SLICING-BASED RESOURCE ALLOCATION AND MOBILITY MANAGEMENT FOR EMERGING WIRELESS NETWORKS

ALI SAEED DAYEM ALFOUDI

BSc, MSc

A thesis submitted in partial fulfilment of the requirements of Liverpool John Moores University for the degree of Doctor of Philosophy

June 2018

DECLARATION

I, Ali Saeed Dayem Alfoudi, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm this has been indicated in the thesis.

Ali Saeed Dayem Alfoudi

ACKNOWLEDGEMENT

PhD is a rewarding but challenging journey, which would not be possible without the help of many people.

First and foremost, I would like to express my sincere gratitude to my supervisors, Dr. Gyu Myoung Lee, Dr. Rubem Pereira and Dr.Thar Baker Shamsa for the continuous support, advice, and guidance throughout my candidature. They have built and directed an environment that granted me an opportunity to learn and practice research skills, meet and collaborate with brilliant researchers, and transfer the long journey of PhD to a great and lovely experience.

Special thanks go to Dr. Abayomi Otebolaku. He offered me guidance when I was struggling to choose an approach for implementing feasible solutions. Besides, I would like to thank Dr. S.H. Shah Newaz for proof-reading, discussion and his comments on my work. His critical comments helped me to improve my writing skills.

I would like to thank my friends and colleagues in Liverpool John Moores University. It has been such a pleasure and a privilege to work with you all.

I appreciate the funding source that made my PhD work possible, the Ministry of Higher Education in Iraq and its representative in the United Kingdom the Iraqi cultural attaché in London.

I sincerely thank my family: my parents and to my brothers and sister for supporting me spiritually throughout writing this thesis and my life in general.

Last but not the least, I would like to extend my heartfelt thanks to my lovely kids for making my life better and my wife for her encouragement during rough times, as we taste all the sweetsness and bittersness of pursuing a PhD. I cannot imagine a greater fortune other than having them in my life. Their love and care are indispensable to any of my achievements.

ABSTRACT

The proliferation of smart mobile devices and user applications has continued to contribute to the tremendous volume of data traffic in cellular networks. Moreover, with the feature of heterogeneous connectivity interfaces of these smart devices, it becomes more complex for managing the traffic volume in the context of mobility. To surmount this challenge, service and resource providers are looking for alternative mechanisms that can successfully facilitate managing network resources and mobility in a more dynamic, predictive and distributed manner. New concepts of network architectures such as Software-Defined Network (SDN) and Network Function Virtualization (NFV) have paved the way to move from static to flexible networks. They make networks more flexible (i.e., network providers capable of on-demand provisioning), easily customizable and cost effective. In this regard, network slicing is emerging as a new technology built on the concepts of SDN and NFV. It splits a network infrastructure into isolated virtual networks and allows them to manage network resources based on their requirements and characteristics. Most of the existing solutions for network slicing are facing challenges in terms of resource and mobility management. Regarding resource management, it creates challenges in terms of provisioning network throughput, end-to-end delay, and fairness resources allocation for each slice, whereas, in the case of mobility management, due to the rapid change of user mobility the network slice operator would like to hold the mobility controlling over its clients across different access networks, rather than the network operator, to ensure better services and user experience.

In this thesis, we propose two novel architectural solutions to solve the challenges identified above. The first proposed solution introduces a Network Slicing Resource Management (NSRM) mechanism that assigns the required resources for each slice, taking into consideration resource isolation between different slices. The second proposed

solution provides a Mobility Management architecture-based Network Slicing (MMNS) where each slice manages its users across heterogeneous radio access technologies such as WiFi, LTE and 5G networks. In MMNS architecture, each slice has different mobility demands (e.g,. latency, speed and interference) and these demands are governed by a network slice configuration and service characteristics.

In addition, NSRM ensures isolating, customizing and fair sharing of distributed bandwidths between various network slices and users belonging to the same slice depending on different requirements of each one. Whereas, MMNS is a logical platform that unifies different Radio Access Technologies (RATs) and allows all slices to share them in order to satisfy different slice mobility demands.

