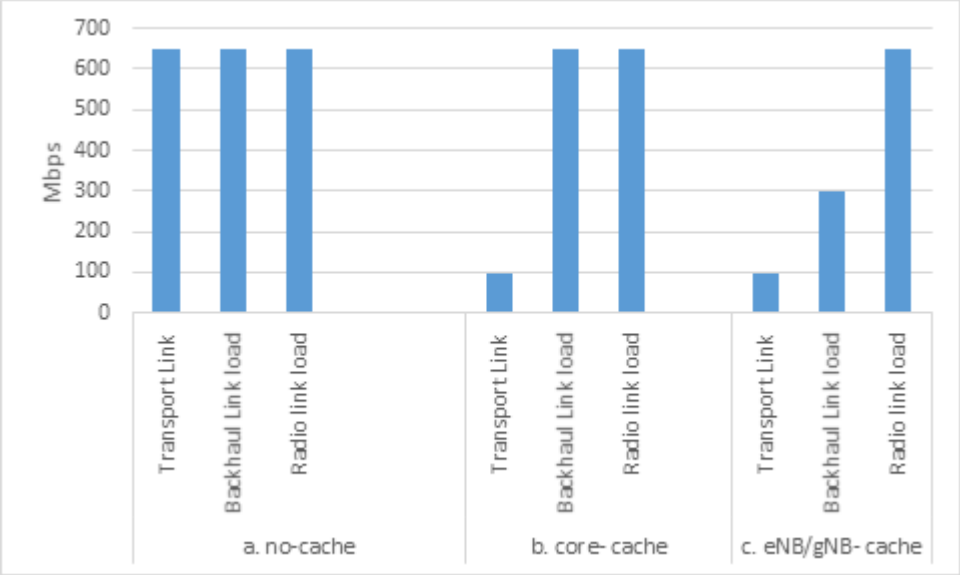

Figure 5-14 link utilization a. No-cache, b. core-cache, c. eNB/gNB cache

Figure 5-15 a shows links utilization without the deployment of cache servers within the network, where all links utilize their maximum capacity to cope with the load requirement for the UEs. Figure 5-15 b shows significant reduction in transport link utilization due to deploying core proxy-cache servers. While Figure 5-15 c shows the saving on both links (transport and backhaul) by deploying local cache servers. Reducing the links utilization means saving the cost of the links, by reducing the need for renting links with large capacities.

5.7.4 Analysis

The obtained results support the "network of services" concept, which shows deploying a caching service at the mobile edge in the form of VNF brings the optimal network performance. As shown in Figure 5-10 – 5-15, deploying cache servers at different locations, reduced the overall latency relative to the un-cached content architecture. Referring to equation 5.1, deploying cache at the core network, eliminates the $T_{transport}$ and makes the overall downlink time equal to: $T = T_{core} + T_{backhaul} + T_{radio}$. Similarly, deploying cache servers at the

eNB/gNB, eliminates the T_{core} and $T_{backhaul}$, therefore, the overall delay of the traffic becomes $T = T_{radio}$. For the current test scenario, the distribution of UEs is based on serving cache servers shown in Figure 5-15. Proxy-cache servers deployed at the core network, serve far more UEs than local servers deployed at eNBs.

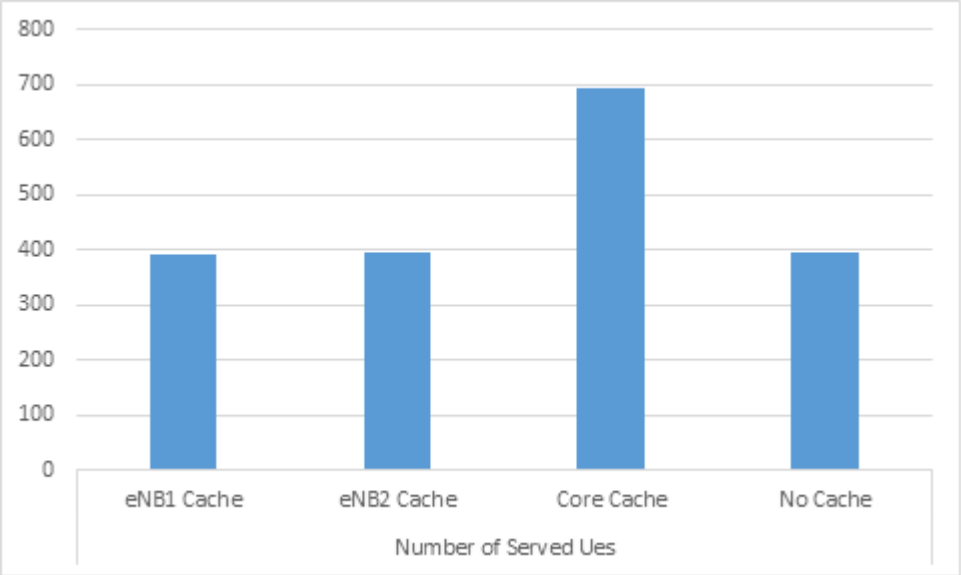

Figure 5-15 number of UEs served by different cache servers

However, the downlink time for contents cached at core servers equals to $T = T_{core} + T_{backhaul} + T_{radio}$, while, deploying the servers at the eNB/gNB brings the downlink time to $T = T_{radio}$ only, yet serves the lowest number of UEs. Therefore, the optimal solution is to deploy cache server as close as possible to UEs, while maintaining high server utilization. IoRL-C offers a good trade-off between latency reduction and resource utilization that satisfies efficient network performance and resource utilization. All the obtained results are based on the performed test scenario and changing the test scenario could lead to different results; however, it will only change the numerical figures not the concept. For instance, deploying cache servers at eNBs will always reduce the cost of backhaul links and transport links regardless of the number of deployed eNB with caching servers.

5.8 Conclusion

Cache servers are being deployed within the mobile networks at various locations to enhance the network performance for both of parties, end users and network operators, by reducing the contents latency for UEs and OPEX for MNOs respectively. IoRL is an indoor small cell solution to enhance the network QoS and QoE for UEs. In this chapter, IoRL-C was introduced as a MEC-based caching solution for IoRL gNB deployments. It exploits SDN and NFV technologies to

provide caching solution for MNOs that is close to the UEs in flexible efficient deployment. IoRL-C provides an efficient trade-off between reduced latency and resource utilization, thereby, supporting the concept of deploying IoRL as 5G solution for indoor environments. Based on the obtained results, IoRL-C provided 300% server efficiency improvement compared to local cache server deployment.

6 Virtual Gateway: Mobility management for 5G Internet of Radio Light gNB

6.1 Introduction

Mobile networks are evolving constantly by incorporating new technologies to cope with the exponential increase in user's demands for higher data traffic, which is expected to increase to 20000 times by 2030 [82]. Researchers have discussed the evolution of the current mobile networks towards more flexible networks by utilizing new concepts such as software defined networking, virtualization and slicing. These technologies enable the emergence of network of functions instead of the current network of entities [41]. The main targets for mobile network operators and other service providers are, increasing the capacity, improving the data rate and expanding the service coverage of the wireless network. There are different approaches for achieving these targets, such as deploying Cloud Radio Access Network (C-RAN) and deploying small cells within the RAN of the MNO.

C-RAN solution aims to maintain the coverage area while merging the baseband processing units for number of cell sites into centralized unit, thereby reducing the capital expenditure and offer better services [83]. C-RAN network architecture enables dynamic reconfiguration of the computing and spectrum resources, for instance, by reducing the processing runtime of the Based Band Units (BBUs), this enables longer distances between BBU-pool and the antennas, which in turn, improves network capacity and design flexibility [84]. However, the main drawback of the C-RAN network architecture is the need for high front-haul links capacity [85].

On the other hand, small cell deployment is another solution for improving network performance. Figure 6-1depicts an example of the deployment of small cells within RAN in LTE mobile networks.

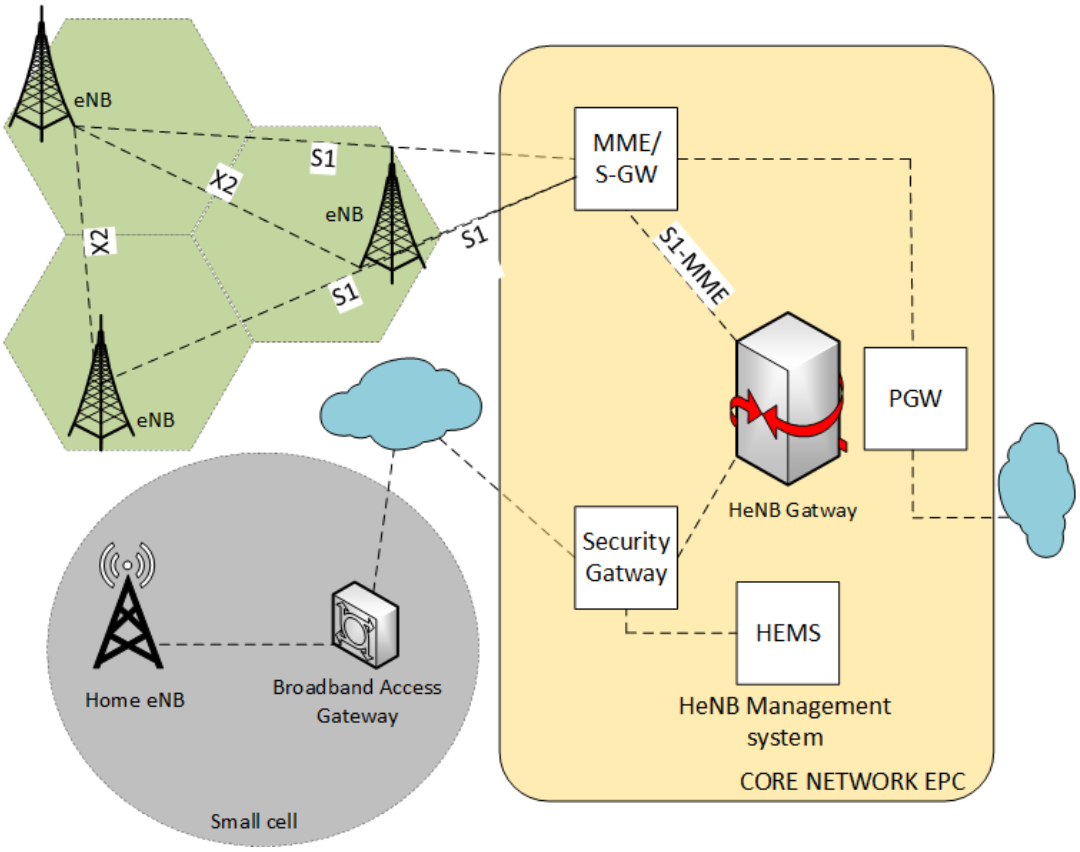

Figure 6-1 small cell deployment

Small cells target the improvement of system capacity for indoor or local area (outdoor) networks with little cost for network operators [86][87]. In fact small cells solution is becoming more interesting because it targets indoor environments, where data traffic usage represents (70-90 %) of the overall data traffic [88][89]. Yet, there are many challenges related to the deployment of small cells, e.g. signal interference as well as increased overhead signalling due to inter and intra smallcell – macrocell handovers. Some solutions proposed to overcome these challenges, such as optimizing handover decisions by considering context parameters (e.g. speed, channel gain…etc.) into handover algorithms, thereby reducing the overhead signalling [84]. Other solution proposed utilizes SDN-based handover decisions, to reduce the number of required handovers and the overhead signalling, which in turn increase throughput [90]. At the meantime, works based on dual connectivity framework [86] provide solutions to optimize the handover in small cells deployment, where UEs are able to perform fast switching for fast handover among small cells and eNBs [87]. In view of these ongoing research activities, the aim of this chapter is to propose the Virtual Gateway (VGW), which is a virtual network entity that enables efficient IoRL deployment, the objectives being, providing detailed analysis about the signaling involved in the mobile operations, to appreciate the role of the VGW in IoRL deployment and the benefits

it offers. VGW can potentially improve the network performance for more than three quarters of the data traffic (indoor data traffic). VGW provides processing offloading as well as reduces the essential handover signalling. VGW is designed specifically for IoRL indoor small cells and it is deployed on a platform that hosts multiple services that are available as add-on features for a cluster of IoRL smallcells. i.e. promotes resource slicing.

6.2 Related work

Researchers from academia and industry are always investigating into new approaches to achieve better resource utilization and network performance. Fifth Generation networks are densely deployed smallcell networks, which bring challenges into network architecture, management and resources utilization. Mobile networks include various technologies and mechanisms to address the aforementioned challenges e.g. Cloud-RAN, small cells, SDN/NFV technologies … etc. C-RAN provides a low-cost front-haul wireless network, where the central base station communicates with distributed antennas of multiple small cells [88]. In [89] authors presented the structure of C-RAN, highlighting the existence of service cloud and the separation of the Remote Radio Heads (RRHs) from the Base-Band Units (BBUs), which enables efficient processing of traffic from hundreds of RRH in a centralized BBU.

On the other hand, C-RAN requires high speed front-haul links (usually optical fibre) and does not tolerate latencies, which contributes to higher costs of Operational Expenditure (OPEX) due to high bandwidth front-haul links rentals. Meanwhile, researchers working on enhancing the coordination among radio and transport resources, to achieve better network performance, in this regards, authors of [84] presented an SDN-based cross-domain (across radio and transport networks) orchestration architecture. They have pointed out the benefits of the proposed orchestration for both infrastructure providers and service providers in terms of service agility and efficient resource utilization.

Network slicing, is another important concept being leveraged by mobile networks, since it enhances resource utilization. Network slicing along with software defined networking, management and orchestration, are the enablers for designing mobile networks of functions rather than networks of devices [90]. In [91] and [92], authors discussed the mobility management schemes over sliced-resources networks, and they proposed changes to the handover mechanisms to support smother handovers. As previously mentioned, small cells are designed and deployed to improve indoor/local area system capacity and improve network

performance in terms of handling the traffic overloads at acceptable costs [93]. The successful deployment of small cells depends on the integration of these cells with lowest overhead and cost [28]. Therefore, researchers proposed various approaches for utilizing small cell concept, for instance, authors in [94] have proposed enabling a group of selected UEs to play the role of virtual small cell BS for other UEs, for the purpose of increasing network capacity while reducing the number of communication links and improving resource utilization. Researchers in [95] have proposed modifying the mobile network architecture by introducing OF protocol, thereby enabling efficient programmability of the network entities, which enhances the network management ultimately. Work presented in [96] provided a survey of the OpenFlow-enabled architectures, and provided detailed analysis of the common operational procedures, they highlighted many benefits of the SDN-based network architecture namely, signalling reduction, architecture flexibility and simplifies configuration and network management.

6.3 System architecture

VGW-based architecture is depicted in Figure 6-2, where the secondary cells are set of gNBs designed based on IoRL indoor gNB [97]. The control plane of the Secondary gNBs (SgNBs) is incorporated in a logically centralized SDN controller while the data plane forwarding is handled by the SDN forwarding devices. The data plane traffic is transported via the network, which consists of a set of forwarding devices and VGW-D. VGW-D is connected to the Core 5G Network (CN), to traverse the ingress and egress traffic in both Uplink (UL) and Downlink (DL). Unlike the traditional DC network architecture, where the Master eNB (MeNB) is responsible for the control plane processing for all SgNBs in its coverage area [98], in VGW-based architecture, some of the control functionalities are offloaded from the MeNB, to run as a separate software applications on top of the controller, in a virtualized platform. This set-up is designed as plug-and-play add-on to the traditional MeNB architecture. It contributes to providing an easier upgrade process of individual services without the need to upgrade the entire physical device. The VGW entity comprises of multiple slices that perform various services utilizing the same physical hardware, to achieve optimum resource utilization. The signalling messages have been restructured and formatted to be sent from the SDN controller using southbound protocol [99].

Contrary to the traditional DC architecture in the case of CN split, VGW-based solution does not require the involvement of the CN as part of the intra-SgNB, which enhances the network performance in terms of delays. VGW solution

eliminates the need for signalling messages for the CN to update the traffic tunnels. VGW-architecture enables all of the Indoors SgNBs to appear as a single IoRL SgNB to the rest of the network entities. To achieve a conceptual understanding and to quantify the performance enhancement by the proposed architecture, a mathematical model is employed as shown in later section of this chapter.

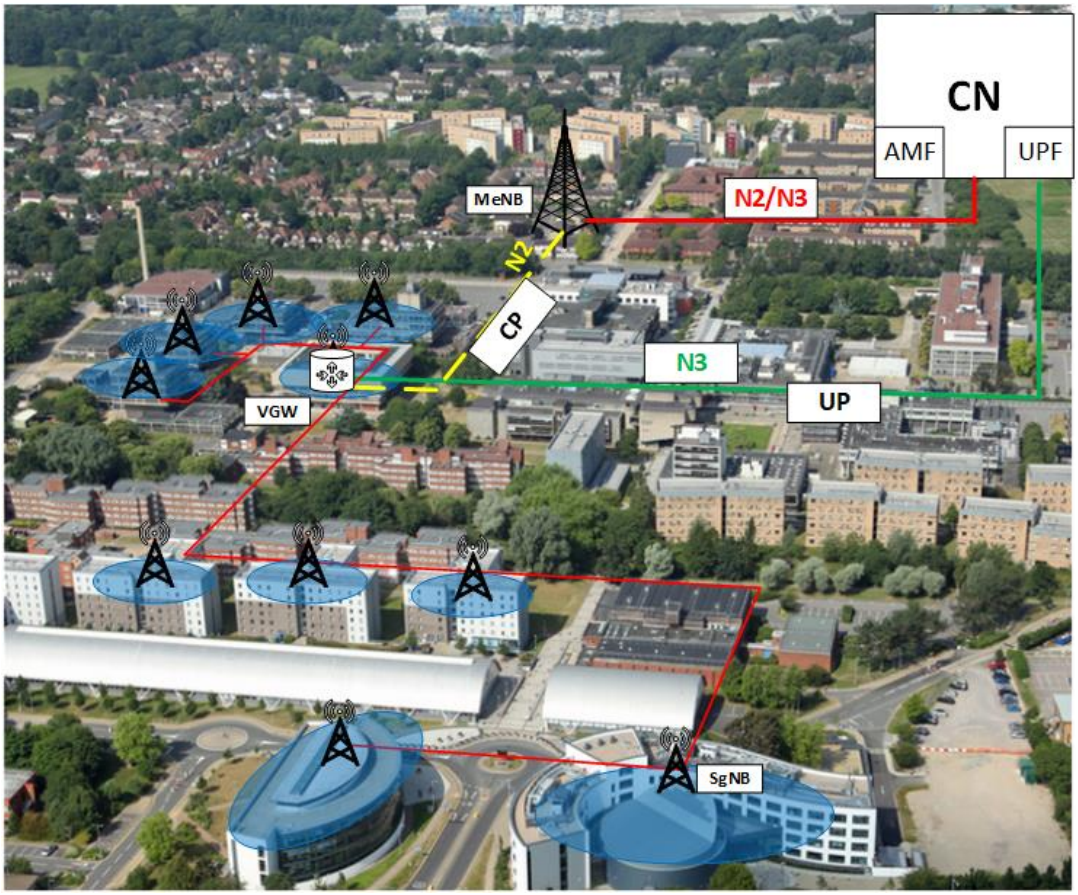

Figure 6-2 VGW-based network architecture

6.4 Architecture description

This section describes the components involved in the proposed deployment solution for the IoRL gNB, which includes Master cell next generation eNB, Secondary cell IoRL gNB, control components e.g. SDN controller, VGW-C and data plane components e.g. VGW-D, IoRL Forwarding Devices (IoRL- FD). The next sub-sections provide detailed description about the aforementioned components.

6.4.1 Master cell eNB

The master Cell eNB is assumed to be a next generation eNB (ng-eNB) [99], which is a network element of NG-RAN. The ng-eNB communicates with CN using New radio (N) interface. In the VGW architecture, the master cell ng-eNB keeps the same

functionality and radio protocol stack as specified by 3GPP standards. For simplicity, the master cell ng-eNB is referred to as (MeNB).

6.4.2 Secondary cell gNB

The secondary cell in this architecture refers to an improved version of the IoRL gNB [100]. The IoRL is designed for indoor coverage with environment-related services. Features include but are not limited to geolocation services, local offloading services, multimedia services …etc. [81]. IoRL gNB communicates with CN using N interface. There are three options for deploying IoRL gNB within mobile RAN according to the specifications in the standards [101]. The proposed VGW-architecture represents the optimum deployment option for IoRL gNB. VGW-based deployment is compatible with 3GPP standards, as well as preserves the unique features of the IoRL system design. For the sake of simplicity, the Secondary IoRL gNB is referred to as (SgNB). The SgNB is deployed with functional split option 2 [102].

6.4.3 Virtual Gateway

The virtual gateway (VGW) represents the centralised entity in the IoRL SgNB deployment architecture. It comprises of various slices dedicated for various services namely, caching services, geolocation services and essential network services. The use of SDN controller provides a flexible control plane and a programmatic platform for controlling the data plane FDs. It has a set of supervisory applications, such as, network monitor and resource optimization apps, which are responsible for maintaining a global view of the SgNB's network resources and distributing the loads across different links in the network (e.g. mmWave, VLC). Other information is maintained in the form of VNFs running on top of the controller, such as geolocation services subscription list, local databases and other virtualized services. The load balancer VNF relies on the information of the other applications such as the monitoring applications to establish links' status information and perform routing decisions. The available rich information is sufficient for establishing a detailed view of the links' status. Queries such as OFP-Table-Stats-Request, OFP-Flow-Stats-Request, OFP-Port-Flow-Request, OFP-Queue-Stats-Request, OFP-Group-Stats-Request, OFP-Meter-Stats-Request …etc. are used to acquire information about the forwarding devices and make them available for other services such load balancer and handover services.

6.4.3.1 VGW forwarding plane

The forwarding plane of the VGW architecture represents a set of OpenFlow switches, one in each IoRL coverage area, each of which consists of flow tables, group tables and meter tables. A remote controller (which is part of the VGW) controls each switch using the OpenFlow Protocol. A flow-table contains multiple flow entries; each flow-entry consists of:

- Matching rules, each match rule contains a combination of ingress port and L2/L3/L4 header fields, which may be variously wild-carded or masked.
- Each match rule contains one or more instructions attached.
- Counters for collecting statistics about the flows.

Meter tables are utilised to provide different QoS treatment for different types of traffic. This information is exploited by the load balancer VNF, to achieve an optimum link utilization achieved across the IoRL coverage area.

6.4.3.2 VGW control plane

The VGW control plane (VGW-C) is responsible for building the routing behaviour for the forwarding devices, to route the traffic amongst MeNB, SgNBs, CN and local services smoothly and efficiently. VGW-C offloads some of the control functionality performed by MeNB in DC mode. VGW-C exploits SDN and NFV technologies to create an add-on solution to be used for the deployment of the IoRL indoor small cells. In the view of the layered structure of the communication systems, VGW-C performs the processing of several layers, namely: Packet Data Convergence Protocol (PDCP) and Radio Resource Control layer. The services are deployed as VNFs to provide optimum flexibility and resource utilization. The use of VNFs for the processing enables for the upgrade of these services easily, e.g. adding AI in the handover processing, load balancing … etc. Another feature of the VGW architecture is the resource slicing as shown in the next subsection.

6.4.3.3 VGW slices

The motive behind VGW slicing is to achieve optimum utilization of the available resources. In the VGW architecture, the resources are sliced into three categories as depicted in Figure 6-3 network services slice, geolocation service slice and caching & streaming services slice. The management of resources, quotas and management policies for each slice is enforced by utilizing a VIM such as Openstack [52].

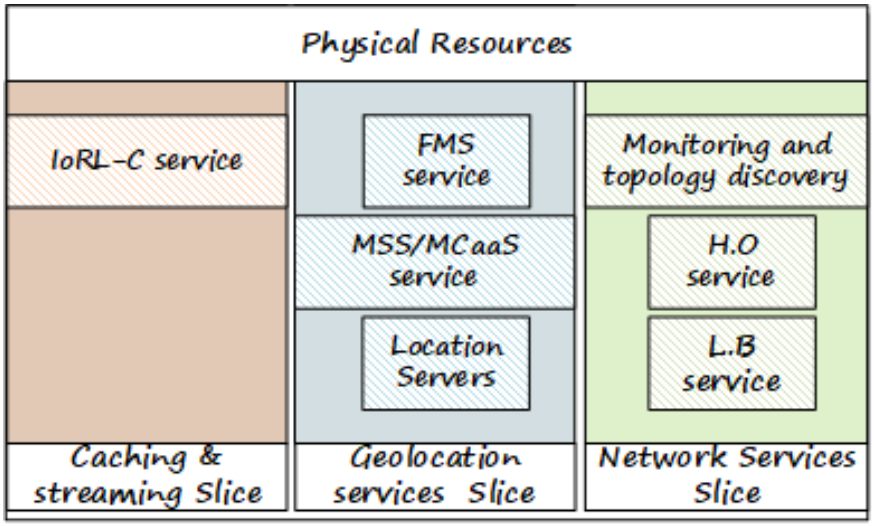

Figure 6-3 VGW slices

The network services slice comprises of MNO related services, e.g. load balancing monitoring and topology discovery, handover H.O. service between the available mmWave and VLC links. The load balancing service follows the MNO policies, which are preloaded to the service by the orchestration tools.

Similarly, Intra-SgNB handover process is performed locally by H.O service. The geolocation services slice hosts the various location database multimedia services e.g. FMS and Media casting as a Service (MCaaS) [81][103]. The other slice of the VGW resources is dedicated for deploying a local caching service to the UEs within the coverage area of the IoRL gNBs. IoRL-C as introduces in chapter 5 is an efficient caching solution that achieves a balanced trade-off between server utilization and latency performance.

6.5 Overview of the VGW layers

This subsection provides an overview of the network service slice in the context of the mobile network stack layers. Figure 6-4 depicts the overall VGW architecture, highlighting the network entities with their related layers to clarify the role of the VGW and the processes performed by its services. The VGW architecture is designed by exploiting SDN and NFV technologies, therefore, the processes implemented in the form of VNFs running on top of an SDN controller, by the help of VIM and IaaS tool.

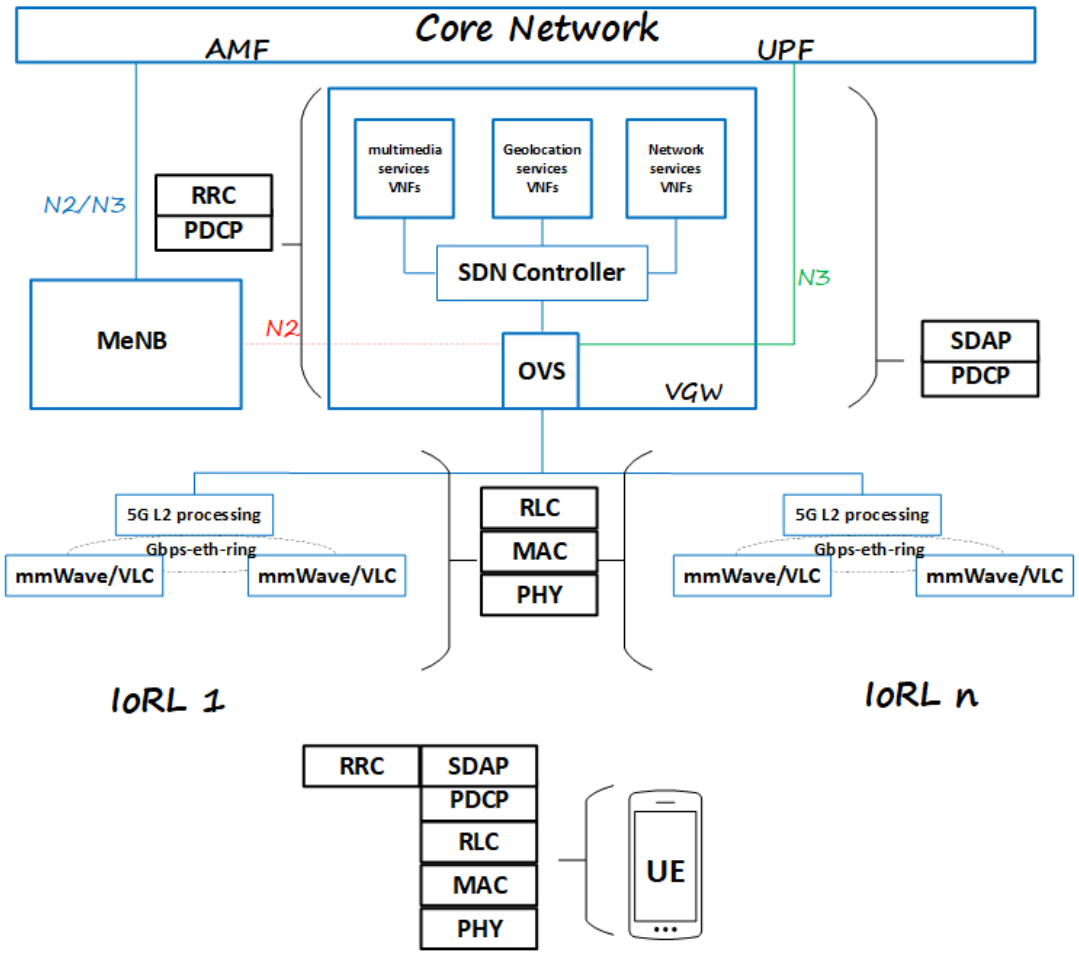

Figure 6-4 VGW-based architecture with its related mobile network layers

6.5.1 Service Data Adaptation Protocol

The Service Data Adaptation Protocol (SDAP) layer comprises of SDAP entities. Each UE might have multiple Packet Data Unit (PDU) sessions; each PDU session is assigned an SDAP entity. SDAP layer are designed to perform specific functions, mainly, transferring User Plane (UP) data, mapping QoS flows to Data Radio Bearers (DRBs) and identifying QoS flow IDs for Up-Link (UL) and Down-Link (DL) traffic, in order to perform reflective QoS mapping (applying QoS parameters for the uplink PDUs similar to the QoS obtained from the downlink Service Data Unit (SDU)). In the VGW-architecture, the process of the SDAP layer is performed at the VGW application layer. The SDAP of the VGW is logically centralized and processes the UP traffic for all the UEs within the coverage of all of the IoRL SgNBs. At the gNB, when the DL UP data SDU gets to the SDAP layer, SDAP entity constructs the corresponding SDAP PDU and submits it to lower layer, while in UL, the SDAP receives the SDAP data PDU from the lower layers, and the SDAP entity retrieves the corresponding SDAP SDU and delivers it to upper layers [104].