We considered two software simulations, namely OPNET Modeler and OMNET++, to validate the performance evaluation of the thesis contributions. The simulation results for both proposed architectures show that, in case of NSRM, the resource blocking is approximately 35% less compared to the legacy LTE network, which it allows to accommodate more users. The NSRM also successfully maintains the isolation for both the inter and intra network slices. Moreover, the results show that the NSRM is able to run different scheduling mechanisms where each network slice guarantee perform its own scheduling mechanism and simultaneously with other slices.

Regarding the MMNS, the results show the advantages of the proposed architecture that are the reduction of the tunnelling overhead and the minimization of the handover latency. The MMNS results show the packets delivery cost is optimal by reducing the number of hops that the packets transit between a source node and destination. Additionally, seamless session continues of a user IP-flow between different access networks interfaces has been successfully achieved.

List of Publications

- **Journals**

A. S. D. Alfoudi, S. H. S. Newaz, A. Otebolaku, G. M. Lee, and R. Pereira, "An Efficient Resource Management Mechanism for Network Slicing in LTE Network," Computer Communications, pp. 0–42. **(Under Review)**

- **International Conferences and Workshops**

A. S. D. Alfoudi, M. Dighriri, G. M. Lee, R. Pereira, and F. P. Tso, "Traffic management in LTE-WiFi slicing networks," in Innovations in Clouds, Internet and Networks (ICIN), 2017 20th Conference on, 2017, pp. 268–273. **(Awarded as a Best Paper)**

A. S. D. Alfoudi, M. Dighriri, A. Otebolaku, R. Pereira and G. M. Lee, "Mobility Management Architecture in Different RATs Based Network Slicing," in The 32-nd IEEE International Conference on Advanced Information Networking and Applications, 2018.

A. S. D. Alfoudi, M. Dighriri and G. M. Lee, "Seamless LTE-WiFi Architecture for Offloading the Overloaded LTE with Efficient UE Authentication," in Developments in eSystems Engineering (DeSE), 2016 9th International Conference on, 2016, pp. 118–122.

M. Dighriri, A. S. D. Alfoudi, G. M. Lee, and T. Baker, "Resource Allocation Scheme in 5G Network Slices," in The 32-nd IEEE International Conference on Advanced Information Networking and Applications, 2018.

M. Dighriri, A. S. D. Alfoudi, G. M. Lee, T. Baker, and R. Pereira, "Comparison Data Traffic Scheduling Techniques for Classifying QoS over 5G Mobile Networks," in Advanced Information Networking and Applications Workshops (WAINA), 2017 31st International Conference on, 2017, pp. 492–497.

M. Dighriri, A. S. D. Alfoudi, G. M. Lee, and T. Baker, "Data Traffic Model in Machine to Machine Communications over 5G Network Slicing," in Developments in eSystems Engineering (DeSE), 2016 9th International Conference on, 2016, pp. 239–244.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

3GPP	3rd Generation Partnership Project
4G	Fourth Generation
5G	Fifth Generation
5GPPP	5G Infrastructure Public Private Partnership
AP	Access Point
API	Application Programming Interfaces
AuC	Authentication Centre
BE	Best effort
BEMG	BE with Minimum Guarantee
BID	Binding Identity
BS	Base Station
BSS	Business Support System
BSSID	Basic Service Set Identifier
CDPI	Control-Data-Plane Interface
CLI	Command Line Interface
CMaaS	Connectivity Management as a Service
CN	Core Network
COTS	Commercial Off-The-Shelf
CP	Control Plane
CSP	Cloud Service Provider
DG	Dynamic Guarantee
EPC	Evolved Packet Core
ePGW	enhance PGW
E-UTRAN	Evolved Terrestrial Radio Access Network
FDD	Frequency Division Duplex

FG	Fixed Guarantee
GBR	Guarantee Bit Rate
GSM	Global System for Mobile communications
HA	Home Agent
HLR	Home Location Register
HMIPv6	Hierarchical Mobile IPv6
HoA	Home Address
HSS	Home Subscribe System
IETF	Internet Engineering Task Force
IKEv2	Internet Key Exchange
IMM	Individual Mobility Model
IMS	IP Multimedia Subsystem
IMSI	International Mobile Subscriber Identity
IMT-2020	International Mobile Telecommunication system–beyond 2020
InP	Infrastructure Provider
KPI	Key Performance Indicator
LMA	Local Mobility Anchor
LSCM	LTE Slice Controller Manager
LTE	Long Term Evolution
LVAP	Light Virtual Access Point
LWCF	LTE-WiFi Controller Flow
MAC	Medium Access Control
MAG	Mobility Access Gateway
MANO	Management and Network Orchestration
MCS	for Modulation and Coding Schemes
MME	Mobility Management Entity