6.5.2 Radio Resource Control

The Radio Resource Control (RRC) layer is responsible for many CP services; it is responsible for many functionalities such as broadcasting of system information, RRC connection control and RRC connection mobility. In case of DC, it is responsible for SgNB change, as well as the UE measurement reports handling. In VGW proposal, some of the functionalities as specified in [105] are performed in the VGW RRC layer VNF to off-load the MeNB RRC layer and to achieve optimized performance.

6.5.3 Packet Data Convergence Protocol

The Packet Data Convergence Protocol (PDCP) layer comprises of PDCP entities. It provides several services to other layers, e.g. transfers both of UP and CP data, header compression, ciphering and integrity protection. PDCP layer also performs many functions, e.g. header compression and decompression of IP data flows, maintaining PDCP sequence numbers, in-sequence delivery of upper layer PDUs at re-establishment of lower layers and duplicate elimination of lower layer SDUs at re-establishment of lower layers for radio bearers mapped on RLC acknowledged mode. One UE may be allocated several PDCP entities, where the data carried by one radio bearer is mapped to one PDCP entity [106]. In VGW-architecture, PDCP layer/service maintains the in-sequence delivery of the traffic when UEs move between IoRL SgNBs, as it remains in the logically centralized location, thereby maintaining global view of the SgNBs.

6.6 VGW in operation

In VGW-based architecture, the topology discovery application builds a graph database from the network nodes and links (set of IoRL-based SgNBs). The monitoring application tracks the network loads and links utilization, to gather detailed statistics about the different SgNBs and their links in the network. The obtained statistics are used for routing the traffic over the available links evenly utilising the MPLS tags, by allocating an MPLS tag for each path between the ingress and egress nodes, thereby allowing the controller to provide fair distribution of the network resource. The controller conducts and installs OpenFlow rules over the forwarding devices of the SgNBs to make a forwarding decision solely based on the received packet MPLS tag. The VGW-D utilizes the MPLS LABEL to specify the output port leading to the next hop. The controller installs the forwarding rules either at the network initialization stage or upon the reception of the first packet

from new SgNB, depending on the operator's requirements. The following sample procedures provides an insight about the operation of the VGW in the VGW-based architecture.

6.6.1 UE attachment procedure

It starts with Attach-request message sent by the UE to the SgNB, where the latter forwards it to the VGW, based on the MPLS tag, and VGW forwards it to the MeNB to perform the normal procedure, specifically after authenticating the UE successfully, VGW forwards the attach response with assigned IP address to the serving SgNB.

6.6.2 Uplink traffic

UE sends the flow to its serving SgNB. Upon receiving the first packet by the serving SgNB, the SgNB forwards it to the VGW. The VGW checks to see if it has a flow-entry programmed on its flow-tables, and if there is, the VGW forwards it accordingly, otherwise, a Packet-In message is sent to the controller requesting instructions to handle this type of traffic. The controller uses the packet header information to identify the traffic type and the destination address to identify the egress ports. The controller determines the path that this packet and all upcoming packets related to the same session should take through the network. The controller then installs a new flow-entry in the VGW-D to push UE traffic to the CN according to the specified procedure in the 3GPP standards. As the VGW-C has a visibility over multiple sgNB, VGW enables a direct communication path for the source and destination UEs, if both are within the coverage area of VGW SgNBs. At the same time, it maintains necessary statistics that are required for the billing proposes. This process reduces the load on the CN and backhaul links, by providing forwarding mechanism that reduces the bandwidth requirements by the backhaul links as well as reducing the latency for the UE-to-UE traffic. The same scenario applies when the destination address belongs to one of the local multimedia services. It is worth mentioning that the work in [107] is designed by following a similar procedures as described.

6.6.3 Downlink traffic

The downlink traffic procedure is performed similar to the uplink traffic procedure. The VGW receives the traffic for the UEs within one of its serving SgNBs, then forwards the downlink traffic to the serving SgNB accordingly. The session type and links' loads are considered in the process to determine the best path, specifically whether to send it via mmWave link or VLC link.

6.6.4 Intra-handover procedure

The handover procedure is explained in more details in the next sub-section to provide a complete understanding of the VGW operation, by comparing the signalling messages in the traditional architecture to the signalling messages in VGW-based architecture. The proposed algorithm for triggering the handover decision is provided below.

Mobile networks with DC deployment are designed to process the CP signalling for intra-handover at the MeNB, in contrast, these processes performed locally at the VGW in the proposed architecture. This relocation of processes from the MeNB to VGW leads to faster and more optimized Intra-handover decisions.

UEs send the "received signal strength" reports constantly, which helps monitoring channel quality. Channel monitoring is an essential process that affects other parameters based on the obtained measurements such as, used modulation, coding, path selection…etc. and for the sake of simplicity, it assumed that UE is reporting measurements of two radio entities, 1 and 2, for each Radio Access Technology (RAT) i.e. VLC1, VLC2 mmWave1 and mmWave2. The measurement message of the Signal to Interference Noise Ratio (SINR) contains a unique identifier (ID) of the VLC light transmitter, or mmWave transceiver. Based on the VLC/ mmWave RRU ID, the VGW is able to identify whether the UE requires intra-handover or is still within the same gNB coverage. The handover flow procedure is depicted in Figure 6-5. The next subsections presents the intra-SgNB handover procedure in details, with and without MeNB change. The required signalling messages for the intra-handover are presented for both deployments namely, traditional deployment and VGW deployment.

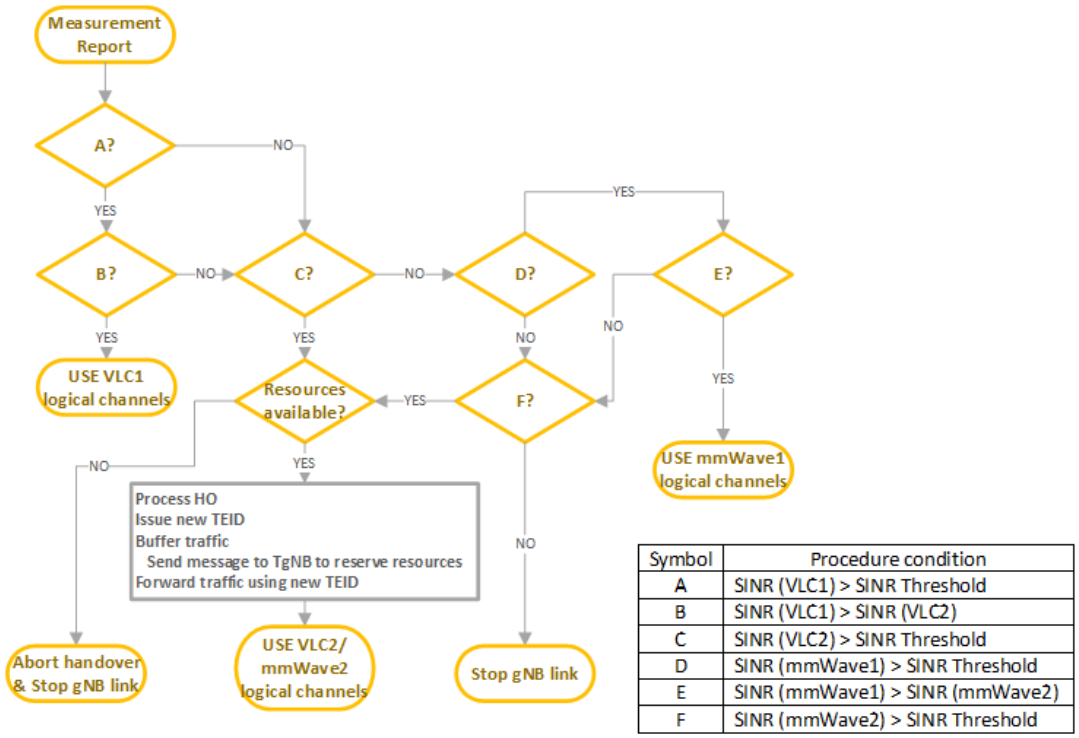

Figure 6-5 handover procedure flow chart

As shown in the HO decision-making algorithm below, the priority given to the VLC RRU over the mmWave, is due to the high bandwidth of the VLC.

1> Check if SINR (VLC1) > SINR (Threshold)

2> Check if SINR (VLC1) > SINR (VLC2)

3> Check if resources available

4> H.O decision made

4> Buffer traffic

4> Send flow-mod msg to forward traffic to VLC1 using TEID

2> Else if SINR (VLC2) > SINR (Threshold)

3> Check if resources available

4> H.O decision made

4> Buffer traffic

4> Send flow-mod msg to forward traffic to VLC2 using TEID

1> Else if SINR (VLC2) > SINR (Threshold)

2> Check if resources available

3> H.O decision made

3> Buffer traffic

3> Send flow-mod msg to forward traffic to VLC2 using TEID

1> Else if SINR (mmWave1) > SINR (Threshold)

2> Check if SINR(mmWave1) > SINR(mmWave2)

3> Check if resources available

4> H.O decision made

4> Buffer traffic

4> Send flow-mod msg to forward traffic to mmWave1 using TEID

2> Else if SINR (mmWave2) > SINR (Threshold)

3> Check if resources available

4> H.O decision made

4> Buffer traffic

4> Send flow-mod msg to forward traffic to mmWave2 using TEID

1> Else if SINR (mmWave2) > SINR (Threshold)

2> Check if resources available

3> HO decision made

3> Buffer traffic

3> Send flow-mod msg to forward traffic to mmWave2 using TEID

1> Else stop gNB link and notify Master cell eNB (no resources available)

6.6.4.1 Intra-SgNB within same MeNB (Traditional architecture)

The signalling messages provided below are based on the 3GPP specification [86], for the handover procedure depicted in Figure 6-6.

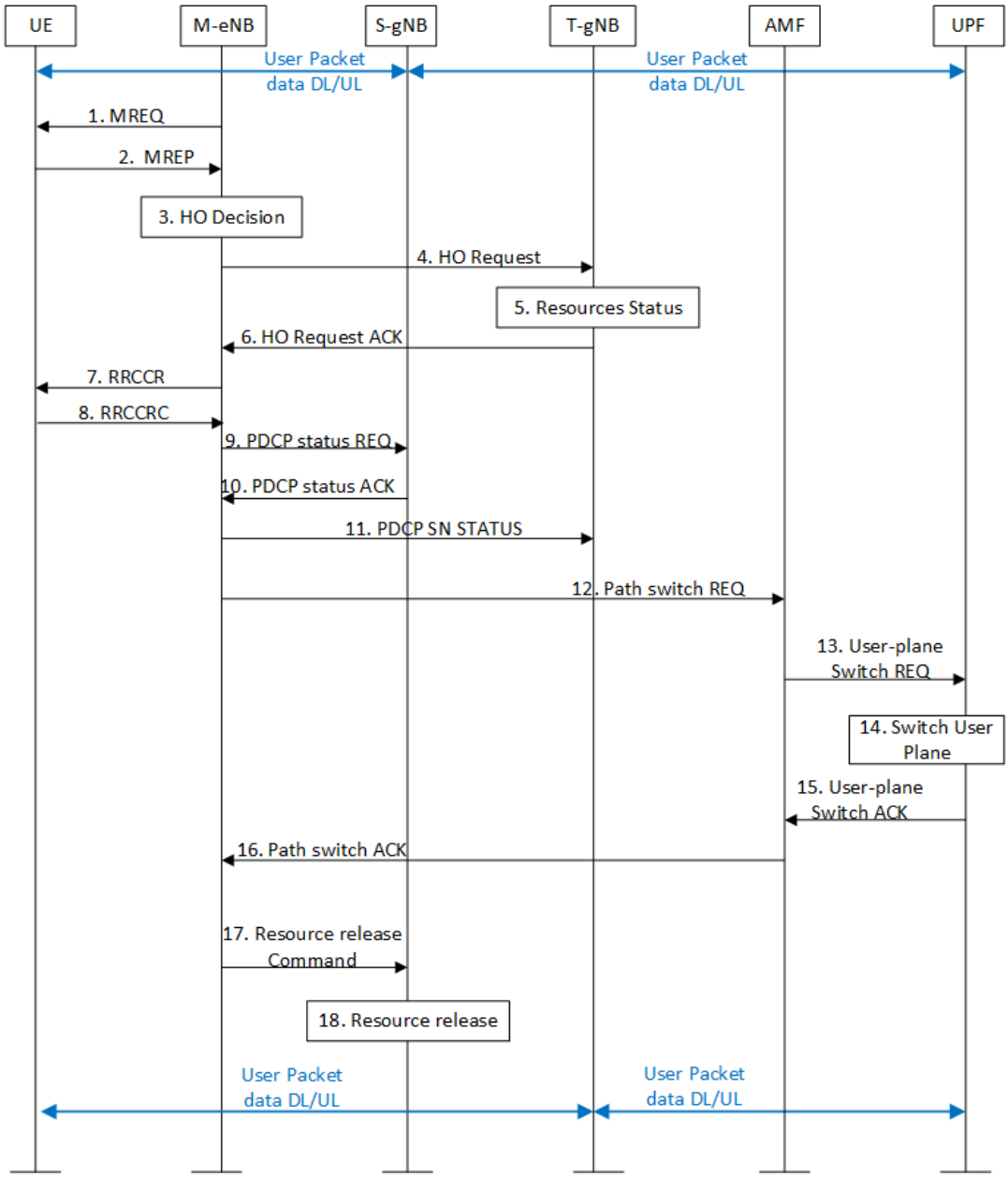

Figure 6-6 signalling flow for Intra-SgNB handover

- (1, 2) M-eNB acquires the channel information by the means of the Measurement REQuest (MREQ) and Measurement REsPonse (MREP).
- (3) Based on the obtained information, MeNB takes HO decision according to the HO algorithm.
- (4, 5, 6) The MeNB initiates the SN change by requesting the target SN (T-gNB) to allocate resources for the UE by means of the S-gNB Addition procedure. The MeNB may include measurement results related to the TgNB. TgNB ensures availability of resources for the UE. If configured with bearers requiring SCG radio resources, the UE synchronizes to the TgNB. TgNB notify the MeNB if it was successful.

- (7, 8) The MeNB triggers the UE to apply the new configuration. The MeNB indicates to the UE the new configuration in the RRCConnectionReconfiguration (RRCCR) message including the NR RRC configuration message generated by the target SgNB. The UE applies the new configuration and sends the RRCConnectionReconfigurationComplete (RRCCRC) message, including the encoded NR RRC response message for the target SgNB, if needed. In case the UE is unable to comply with (part of) the configuration included in the RRCCR message, it performs the reconfiguration failure procedure.
- (9, 10, 11) MeNB requests the PDCP sequence number of the traffic. If data forwarding is needed, the MeNB provides data forwarding addresses to the SgNB. Reception of the PDCP status Request message triggers the SgNB to stop providing user data to the UE and, if applicable, to start data forwarding.
- (12-16) if one of the bearer was terminated at the SgNB, path update is triggered by the MeNB.
- (17) After receiving the acknowledgment for path switch request, MeNB sends Resource release Command to the SgNB.
- (18) Upon reception of the Resource release command, the SgNB can release the related resource associated to the UE context. Any ongoing data forwarding may continue.

6.6.4.2 Intra-SgNB with MeNB change (Traditional architecture)

Intra-SgNB with MeNB change is when the UE moves from source SgNB to destination SgNB, but the source and destination SgNB belong to different MeNB. The signalling flow is depicted in Figure 6-7. The signalling is following the 3GPP specification release 15 [86].

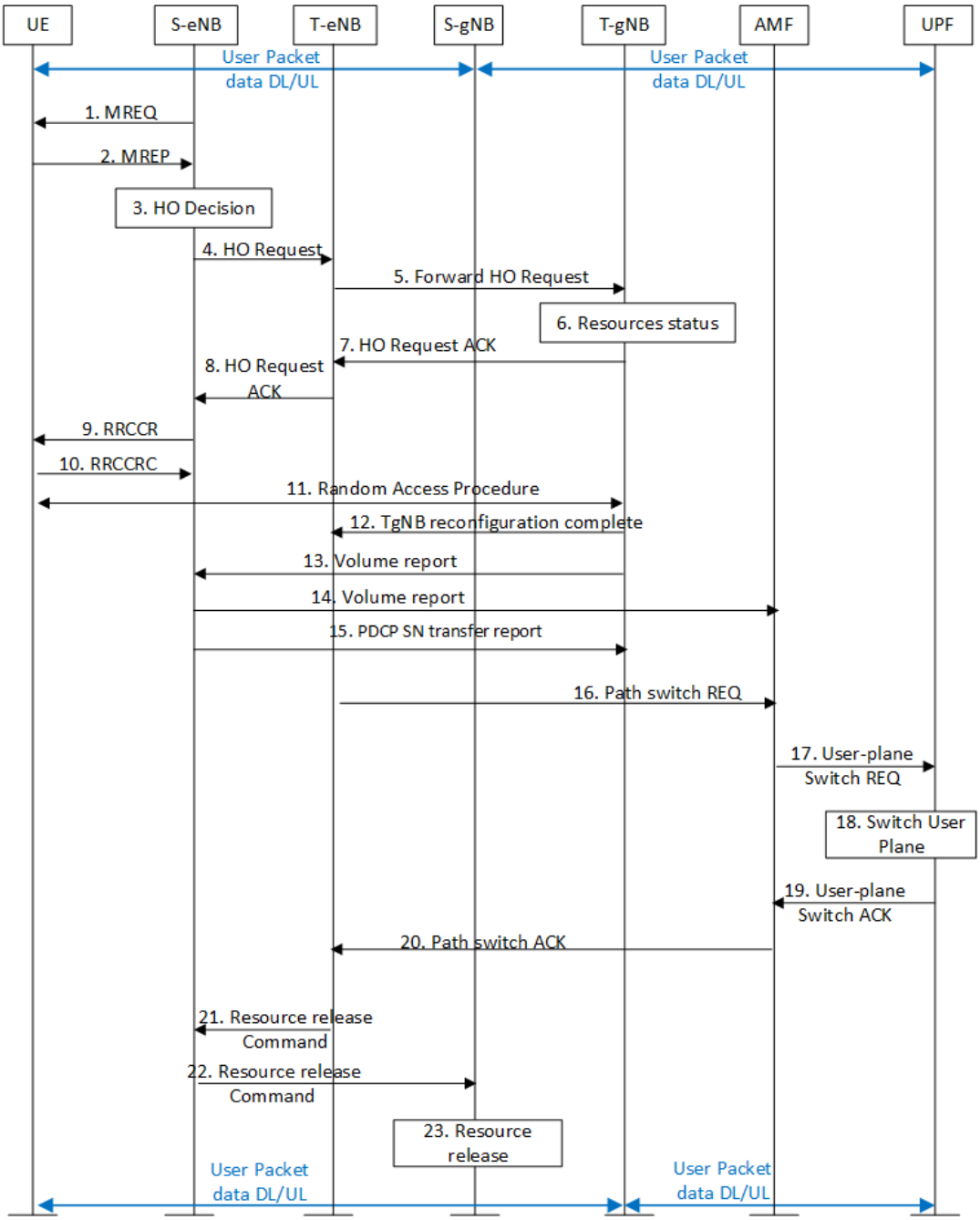

Figure 6-7 signalling flow for Intra-SgNB, inter-MeNB handover

- (1, 2) M-eNB acquires the channel information by the means of the Measurement REQuest (MREQ) and Measurement REsPonse (MREP).
- (3) Based on the obtained information, MeNB takes HO decision according to the HO algorithm.
- (4) The Source MN (S-eNB) starts the handover procedure by initiating the Xn Handover Preparation procedure including both MCG and SCG configuration. S-eNB includes the (source) SN (SgNB) UE Xn-C ID, SgNB ID and the UE context in SgNB in the Handover Request message.

- (5, 6, 7) TeNB forwards the request to the TgNB in the form of Addition Request including the UE context in the SgNB that was established by the S-eNB. The latter requests the T-gNB to allocate resources for the UE. TgNB ensures availability of resources for the UE. If configured with bearers requiring SCG radio resources, the UE synchronizes to the TgNB. TgNB notifies the TeNB if it was successful.
- (8) TeNB includes within the Handover Request Acknowledge message a transparent container to be sent to the UE as an RRC message to perform the handover, and may also provide forwarding addresses to the S-eNB.
- (9, 12) The MeNB triggers the UE to apply the new configuration. The MeNB indicates to the UE the new configuration in the RRCConnectionReconfiguration (RRCCR) message including the NR RRC configuration message generated by the target SgNB. The UE applies the new configuration and sends the RRCConnectionReconfigurationComplete (RRCCRC) message, including the encoded NR RRC response message for the target SgNB, if needed. In case the UE is unable to comply with (part of) the configuration included in the RRCCR message, it performs the reconfiguration failure procedure.
- (13) The TgNB sends volume Report message to the S-eNB, which includes the data volumes delivered to the UE over the NR radio for the related E-RABs.
- (14) S-eNB forwards the volume Report message to AMF to provide information on the used NR resource.
- (15) S-eNB sends the PDCP sequence number of the traffic to the TgNB in PDCP SN transfer report. Reception of the PDCP SN transfer report triggers the SgNB to stop providing user data to the UE and, if applicable, to start data forwarding.
- (16-20) TeNB initiates the N2 Path Switch procedure.
- (21) After receiving the acknowledgment for path switch request, TeNB sends Resource release Command to the S-eNB.
- (22) S-eNB forwards the Resource release command to SgNB.
- (23) Upon reception of the Resource release command, the SgNB can release the related resource associated to the UE context. Any ongoing data forwarding may continue.

6.6.4.3 Intra-SgNB within same MeNB (VGW-based architecture)

The signalling messages flow is depicted in Figure 6-8, it is compatible with previous works in [107], with fine-tuning to reflect the VGW-based architecture.

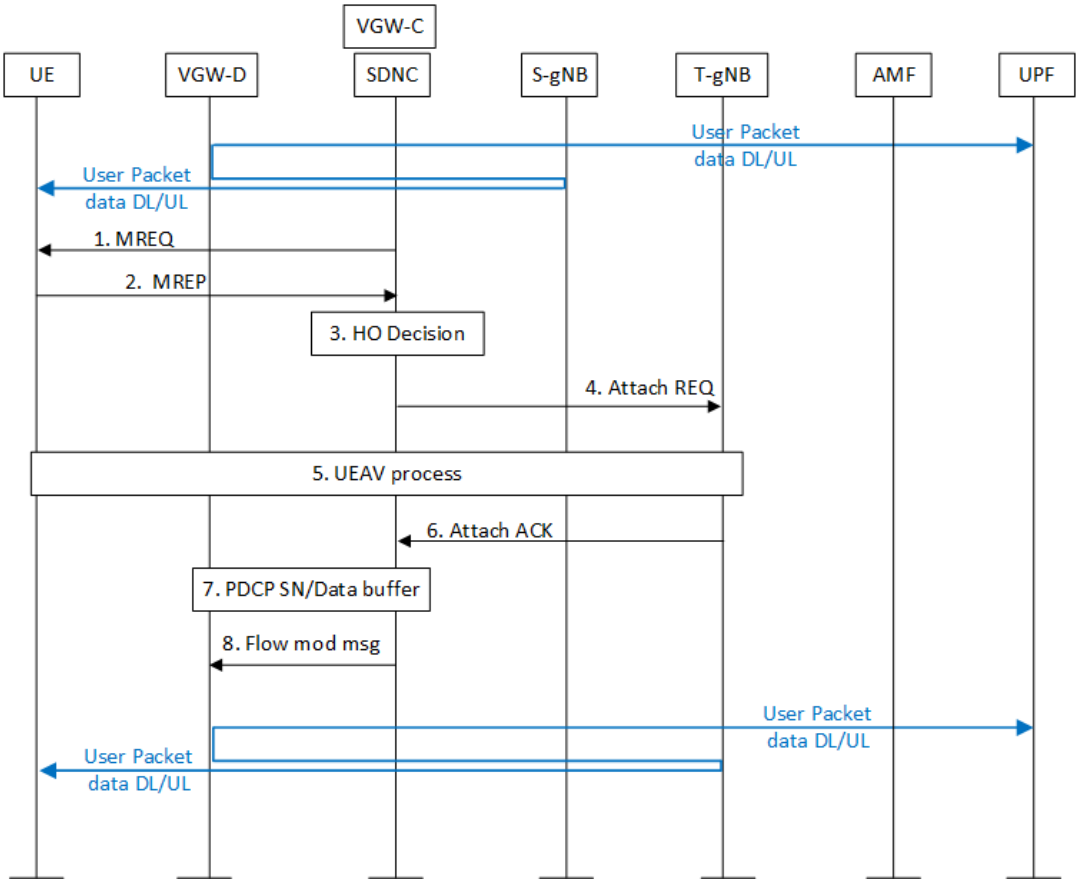

Figure 6-8 signalling flow for Intra-SgNB handover (VGW-based architecture)

- (1, 2) VGW-C acquires the channel information by the means of the Measurement REQuest (MREQ) and Measurement REsPonse (MREP).
- (3) Based on the obtained information, VGW-C takes HO decision, according to the HO algorithm explained earlier.

Note: VGW-C is registered as listener for the Forwarding Devices' statistics reports of all the SgNBs, to establish the awareness of the resources' availability at each gNB. Thereby VGW-C can perform handover decision without the need for TgNB resource checking messages.

- (4, 5, 6) The VGW-C initiates the SN change by performing UE Attachment Validation (UEAV) process. This process includes controller instruction for the TgNB to send a short message of few packets to the UE on the IP address given by the VGW-C and receiving an echo reply.

Note: TgNB piggybacks the UE response message into OFPT_PACKET_IN message, to inform the VGW-C about the status of the process to confirm the attachment of UE.

- (7) VGW-C establishes the data forwarding process internally, starts by marking the PDCP sequence number of the traffic. If data forwarding is needed (in case of inter-eNB handover), the VGW-C begins to buffer the data, while obtaining the data forwarding addresses to the TgNB from the S-eNB.
- (8) SDN-C sends Flow Mod MSG to VGW-D to install new flow rule to forward the UE DL traffic towards the TgNB.

Note: The previous flow rule removed after hard-time-out counter expires, without the need for further release instructions.

6.6.4.4 Intra-SgNB with MeNB change (VGW-based architecture)

The following messages flow show the procedure when the UE handed over from one gNB to another gNB, both of which are secondary gNBs working with different MeNB, and communicating via different VGWs. Figure 6-9 depicts the messages flow in case of Intera-SgNB, Inter-MeNB handover process.

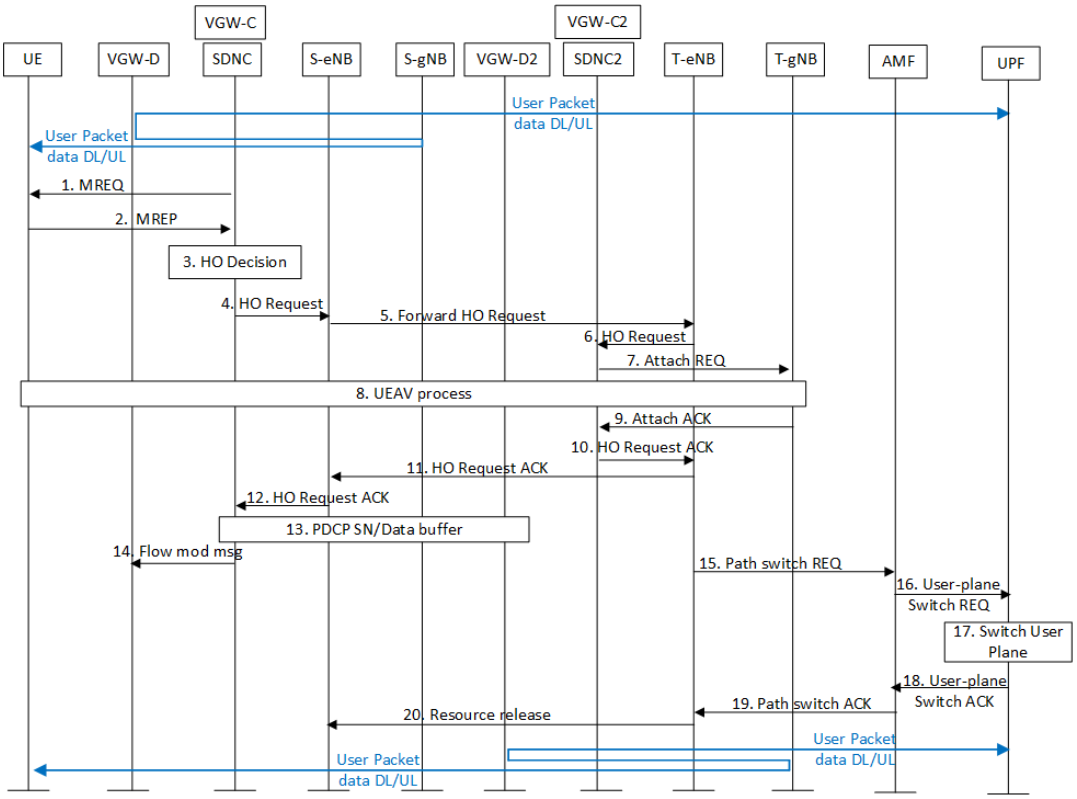

Figure 6-9 signalling flow for Intra-SgNB, inter-MeNB handover (VGW-architecture)

- (1, 2) VGW-C acquires the channel information by the means of the Measurement REQuest (MREQ) and Measurement REsPonse (MREP).
- (3) Based on the obtained information, VGW-C takes HO decision, according to the HO algorithm explained earlier.