MMNS	Mobility Management Network Slicing
MN	Mobile Node
M-SCTP	Mobile Stream Control Transmission Protocol
MT	Mobile Terminal
Mux/DeMux	Multiplexing/DeMuliplexing
NBI	Northbound Interface
NFV	Network Function Virtualization
NFVI	NFV Infrastructure
NGBR	None Guarantee Bit Rate
NGMN	Next Generation Mobile Networks
NSI	Network Slice Instance
NSRM	Network Slicing Resource Management
NVS	Network Virtualization Substrate
Ofcom	Office of Communications
OFDMA	Orthogonal Frequency Division Multiple Access
ONF	Open Networking Foundation
OSS	Operations Support System
PaaS	Platform as a Service
PCRF	Policy and Charging Rules Functions
PDCP	Packet data convergence protocol
P-GW	Packets Data Network GateWay
PHY	Physical Layer
PMIPv6	Proxy Mobile IPv6
PRB	Physical Resource Block
PRR	Priority Round Robin
PSRM	Per Slice Resource Management

QAM	Quadrature Amplitude Modulation
QCI	Quality of service Class Identifier
QoE	Quality of Experience
QoS	Quality of Service
RAN	Radio Access Network
RAP	Resource Allocation Policy
RATs	Radio Access Technologies
RC	Resource Computing
RCPU	Resource Computing Per User
RE	Resource Element
RLC	Radio link Control
RRC	Radio resource control
RSRP	Reference Signals Received Power
RSRQ	Reference Signal Received Quality
RSS	Radio Signal Strength
RTT	Round Trip Time
RWP	Random Waypoint
SA	Slice Allocation
SaaS	Software as a Service
SBI	Southbound Interface
SBS	Small cell Base Stations
SC-FDMA	Single Carrier-Frequency Division Multiple Access
SDN	Software-Defined Network
SDU	Service Data Units
SGI	Statistics Gathering Information
S-GW	Serving GateWay

SINR	Signal to Interference & Noise Ratio
SIP	The Session Initiation Protocol
SLA	Service Level Agreement
SNMP	Simple Network Management Protocol
SP	Service Provider
SRT	Slice Resource Tracker
SSID	Service Set Identifier
TCP	Transmission Control Protocol
TDD	Time Division Duplex
TTI	Transmission Time Interval
UDP	User Datagram Protocol
UE	User Equipment
UICC	Universal Integrated Circuit Card
UID	User Information Database
UP	User Plane
UR	User Requests
USIM	Universal Subscriber Identity Module
VB	Virtual Bearer
VM	Virtual Machine
VNF	Virtual Network Functions
VNO	Virtual Network Operator
VXLAN	Virtual Extensible LAN
WF	Weighted Fair

List of Notations

Symbol	Explanation
X	A set of base stations
V	A set of slices
U	A set of users
B_x	The base station spectrum bandwidth
$\eta_{u_i\, x}$	The spectrum bandwidth for user within x
$S\ and\ N$	Represents the average signal and noise power
$L_{u_i\, x}$	The indication of user associated to x
$Y_{u_i\, x}$	The percentage of radio resources allocated to user u_i by BS x
$R_{u_i\, x}$	The instantaneous user u_i data rate
δ_{u_i}	The total number of virtual bearers assigned to a user
ΔT	Observation period
ρ_{u_i}	The total user bearer data rate over the period ΔT
$Q_{u_i x}$	The actual data rate load of a user bearer in slice
$(\rho_{u_i})_{t+1}$	The next time round of scheduling allocation to a user data rate
v_B	A slice bandwidth capacity over base station x
V_B	The total slices bandwidth in the base station x
λ_{t+1}	the estimate of PRBs during the $(t+1)$ interval time
α	A smoothing constant

FF_v	The fairness factor of a slice v
ω	The total estimated bandwidth of all slices
φ_v	The total number of PRBs allocated for each BE slice v
n	The number of users in a slice v
z	Represents the excess bandwidth for individual user u in v