- (4, 5) VGW-C at SeNB indicates to the S-eNB the need for handover to gNB not within its group. S-eNB starts the handover procedure by forwarding the request to the TeNB. S-eNB includes the (source) SN (VGW-D) address in the Handover Request message.
- (6) TeNB forwards the request to the VGW-C2 at the TeNB
- (7, 8, 9) The VGW-C2 initiates the SN change by performing UE Attachment Validation (UEAV) process. This process includes controller instruction for the TgNB to send a short message of few packets to the UE and receiving an echo reply using IP address assigned by source VGW-C. Furthermore, the message includes new IP address assigned by VGW-C2 to used onwards.
- (10, 11, 12) VGW-C2 sends an acknowledgment to the TeNB, the later forwards this acknowledgment message to the S-eNB, which in turn forwards it to VGW-C.

Note: VGW buffers the traffic until obtaining the forwarding address of the next VGW-D

- (13, 14) VGW-C establishes the data forwarding process locally with VGW-D2 (using the address received earlier). The process starts by marking the PDCP sequence number of the traffic as well as delivering the buffered data.
- (15-19) TeNB initiates the N2 Path Switch procedure.
- (20) After receiving the acknowledgment for path switch request, TeNB sends Resource release Command to the S-eNB.

Note: The previous flow rule is removed after hard-time-out counter expires, without the need for further release instructions.

6.7 Signalling load analysis

As mentioned earlier, the same evaluation method from [108] is used, to calculate signalling load. The signalling flow messages were drawn to calculate the exchanged signalling messages for each handover scenario. The traditional architecture is based on 3GPP specification release 15 [86]. Furthermore, the same concept applies to analyze the handover procedure for the proposed VGW architecture.

6.7.1 Intra-SgNB without MeNB change

From Figure 6-6, the total number of messages (entering and leaving) processed at AMF per hour due to UE intra-SgNB handover without MeNB change is given by:

$$\mathrm{SL1}_{\mathrm{AMF}}(k) = [4\ (1 - P_r) + 6\ P_r]\ R\ (1 - P_i)C \qquad 6.1$$

Where $SL1_{AMF}$ is the total number of signaling messages, k is the application types, R is the mobile crossing rate out of enclosed region, P_i is the probability of UE in idle state, C is the total number of MeNBs. The numbers 4 and 6 represent the number of messages at AMF without and with UPF change respectively as can be observed in Figure 6-6. Finally, P_r is the relocation probability of UPF, which is approximately equals to:

$$P_r = \frac{1}{\sqrt{Ca}} \qquad 6.2$$

Where Ca is the number of eNB per tracking area.

R, is the mobile crossing rate out of enclosed region with perimeter length L, it is given by:

$$R = \frac{\rho\, V L}{\pi} \qquad 6.3$$

Where ρ is UE density UE/Km^2

V is the average UE velocity Km/hr.

L is the perimeter length of a cell Km

P_i is the probability of the UE being in idle state, the work in [108] shows the proof for the probability of UE being in connected state equals to ($1 - Pi$). All parameters included in Table 6-1.

6.7.2 Intra-SgNB with eNB change

From Figure 6-7, the total number of messages (entering and leaving) processed at AMF per hour due to UE intra-SgNB handover with MeNB change is given by:

$$\mathrm{SL2}_{\mathrm{AMF}}(k) = [5\ (1 - P_r) + 7\ P_r]\ R\ (1 - P_i)C \qquad 6.4$$

Where $SL2_{AMF}$ is the total number of messages processed at AMF. Similarly, the numbers 5 and 7 represent the number of messages at AMF without and with UPF change respectively as can be observed in Figure 6-7.

6.7.3 VGW-Based architecture

Applying the same concept to calculate the signalling load at the AMF, following the signalling messages for Intra-SgNB handover as depicted in Figure 6-8, the total number of messages processed at the AMF without and with MeNB change is given by:

$$\mathrm{SL1v}_{\mathrm{AMF}}(k) = 0 \qquad 6.5$$

Where $SL1v_{AMF}$ is the total number of signaling messages processed at AMF with VGW deployment without MeNB change.

And :

$$\mathrm{SL2v_{AMF}}(k) = [4\ (1 - P_r) + 6\ P_r]\ R\ (1 - P_i)C \qquad 6.6$$

The numbers 4 and 6 represent the number of messages at AMF without and with UPF change respectively as can be observed in Figure 6-9.

6.7.4 Signalling load at eNB/ VGW-C

In the DC mode, the first control entity for the SgNBs is the MeNB, where the N2 CP interface signalling terminates at the MeNB. Therefore, signalling load at the MeNB could be calculated in the traditional architecture and compared with the equivalent control entity at the VGW architecture that is the VGW-C. From Figure 6-6 through Figure 6-9, the total number of processed messages per hour at the MeNBs can be calculated as the following:

$$\mathrm{SL3_{eNB}}(k) = [12\ (1 - P_{reNB}) + 18\ P_{reNB}]\ R\ (1 - P_i)C_g \qquad 6.7$$

Where $\mathrm{SL3_{eNB}}$ is the total number of processed messages per hour at the MeNB, C_g is the total number of SgNBs in considered region. The numbers 12 and 18 represent the number of messages at MeNBs without and with MeNB change respectively as can be observed in Figure (6-6) and Figure (6-7). Finally, P_{reNB} is the probability of the MeNB relocation, which is approximately calculated by:

$$R_{reNB} = \frac{1}{\sqrt{C_{ag}}} \quad 6.8$$

Where C_{ag} is the number of SgNB per eNB coverage area.
Moreover, total number of messages per hour at VGW-C is given by:

$$\mathrm{SL3v_{vgw-c}}(k) = [5\ (1 - P_{reNB}) + 9\ P_{reNB}]\ R\ (1 - P_i)C_g \qquad 6.9$$

$\mathrm{SL3v_{vgw\text{-}c}}$ is the total number of messages per hour at VGW-C. The numbers 5 and 9 represent the number of messages at VGW-C without and with MeNB change respectively as can be observed in Figure (6-8) and Figure (6-9).

6.8 Numerical results

This section presents numerical results using all the equations described in the previous sections, to quantify the efficiency of the proposed VGW architecture. Table 6-1 presents the values of the parameters in equation 6.1 through equation 6.9, some of the parameters could be inferred from other parameters, so they have no values in that table. The equations are based on the work of [108][109]. Evaluate the total number of signalling messages that are processed by AMF, the VGW controller, and MeNBs. The impact of these values on the amount of signalling messages at various cell densities are shown.

Parameter	Description	Value
P_r	Gateway relocation probability	Scenario based
P_{reNB}	MeNB relocation probability	Scenario based
R	Crossing rate out of a cell (UEs/h)	Depend on equation 6.3
ρ	UEs density (UEs/Km2)	Variable
V	UEs' velocity (Km/h)	5Km/h (except scenario 4)
L	Perimeter length of a cell Km	2 π r
r	Cell radius Km	variable = $\gamma\sqrt{\frac{S}{C\pi}}$
γ	Overlapping factor	1.2
P_i	UE idle state probability	$= \frac{1}{1 + [Av.arrival\ rate * Av.seccion\ duration]}$
C	Total number of MeNBs in considered region	500 (scenarios 1-4)
C_g	Total number of SgNBs in considered region	20 (scenarios 5, 6)
S	Area of a considered region	500 Km2 (scenarios 1-4), 2Km2 (scenarios 5, 6)

Table 6-1 evaluation tests parameters

6.8.1 Scenario 1

In this scenario, the total number of processed signalling messages per hour at the AMF due to intra-SgNB handover are calculated. The scenario does not involve MeNB change. Using equation 6.1 to calculate the total number of messages at AMF in both network architectures namely, traditional and VGW-based.

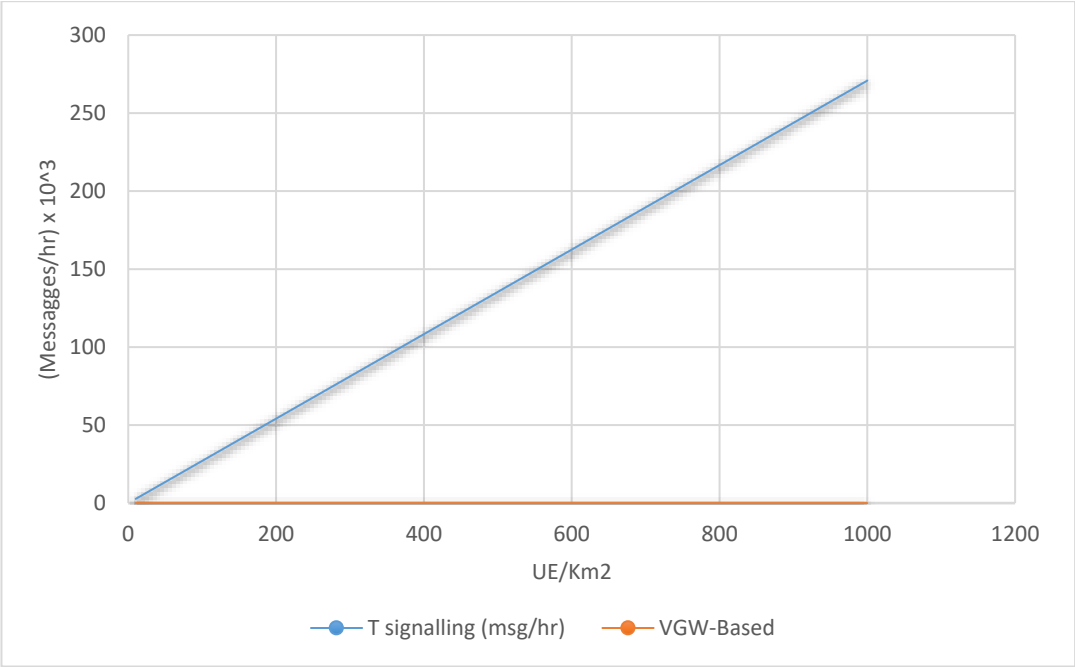

Figure 6-10 total amount of signals at AMF (intra-SgNB, without MeNB change)

As can be seen from Figure 6-10, the amount of signalling increases linearly with UE densities in the traditional network architecture, however, adopting the VGW architecture, the AMF would be offloaded and there are no signalling messages at the AMF regardless of UE density. It is worth mentioning that, the probability of changing the UPF equals zero in the equation, since it is not possible to have two SgNBs within the coverage of one MeNB, while served by different UPFs.

6.8.2 Scenario 2

In this scenario, UEs undergo intra-SgNB handover with MeNB change, while both SgNBs served by the same UPF.

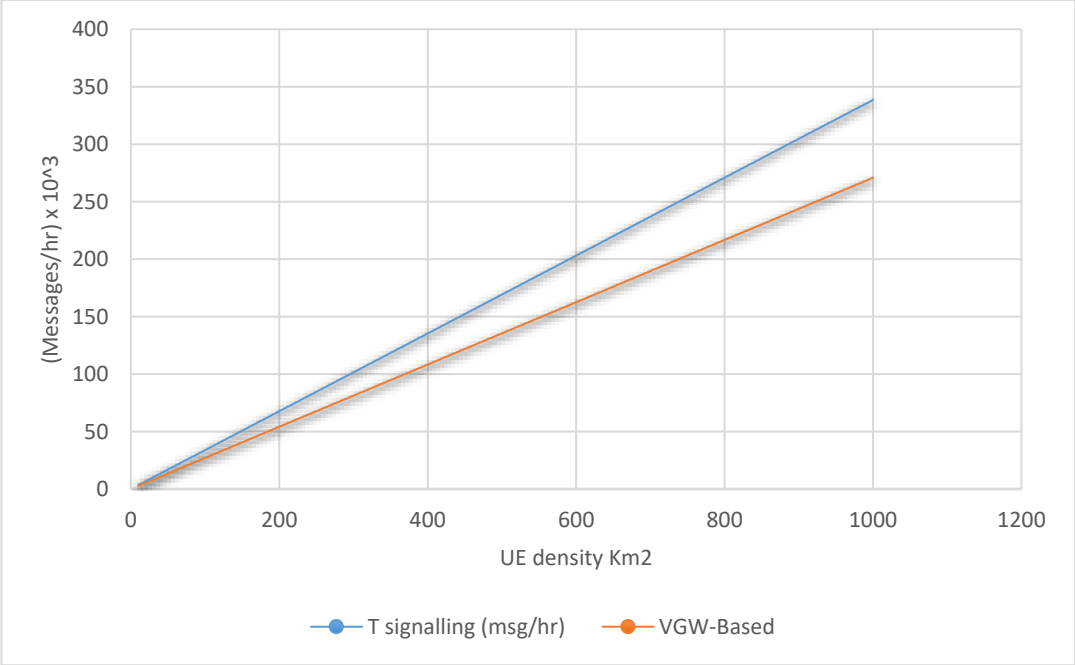

Figure 6-11 total amount of signals at AMF (intra-SgNB, with MeNB change)

Figure 6-11 depicts the amount of signalling per hour processed at AMF during intra-SgNB while changing the MeNB. VGW-based network architecture improves the network performance, by means of reducing the amount of signalling messages processed at the AMF. The effect of the VGW-based architecture is not so evident when the UE densities are relatively low e.g. $\rho < 200$ UE/Km2, however it is more effective at higher densities.

6.8.3 Scenario 3

The third scenario also considering intra-SgNB handover with MeNB change, but in the case of serving UPF change for the target MeNB from the serving UPF for the source MeNB.

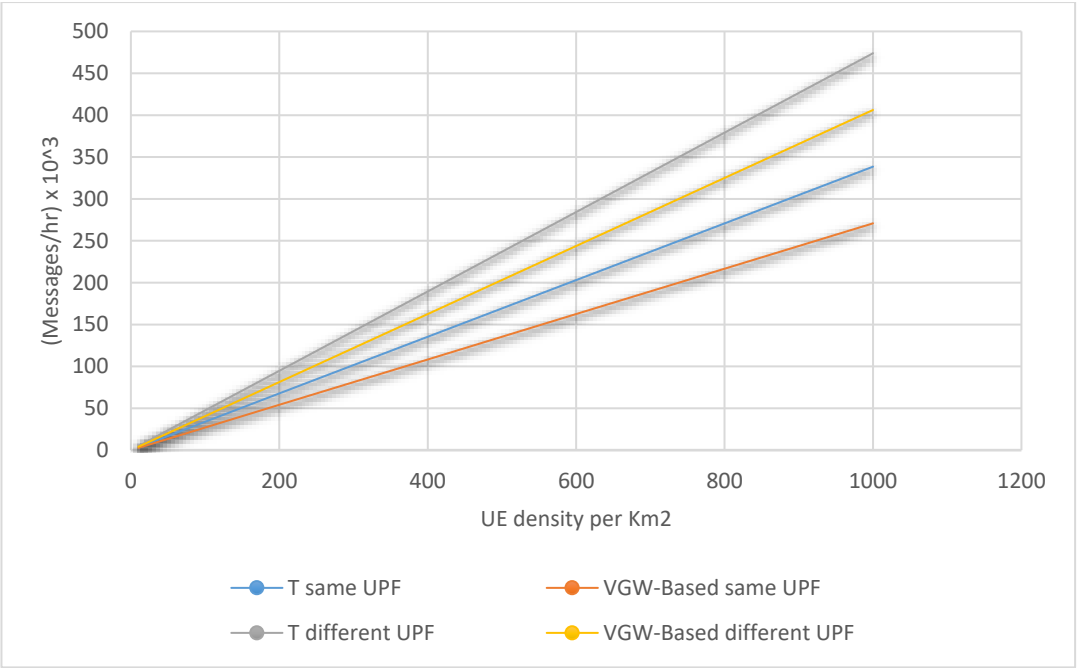

Figure 6-12 total amount of signals at AMF (intra-SgNB, with MeNB change, with/without UPF change)

As shown in Figure 6-12, although VGW-based architecture does not offload the signalling from the AMF as effectively as it did without UPF change, but is still showing an improvement compared to the traditional network architecture with changing UPF scenario.

6.8.4 Scenario 4

This scenario compares the performance of both system architectures at different UE mobility rates, using the measurements based on the second scenario parameters for intra-SgNB handover and inter-MeNB while being served by the same UPF. The UE density is assumed to be 100 UE/Km2.

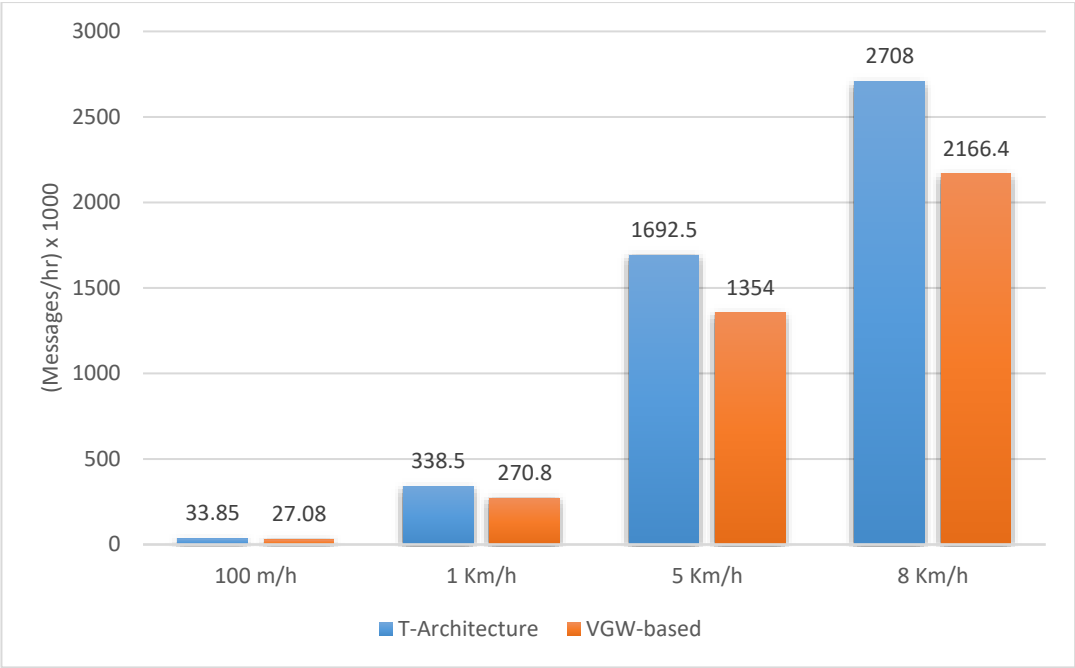

Figure 6-13 signalling messages at AMF (intra-SgNB, inter-MeNB and same UPF at different UE mobility)

As depicted in Figure 6-13, the signalling messages increase linearly with UE mobility rate. The number of the offloaded signals are proportional to the UEs' mobility speed, where VGW-based architecture offloads $541.6 * 10^3$ message per hour compared to the traditional architecture.

6.8.5 Scenario 5

In this scenario, calculating the amount of signalling messages processed at the first control entity that is responsible for the control signalling of the S-gNBs namely, M-eNB and VGW-C, at the traditional architecture and VGW-based architecture respectively, assuming source and destination SgNBs are within the same Macro-cell eNB coverage area.

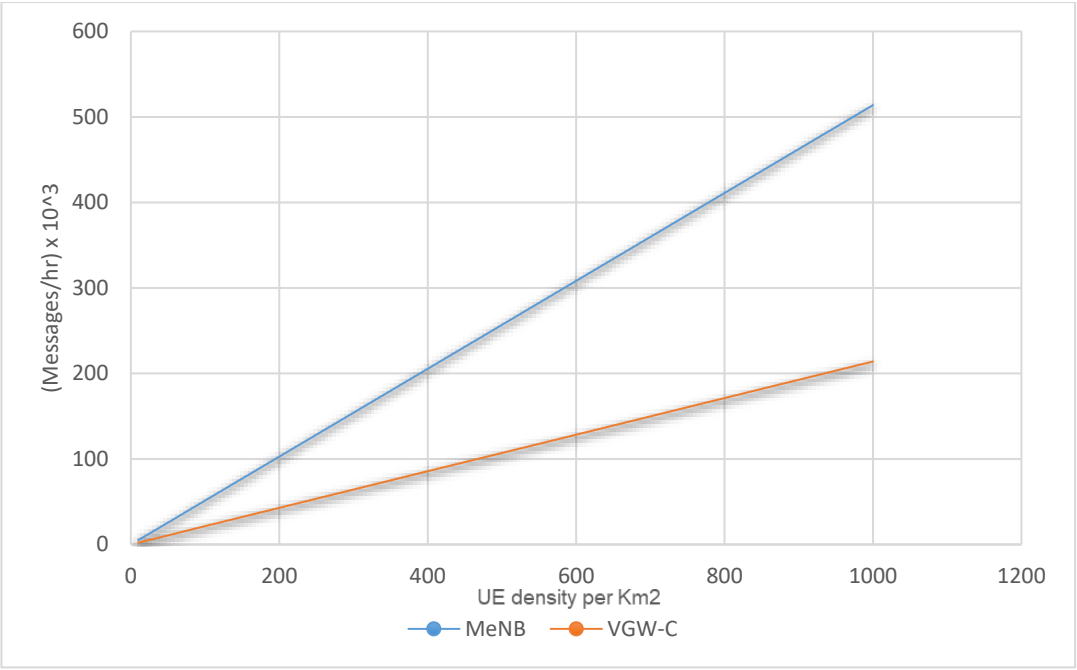

Figure 6-14 total amount of signals at MeNB (intra-SgNB, without MeNB change)

The amount of signalling messages at M-eNB in traditional network architecture and VGW-C at VGW-based architecture is depicted in Figure 6-14. The measurements are in the case of intra-SgNB with no MeNB change scenario. It is apparent that the proposed architecture reduces the amount of required signalling to perform the intra-SgNB handover.

6.8.6 Scenario 6

In this last scenario, calculating the amount of signalling at the source and destination MeNB at the traditional architecture and comparing it with their counterparts at VGW-based architecture. It is worth mentioning in both scenarios (five and six), that the measurements performed assumes both source and destination MeNBs are served by the same UPF.

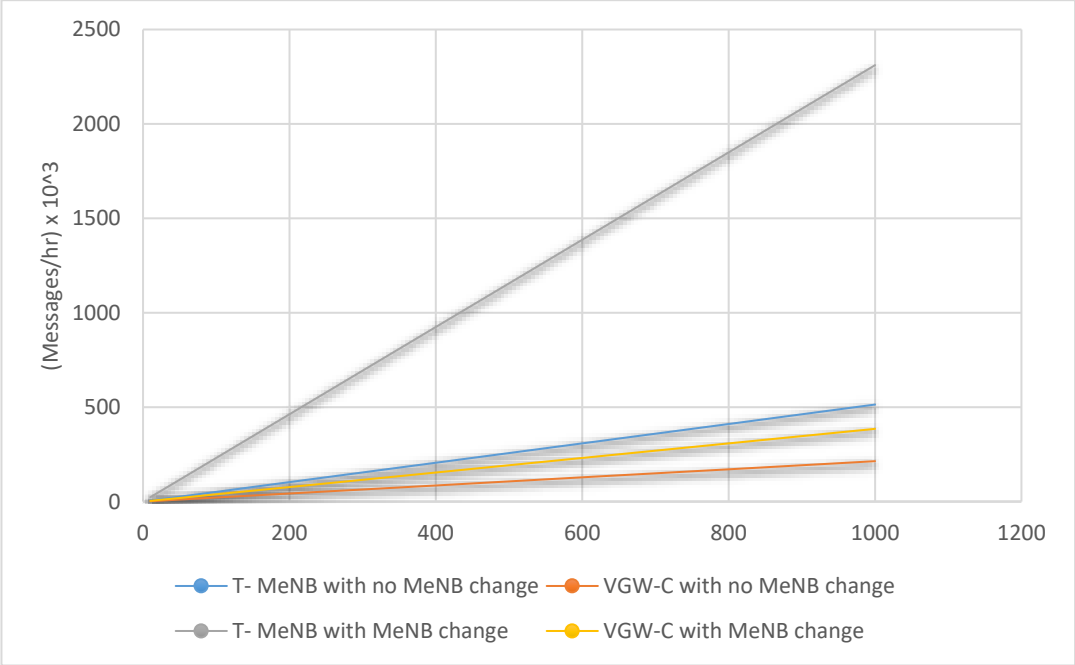

Figure 6-15 total amount of signals at MeNB (intra-SgNB, with and without MeNB change)

The overall signalling at the MeNB and VGW-C in both network architectures are depicted in Figure 6-15. As can be seen, the proposed architecture reduces the required signalling to more than 50 percent and even more, which is due to the ability of the virtualized SDN architecture to perform many sub-processes internally without the need for signalling overhead.

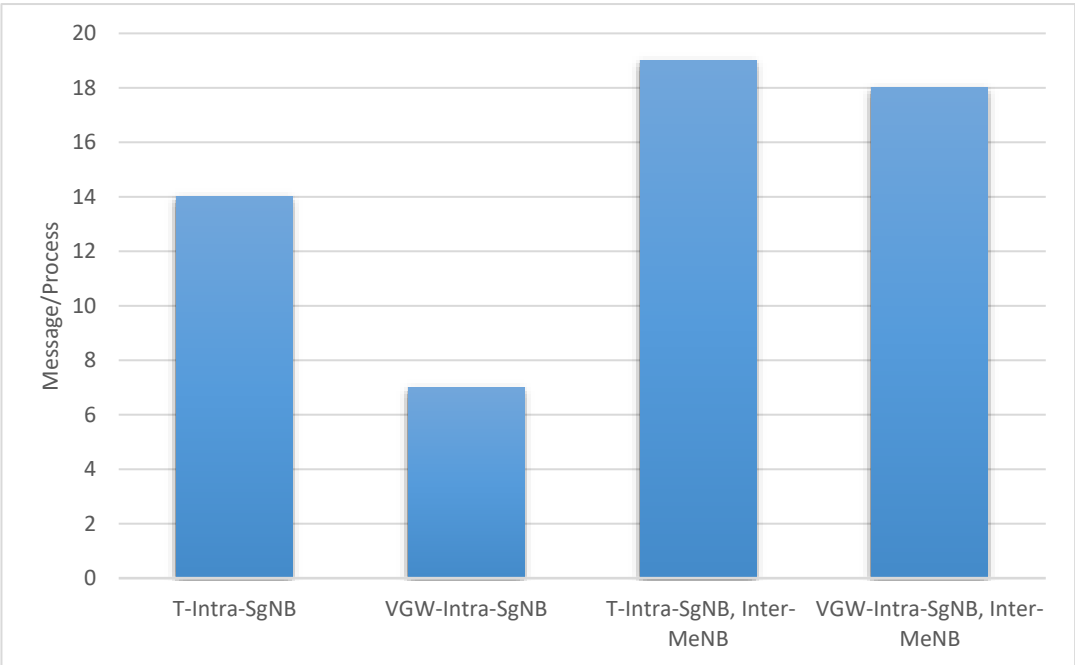

Figure 6-16 overall messages per process

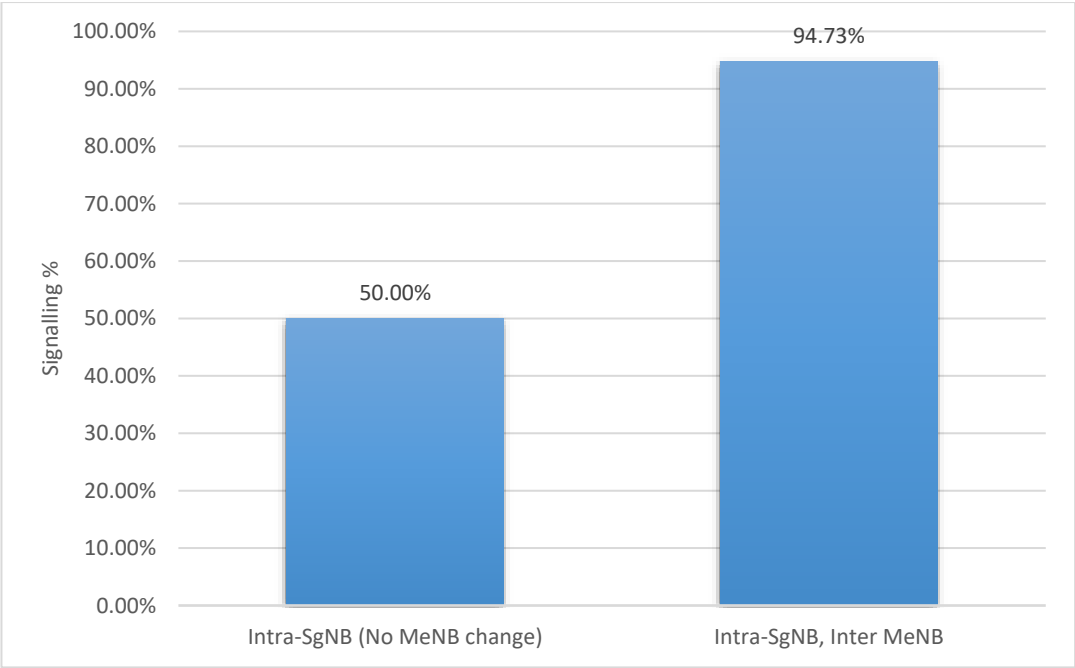

Figure 6-17 required signalling percentage when utilizing VGW compared to traditional architecture

Figure 6-16 depicts the overall messages for Intra-SgNB processes with/without MeNB change for both network architectures; the overhead signalling is much reduced by utilizing the VGW architecture. Figure 6-17 depicts the required signalling percentage with the VGW architecture in comparison with the traditional architecture, which is 50 and 94.73 percent for the case of Intra-SgNB no MeNB change and Intra-SgNB Inter-MeNB respectively. The performance enhancement is measured by means of reduced overhead signalling. In the case of Intra-SgNB, the overhead signalling is reduced by 50%, while in the case of Intra-SgNB with Inter-MenB, the overhead signalling is reduced by only 5.27%. It is clear that VGW architecture offers optimum IoRL deployment solution.

6.9 Conclusion

Mobile networks are constantly evolving to cope with increasing user demands and high expectations. Enhancing the network performance within indoor environment is an essential requirement for a successful service provider. IoRL is one of the next generation gNBs for indoor coverage; it offers various services to its clients. VGW is a virtual entity that facilitates the deployment of IoRL gNBs efficiently by various means mainly, providing a logically centralized entity that hosts multiple back-end components e.g. cache servers, databases…etc. in addition to lowering the un-necessary back-haul traffic especially due intra-handover procedures. VGW provides extraordinary performance during intra-SgNB

handover. However, it offers limited performance enhancement in case of intra-SgNB inter-MeNB. The performance analysis reveals that the VGW is more efficient when UE remains within the same AMF. For the case of Intra-SgNB, the overhead signalling at the AMF is reduced by 100% while the overall signalling is reduced by 50% compared to traditional architecture deployment. On the other hand, the overall signalling reduced by only 5.27% in the case of Intra-SgNB with Inter-MenB. The results reflect the VGW effectiveness in terms of reducing the unnecessary back-haul signalling. VGW brings other benefits to the network architecture e.g. promoting network automation/slicing and resource efficient utilization due to the nature of vitalized design.

7 Conclusion and Future Works

7.1 Conclusion

This thesis presented the design for one of the next generation coverage solutions for indoor environments namely, IoRL, along with multiple multimedia services that are proposed to enhance the network performance as well as the end users' QoE. Caching solution for IoRL is introduced in chapter 5 and deployment solution within MNO is proposed in chapter 6. Several experiments were performed within various testing scenarios, to quantify the service enhancement. The obtained results led to the following conclusions.

- ✓ SDN provides the concept of configurable traffic routing, which is an enabler for the emergence of new services e.g. FMS, MSS. This concept is utilized in developing the geolocation services as can be seen in chapter 4, where these two services designed to offer the services to end users in proactive and reactive manner depending on this feature of SDN.
- ✓ Softwareization of networking is the enabler for resource virtualization, which in turn represents the basis for designing cost-effective flexible platforms. As described in chapter3, where the system design was implemented by utilizing a VIM that enables resource abstraction which enabled the flexibility of the developed platform.
- ✓ Introducing SDN networking promotes new resource provisioning that enables automating resource allocation. The developed platform presented in chapter 3, was developed by exploiting SDN networking, which enabled the deployment of various services such as the load balancer and security services in the form of VNFs by sharing the available resources and coexist on the same platform, as described in chapter3 with more details.
- ✓ Next generation mobile networking is designed to be more flexible and adaptable, by relying more on the open-source solutions and deprecating proprietary solutions. All of the developed platform, services and proposed solutions, adopted open-source technologies e.g. as can be seen in chapters 3, 4, 5 and 6.
- ✓ Next generation mobile network exploit new access technologies to support higher data rates and reduce latency, as can be seen that IoRL small cell

design incorporates VLC and mmWave technologies to achieve higher data rates and offer higher bandwidths.

- ✓ Next generation mobile networks support distributed network management to remove possible bottle necks. Therefore, the proposed architecture for IoRL included the design of the IHIPGW which entails management features and offer data processing and storage although it is deigned to be distributed within the network architecture.
- ✓ Cloud computing concept brings the potential to the mobile networks, by facilitating the deployment of network services that require rich computational computing resources and making them available to end users independent of their devices' resources, for instance, follow me service and multicast sharing services are designed to provide services to end user by elevating all the data processing, video rendering and data storage to the cloud, and only require the user to use his mobile device as lightweight controller and location indicator.
- ✓ Exploiting the synergy between SDN and NFV technologies promotes efficient deployments for intelligent network services. Although, SDN is not dependent on NFV, but they normally work well together, as shown in the proposed architecture which outlined in chapter3, it was designed by adopting virtualized platform that was managed by an open-source VIM, which have exploited SDN networking to perform traffic routing.
- ✓ The concept of virtualization and IaaS enables MNOs and service providers to offer better services at lower costs, by enhancing the resources utilization. This concept represented the backbone of the proposed architecture, where the IHIPGW is developed to be compatible with cloud computing deployments. It was deployed within OpenStack platform to enable the resources to be shared across the various services. One of the contributions of this work is to implement an industrial-like platform of an the IHIPGW over an open-sources IaaS platform.
- ✓ The use of VIM e.g. OpenStack, empower network managers with powerful tools, which enable them to manage the available pools of resources more effectively and intelligently.
- ✓ Relying on cloud computing within mobile network architecture promotes resource slicing, efficient resource utilization and simplifies network reaction to load fluctuations, by the help of the autonomous scalability features. This concept was one of the motivations for some of the contributions of this work, e.g. designing and proposing small cell caching

for IoRL gNB and VGW. Both of these solutions developed by exploiting the available resources of the proposed IoRL small cell, to make efficient use of the resources and to achieve resource slicing.

- ✓ Exploiting SDN, NFV and cloud computing technologies, facilitates the deployment of new virtual network entities that helps to enhance network performance via providing new VNFs without imposing negative impact on the propagation latency. As can be seen in chapter 6, VGW is designed to be a virtual entity and coexist on the physical host of other network entities that already exists within the network architecture, thereby do not create a new entity and cause further complications.

7.2 Future Works

The adoption of small cells in mobile networks are serving multiple purposes e.g. enhance the QoS, expands the coverage area, enables granule-level tuning to the deployment environment…etc.

As mentioned earlier, enhancing the QoS parameters, result in enhancing the user satisfaction level, which can be considered as an enhancement in the QoE according to the definition of QoE that established throughout this work.

Furthermore, the deployment of intelligent small cells enables MNOs to provide customized services to end users in plug-and-play form. Based on the presented work, there are several aspects of future work that could be performed.

- ✓ Further experimental testing of the proposed services in multiple environments over the fully developed IoRL platform to compare the obtained proof of concepts results with actual results.
- ✓ Further investigation is required to study the capability of IHIPGW to scale in handling the increasing number of users, their mobility, and providing fine-grain access control.
- ✓ Further analysis is required to the proposed services in terms of latency and link loads in a very large network, including the side effect of sending the first packet of each new flow to the controller for service discovery and routing configuration.
- ✓ Include Artificial Intelligence algorithms to analyze the available information about end users, available services, propagation environment…etc. and come up with informed decisions about future expansions and performance enhancements.

- ✓ Performing a simulation study about the proposed VGW, to examine its performance as another step closer to the actual deployment in MNO's environment.
- ✓ Automation solutions need to be introduced to the proposed services and included as an essential part of the IHIPGW platform for enabling the reactive auto-scaling feature for all the services.

References

[1] C. Y. Oh, M. Y. Chung, H. Choo, and T. J. Lee, "Resource allocation with partitioning criterion for macro-femto overlay cellular networks with fractional frequency reuse," *Wirel. Pers. Commun.*, vol. 68, no. 2, pp. 417–432, Jan. 2013.

[2] B. NGMN Alliance, R. El Hattachi, and J. Erfanian, "NGMN 5G White Paper," 2015.

[3] "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2017–2022 White Paper - Cisco." [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/serviceprovider/visual-networking-index-vni/white-paper-c11-738429.html. [Accessed: 03-Dec-2019].

[4] R. Chayapathi, S. F. Hassan, and P. Shah, Network Functions Virtualization (NFV) with a Touch of SDN: Netw Fun Vir (NFV ePub_1. Addison-Wesley Professional, 2016.

[5] M. R. Sama, L. M. Contreras, J. Kaippallimalil, I. Akiyoshi, H. Qian, and H. Ni, "Software-defined control of the virtualized mobile packet core," *IEEE Commun. Mag.*, vol. 53, no. 2, pp. 107–115, 2015.

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

[7] W. Stallings, Foundations of modern networking: SDN, NFV, QoE, IoT, and Cloud. Addison-Wesley Professional, 2015.

[8] A. Bradai, K. Singh, T. Ahmed, and T. Rasheed, "Cellular software defined networking: A framework," *IEEE Commun. Mag.*, vol. 53, no. 6, pp. 36–43, Jun. 2015.

[9] F. Hu, *Network Innovation through OpenFlow and SDN: Principles and Design*. Crc Press, 2014.

[10] P. A. Trautman, *Designing and Building a Hybrid Cloud Deliver Automation, Visibility, and Management Consistency in a Multi-Cloud World*. 2018.

[11] "National Institute of Standards and Technology | NIST." [Online]. Available: https://www.nist.gov/. [Accessed: 09-Dec-2019].

[12] H. T. Dinh, C. Lee, D. Niyato, and P. Wang, "A survey of mobile cloud computing: architecture, applications, and approaches," *Wirel. Commun. Mob. Comput.*, vol. 13, no. 18, pp. 1587–1611, Dec. 2013.

[13] A. Banerjee, X. Chen, J. Erman, V. Gopalakrishnan, S. Lee, and J. Van Der Merwe, "MOCA: A lightweight mobile cloud offloading architecture," in *Proceedings of the Annual International Conference on Mobile Computing and Networking, MOBICOM*, 2013, pp. 11–16.

[14] T. Taleb, K. Samdanis, B. Mada, H. Flinck, S. Dutta, and D. Sabella, "On Multi-Access Edge Computing: A Survey of the Emerging 5G Network Edge Cloud Architecture and Orchestration," *IEEE Commun. Surv. Tutorials*, vol. 19, no. 3, pp. 1657–1681, 2017.

[15] J. Cho, B. Nguyen, A. Banerjee, R. Ricci, J. der Merwe, and K. Webb, "SMORE: software-defined networking mobile offloading architecture," in *Proceedings*

of the 4th workshop on All things cellular: operations, applications, & challenges, 2014, pp. 21–26.
[16] M. T. Beck, S. Feld, A. Fichtner, C. Linnhoff-Popien, and T. Schimper, "ME-VoLTE: Network functions for energy-efficient video transcoding at the mobile edge," in *2015 18th International Conference on Intelligence in Next Generation Networks, ICIN 2015*, 2015, pp. 38–44.
[17] X. Chen, L. Jiao, W. Li, and X. Fu, "Efficient Multi-User Computation Offloading for Mobile-Edge Cloud Computing," *IEEE/ACM Trans. Netw.*, vol. 24, no. 5, pp. 2795–2808, Oct. 2016.
[18] N. Takahashi, H. Tanaka, and R. Kawamura, "Analysis of process assignment in multi-tier mobile cloud computing and application to edge accelerated web browsing," in *Proceedings - 2015 3rd IEEE International Conference on Mobile Cloud Computing, Services, and Engineering, MobileCloud 2015*, 2015, pp. 233–234.
[19] D. O. Mau, T. Taleb, and M. Chen, "MM3C: Multi-source mobile streaming in cache-enabled content-centric networks," in *2015 IEEE Global Communications Conference (GLOBECOM)*, 2015, pp. 1–6.
[20] X. Sun and N. Ansari, "EdgeIoT: Mobile Edge Computing for the Internet of Things," *IEEE Commun. Mag.*, vol. 54, no. 12, pp. 22–29, Dec. 2016.
[21] L. Gu, D. Zeng, S. Guo, A. Barnawi, and Y. Xiang, "Cost efficient resource management in fog computing supported medical cyber-physical system," *IEEE Trans. Emerg. Top. Comput.*, vol. 5, no. 1, pp. 108–119, Jan. 2017.
[22] M. Satyanarayanan *et al.*, "Edge analytics in the internet of things," *IEEE Pervasive Comput.*, vol. 14, no. 2, pp. 24–31, Apr. 2015.
[23] J. O. Fajardo, I. Taboada, and F. Liberal, "Improving content delivery efficiency through multi-layer mobile edge adaptation," *IEEE Netw.*, vol. 29, no. 6, pp. 40–46, Nov. 2015.
[24] "GSMA Intelligence." [Online]. Available: https://www.gsmaintelligence.com/. [Accessed: 20-Dec-2019].
[25] "Population Clock." [Online]. Available: https://www.census.gov/popclock/?sec_ak_reference=18.04efdd17.1576882046.c8d4b908. [Accessed: 20-Dec-2019].
[26] C. E. Shannon, "Communication in the Presence of Noise," *Proc. IRE*, vol. 37, no. 1, pp. 10–21, 1949.
[27] "Game changing economics for Small Cell Deployment. Amdocs OSS White Paper October PDF Free Download." [Online]. Available: https://www.census.gov/popclock/?sec_ak_reference=18.04efdd17.1576882046.c8d4b908. [Accessed: 20-Dec-2019].
[28] H. Claussen, D. Lopez-Perez, L. Ho, R. Razavi, and S. Kucera, *Small cell networks: deployment, management, and optimization*. John Wiley & Sons, 2017.
[29] H. Claussen, L. T. W. Ho, and L. G. Samuel, "An overview of the femtocell concept," *Bell Labs Tech. J.*, vol. 13, no. 1, pp. 221–245, May 2008.
[30] V. Chandrasekhar, J. Andrews, and A. Gatherer, "Femtocell Networks: A Survey," Mar. 2008.
[31] J. G. Andrews, H. Claussen, M. Dohler, S. Rangan, and M. C. Reed, "Femtocells: Past, Present, and Future," *IEEE J. Sel. AREAS Commun.*, vol. 30, no. 3, p. 497, 2012.
[32] "BeFEMTO: NEC Laboratories Europe | NEC." [Online]. Available:

https://uk.nec.com/en_GB/emea/about/neclab_eu/projects/befemto.html. [Accessed: 20-Jan-2020].
[33] "Distributed computing, storage and radio resource allocation over cooperative femtocells | TROPIC Project | FP7 | CORDIS | European Commission." [Online]. Available: https://cordis.europa.eu/project/id/318784. [Accessed: 20-Jan-2020].
[34] "Small Cell Forum Releases." [Online]. Available: https://scf.io/en/index.php. [Accessed: 20-Dec-2019].
[35] H. Holma, A. Toskala, and J. Reunanen, "LTE Small Cell Optimization: 3GPP Evolution to Release 13," p. 464, 2016.
[36] M. Pathan, R. K. Sitaraman, and D. Robinson, *Advanced content delivery, streaming, and cloud services*. John Wiley & Sons, 2014.
[37] F. Al-Turjman, E. Ever, and H. Zahmatkesh, "Small cells in the forthcoming 5G/IoT: Traffic modelling and deployment overview," *IEEE Commun. Surv. Tutorials*, vol. 21, no. 1, pp. 28–65, Jan. 2019.
[38] S. Shew Ciena Canada, "ITU-T TELECOMMUNICATION STANDARDIZATION SECTOR OF ITU GSTR-TN5G Transport network support of IMT-2020/5G," no. February 2018, 2018.
[39] J. Feng, Q. Zhang, G. Dong, P. Cao, and Z. Feng, "An approach to 5G wireless network virtualization: Architecture and trial environment," in *IEEE Wireless Communications and Networking Conference, WCNC*, 2017.
[40] G. Tseliou, F. Adelantado, and C. Verikoukis, "Scalable RAN Virtualization in Multitenant LTE-A Heterogeneous Networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 8, pp. 6651–6664, Aug. 2016.
[41] E. J. Kitindi, S. Fu, Y. Jia, A. Kabir, and Y. Wang, "Wireless Network Virtualization with SDN and C-RAN for 5G Networks: Requirements, Opportunities, and Challenges," *IEEE Access*, vol. 5, pp. 19099–19115, 2017.
[42] P. A. Frangoudis, L. Yala, A. Ksentini, and T. Taleb, "An architecture for on-demand service deployment over a telco CDN," in *2016 IEEE International Conference on Communications, ICC 2016*, 2016.
[43] S. Retal, M. Bagaa, T. Taleb, and H. Flinck, "Content delivery network slicing: QoE and cost awareness," in *IEEE International Conference on Communications*, 2017.
[44] Y. Jin, Y. Wen, and C. Westphal, "Optimal Transcoding and Caching for Adaptive Streaming in Media Cloud: An Analytical Approach," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 12, pp. 1914–1925, Dec. 2015.
[45] E. Zeydan *et al.*, "Big data caching for networking: Moving from cloud to edge," *IEEE Commun. Mag.*, vol. 54, no. 9, pp. 36–42, 2016.
[46] D. Liu, B. Chen, C. Yang, and A. F. Molisch, "Caching at the wireless edge: Design aspects, challenges, and future directions," *IEEE Commun. Mag.*, vol. 54, no. 9, pp. 22–28, Sep. 2016.
[47] T. Taleb, S. Dutta, A. Ksentini, M. Iqbal, and H. Flinck, "Mobile edge computing potential in making cities smarter," *IEEE Commun. Mag.*, vol. 55, no. 3, pp. 38–43, Mar. 2017.
[48] K. Bian *et al.*, "Learning at the Edge: Smart Content Delivery in Real World Mobile Social Networks," *IEEE Netw.*, vol. 33, no. 4, pp. 208–215, Jul. 2019.
[49] S. Subramanian and S. Voruganti, Software-defined networking (SDN) with OpenStack. Packt Publishing Ltd, 2016.
[50] B. Silverman and M. Solberg, *OpenStack for Architects: Design production-*

ready private cloud infrastructure. Packt Publishing Ltd, 2018.
[51] "Rackspace: Managed Dedicated & Cloud Computing Services." [Online]. Available: https://www.rackspace.com/. [Accessed: 23-Dec-2019].
[52] "The OpenStack Foundation." [Online]. Available: https://www.openstack.org/foundation. [Accessed: 23-Dec-2019].
[53] J. Denton, *Learning OpenStack Networking: Build a solid foundation in virtual networking technologies for OpenStack-based clouds*. Packt Publishing Ltd, 2018.
[54] "OpenStack Docs: Queens." [Online]. Available: https://docs.openstack.org/queens/. [Accessed: 29-Dec-2019].
[55] J. Cosmas *et al.*, "A scaleable and license free 5G internet of radio light architecture for services in train stations," in *24th European Wireless 2018 "Wireless Futures in the Era of Network Programmability", EW 2018*, 2018.
[56] X. Xu, Y. Jiang, T. Flach, E. Katz-Bassett, D. Choffnes, and R. Govindan, "Investigating transparent web proxies in cellular networks," in *International Conference on Passive and Active Network Measurement*, 2015, pp. 262–276.
[57] P. Goransson, C. Black, and T. Culver, *Software defined networks: a comprehensive approach*. Morgan Kaufmann, 2016.
[58] J. H. Cox *et al.*, "Advancing software-defined networks: A survey," *IEEE Access*, vol. 5, pp. 25487–25526, Oct. 2017.
[59] V. G. Nguyen, A. Brunstrom, K. J. Grinnemo, and J. Taheri, "SDN/NFV-Based Mobile Packet Core Network Architectures: A Survey," *IEEE Commun. Surv. Tutorials*, vol. 19, no. 3, pp. 1567–1602, 2017.
[60] G. Shen, R. Zetik, and R. S. Thomä, "Performance comparison of TOA and TDOA based location estimation algorithms in LOS environment," in *5th Workshop on Positioning, Navigation and Communication 2008, WPNC'08*, 2008, pp. 71–78.
[61] R. M. Buehrer, H. Wymeersch, and R. M. Vaghefi, "Collaborative Sensor Network Localization: Algorithms and Practical Issues," *Proceedings of the IEEE*, vol. 106, no. 6. Institute of Electrical and Electronics Engineers Inc., pp. 1089–1114, Jun-2018.
[62] A. Arafa, S. Dalmiya, R. Klukas, and J. F. Holzman, "Angle-of-arrival reception for optical wireless location technology," *Opt. Express*, vol. 23, no. 6, p. 7755, Mar. 2015.
[63] C. Huang and X. Zhang, "Impact and feasibility of darklight LED on indoor visible light positioning system," in *2017 IEEE 17th International Conference on Ubiquitous Wireless Broadband, ICUWB 2017 - Proceedings*, 2018, vol. 2018-January, pp. 1–5.
[64] "Ryu SDN Framework." [Online]. Available: https://osrg.github.io/ryu/. [Accessed: 20-Jan-2020].
[65] C. (Christopher S. . Lewis and S. Pickavance, *Selecting MPLS VPN services*. Cisco, 2006.
[66] M. Amani, A. Aijaz, Naziruddin, and A. H. Aghvami, "On mobile data offloading policies in heterogeneous wireless networks," in *IEEE Vehicular Technology Conference*, 2013.
[67] N. Jawad, M. Salih, and J. Cosmas, "Media Casting as a Service: Industries Convergence Opportunity and Caching Service for 5G Indoor gNB," *IEEE Trans. Broadcast.*, pp. 1–10, 2020.

[68] Erl, Thomas, Ricardo Puttini, and Zaigham Mahmood. Cloud computing: concepts, technology, & architecture. Pearson Education, 2013.
[69] T. Erl, R. Puttini, and Z. Mahmood, *Cloud computing: concepts, technology, & architecture*. Pearson Education, 2013.
[70] D. Huang and H. Wu, *Mobile cloud computing: foundations and service models*. Morgan Kaufmann, 2017.
[71] B. J. Chang, S. H. Liou, and Y. H. Liang, "Cooperative communication in ultra-dense small cells toward 5G cellular communication," in *2017 8th IEEE Annual Information Technology, Electronics and Mobile Communication Conference, IEMCON 2017*, 2017, pp. 365–371.
[72] X. Wang, M. Chen, T. Taleb, A. Ksentini, and V. C. M. Leung, "Cache in the air: Exploiting content caching and delivery techniques for 5G systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 131–139, 2014.
[73] S. Zhang, P. He, K. Suto, P. Yang, L. Zhao, and X. Shen, "Cooperative Edge Caching in User-Centric Clustered Mobile Networks," *IEEE Trans. Mob. Comput.*, vol. 17, no. 8, pp. 1791–1805, 2018.
[74] M. Furqan, C. Zhang, W. Yan, A. Shahid, M. Wasim, and Y. Huang, "A collaborative hotspot caching design for 5g cellular network," *IEEE Access*, vol. 6, pp. 38161–38170, 2018.
[75] K. Poularakis, G. Iosifidis, V. Sourlas, and L. Tassiulas, "Multicast-aware caching for small cell networks," *IEEE Wirel. Commun. Netw. Conf. WCNC*, vol. 3, pp. 2300–2305, 2014.
[76] L. Lei, X. Xiong, L. Hou, and K. Zheng, "Collaborative Edge Caching through Service Function Chaining: Architecture and Challenges," *IEEE Wirel. Commun.*, vol. 25, no. 3, pp. 94–102, 2018.
[77] C. A. Garcia-Perez and P. Merino, "Enabling low latency services in standard LTE Networks," *Proc. - IEEE 1st Int. Work. Found. Appl. Self-Systems, FAS-W 2016*, pp. 248–255, 2016.
[78] Y. C. Hu, M. Patel, D. Sabella, N. Sprecher, and V. Young, "ETSI White Paper #11 Mobile Edge Computing - a key technology towards 5G - etsi_wp11_mec_a_key_technology_towards_5g.pdf," *ETSI White Pap. No. 11 Mob.*, no. 11, pp. 1–16, 2015.
[79] J. Cosmas, N. Jawad, M. Salih, S. Redana, and O. Bulakci, "5G PPP Architecture Working Group View on 5G Architecture," 2019.
[80] J. Cosmas *et al.*, "A 5G Radio-Light SDN Architecture for Wireless and Mobile Network Access in Buildings," *IEEE 5G World Forum, 5GWF 2018 - Conf. Proc.*, pp. 135–140, 2018.
[81] N. Jawad *et al.*, "Smart Television Services Using NFV/SDN Network Management," *IEEE Trans. Broadcast.*, vol. 65, no. 2, pp. 404–413, 2019.
[82] V. Q. Rodriguez and F. Guillemin, "Towards the deployment of a fully centralized Cloud-RAN architecture," *2017 13th Int. Wirel. Commun. Mob. Comput. Conf. IWCMC 2017*, pp. 1055–1060, 2017.
[83] I. Parvez, A. Rahmati, I. Guvenc, A. I. Sarwat, and H. Dai, "A survey on low latency towards 5G: RAN, core network and caching solutions," *IEEE Commun. Surv. Tutorials*, vol. 20, no. 4, pp. 3098–3130, 2018.
[84] A. Rostami *et al.*, "Orchestration of RAN and transport networks for 5G: An SDN approach," *IEEE Commun. Mag.*, vol. 55, no. 4, pp. 64–70, 2017.
[85] A. Burr, M. Bashar, and D. Maryopi, "Ultra-dense Radio Access Networks for Smart Cities: Cloud-RAN, Fog-RAN and "cell-free" Massive

MIMO," Nov. 2018.
[86] TSGR, "TS 137 340 - V15.3.0 - Universal Mobile Telecommunications System (UMTS); LTE; 5G; NR; Multi-connectivity; Overall description; Stage-2 (3GPP TS 37.340 version 15.3.0 Release 15)," 2018.
[87] M. Polese, M. Giordani, M. Mezzavilla, S. Rangan, and M. Zorzi, "Improved Handover Through Dual Connectivity in 5G mmWave Mobile Networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 9, pp. 2069–2084, 2017.
[88] M. B. Krishna and J. L. Mauri, *Advances in mobile computing and communications: perspectives and emerging trends in 5G networks*. CRC Press, 2016.
[89] J. Wu, Z. Zhang, Y. Hong, and Y. Wen, "Cloud radio access network (C-RAN): a primer," *IEEE Netw.*, vol. 29, no. 1, pp. 35–41, 2015.
[90] P. Rost *et al.*, "Mobile network architecture evolution toward 5G," *IEEE Commun. Mag.*, vol. 54, no. 5, pp. 84–91, 2016.
[91] H. Zhang, N. Liu, X. Chu, K. Long, A. H. Aghvami, and V. C. M. Leung, "Network Slicing Based 5G and Future Mobile Networks: Mobility, Resource Management, and Challenges," *IEEE Commun. Mag.*, vol. 55, no. 8, pp. 138–145, 2017.
[92] J. Heinonen, P. Korja, T. Partti, H. Flinck, and P. Pöyhönen, "Mobility management enhancements for 5G low latency services," *2016 IEEE Int. Conf. Commun. Work. ICC 2016*, no. 5GArch, pp. 68–73, 2016.
[93] X. Ge, S. Tu, G. Mao, C. X. Wang, and T. Han, "5G Ultra-Dense Cellular Networks," *IEEE Wirel. Commun.*, vol. 23, no. 1, pp. 72–79, Feb. 2016.
[94] A. Behnad and X. Wang, "Virtual Small Cells Formation in 5G Networks," *IEEE Commun. Lett.*, vol. 21, no. 3, pp. 616–619, 2017.
[95] M. R. Sama, S. B. H. Said, K. Guillouard, and L. Suciu, "Enabling network programmability in LTE/EPC architecture using OpenFlow," in *2014 12th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks, WiOpt 2014*, 2014, pp. 389–396.
[96] D. Kreutz, F. M. V Ramos, P. E. Verissimo, C. E. Rothenberg, S. Azodolmolky, and S. Uhlig, "Software-defined networking: A comprehensive survey," *Proc. IEEE*, vol. 103, no. 1, pp. 14–76, 2014.
[97] J. Cosmas *et al.*, "A Scalable and License Free 5G Internet of Radio Light Architecture for Services in Homes Businesses," in *IEEE International Symposium on Broadband Multimedia Systems and Broadcasting, BMSB*, 2018, vol. 2018-June.
[98] M. G. Kibria, K. Nguyen, G. P. Villardi, K. Ishizu, and F. Kojima, "Next Generation New Radio Small Cell Enhancement: Architectural Options, Functionality and Performance Aspects," *IEEE Wirel. Commun.*, vol. 25, no. 4, pp. 120–128, 2018.
[99] M. Salih, N. Jawad, and J. Cosmas, "Simulation and Performance Analysis of Software-Based Mobile Core Network Architecture (SBMCNA) Using OMNeT++," in *IEEE International Symposium on Broadband Multimedia Systems and Broadcasting, BMSB*, 2018, vol. 2018-June.
[100] K. Ali, A. Alkhatar, N. Jawad, and J. Cosmas, "IoRL Indoor Location Based Data Access, Indoor Location Monitoring Guiding and Interaction Applications," in *IEEE International Symposium on Broadband Multimedia Systems and Broadcasting, BMSB*, 2018, vol. 2018-June.
[101] 5GPPP Architecture Working Group, "view on 5G architecture," *White Pap.*,

no. June, 2019.
[102] “3rd Generation Partnership Project; Technical Specification Group Radio Access Network; NG-RAN; F1 general aspects and principles (Release 15) Keywords,” 2018.
[103] N. Jawad *et al.*, “Indoor Unicasting/Multicasting service based on 5G Internet of Radio Light network paradigm,” in *2019 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, 2019, pp. 1–6.
[104] TSGR, “TS 137 324 - V15.1.0 - LTE; 5G; Evolved Universal Terrestrial Radio Access (E-UTRA) and NR; Service Data Adaptation Protocol (SDAP) specification (3GPP TS 37.324 version 15.1.0 Release 15),” 2018.
[105] TSGR, “TS 136 331 - V15.3.0 - LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC); Protocol specification (3GPP TS 36.331 version 15.3.0 Release 15),” 2018.
[106] Lte, “TS 136 323 - V14.3.0 - LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Packet Data Convergence Protocol (PDCP) specification (3GPP TS 36.323 version 14.3.0 Release 14),” 2017.
[107] M. Salih, “Software defined networking integration with mobile network towards scalable and programmable core,” Brunel University London, 2019.
[108] I. Widjaja, P. Bosch, and H. La Roche, “Comparison of MME signaling loads for long-term-evolution architectures,” in *IEEE Vehicular Technology Conference*, 2009.
[109] V. -G Nguyen and Y. Kim, “Proposal and evaluation of SDN-based mobile packet core networks,” *Eurasip J. Wirel. Commun. Netw.*, vol. 2015, no. 1, Dec. 2015.