Chapter 1: Introduction

Today's and future network providers contend with the exponential growth of network traffic and the heterogeneous network connectivity environment (e.g., LTE and WiFi networks) due to the proliferation of network users and bandwidth-hungry services. According to CISCO, because of the increasing appetite of mobile users for network resources, the mobile network traffic has increased and it is expected to grow to around 70% by 2021 [1, 2]. The unprecedented growth of mobile networks and the intelligence of smart mobile devices pushes the network providers to look for more efficient management mechanisms in terms of network resources and user mobility, in order to introduce the innovative services which are considered as a promising inspiration for the user experience. These smart devices have various wireless interfaces (e.g., LTE and WiFi) to connect to the network for accessing different network services. However, different issues due to the use of traditional mechanisms and protocols inhibit the development of network resources as well as mobility management. For example, the traditional network resource allocation mechanisms are inadequate to meet new services requirement of network slicing where the resource requirements dynamically change according to the change of user requests of a slice; the user mobility demands also fluctuate depending on mobility services and access network conditions. Therefore, in order to cope with the development of a large number of complex and intelligent user applications and network services, it is required to redesign a new network architecture for sharing network resources and management of user mobility services. Since the aim of service providers is to efficiently utilize network resources and connectivity, virtual network architecture has been highlighted as a vital key for abstracting network resources, creating logical end-

to-end network and for complying with network services business and customers' needs. Therefore, it is essential to ensure the isolation sharing resources and network utilization optimization. The virtual architecture helps the cellular Infrastructure Providers (InPs) to run the network in a controlled manner and virtually establish different virtual networks on the same infrastructure in such a way as to avoid unpredicted conditions and service failures. The InP needs to engage an emerging technology such as network slicing for establishing the different virtual networks on the same infrastructure. Network slicing is the conceptual network architecture that enables slicing of logical network resources into different end-to-end logical network across Radio Access Network (RAN) and Core Network (CN).

In order to facilitate such flexible resource allocation, dynamic network configuration and cost-effective operation in a network, Software Defined Network (SDN) and Network Function Virtualization (NFV) open up new opportunities [3]. SDN is an emerging technology where a control plane is decoupled successfully from a data plane, making a network programmable and cost effective. SDN offers several advantages over conventional hardware-centric networks, including on-demand traffic forwarding policy, reduced cost and better QoS. NFV is a revitalizing technology of future networks, which allows a physical network infrastructure to be shared among coexistence of multiple network instances simultaneously. SDN and NFV partition the traditional networks into virtual elements, which are logically linked together [4].

To enable multiple virtual elements to share a common physical network, the network slicing mechanism comes into play. Network slicing enables the slicing of a virtual network across Radio Access Network (RAN) and a Core Network (CN). It is a conceptual architecture, which aims to share a common physical infrastructure among

multiple virtual networks using the same principles applied in SDN and NFV [5, 6]. In particular, there are some important requirements, which should be met when applying network slicing. These requirements are summarized as follows:

Isolation among slices: isolation means the ability of restricting the impact of a slice on other slices in the same network, even if they share the same infrastructure. That is to say, if there is any change of resource status in a slice (e.g., traffic load change), such a change should not influence the allocated resources of other slices.

Customization: resource management of each slice can be operated independently to meet the best individual service requirements. That is, the admission control policy of a slice can be different from the other slices.

Efficient resources utilization: maximizing the utilization of channel resources as much as possible would in turn allow increasing the capacity of a base station and efficiently utilizing a channel transmission.

This thesis focuses on providing two major architectural solutions based on the concept of network slicing. The first solution is the Network Slicing Resources Management (NSRM) architecture. The NSRM aims to ensure the isolation of allocated resources, fair resource sharing and customized network slice configuration. The second solution is the Mobility Management architecture-based Network Slicing (MMNS) in heterogeneous access network environment. The MMNS aims at creating a unified platform of different RATs and allows each slice to control its own users in such a way as to satisfy their different mobility demands.

1.1. Research Motivation and Problem Statement

The emerging wireless networks such as 5G are expected to be built based on the current 4G technology and providing a surplus of network services with different performance requirements. The 5G network is anticipated to support diverse use-cases as well as special service scenarios promising simultaneous satisfaction of different service requirements of these use cases and scenarios. This emerging technology is in respect offering diverse business partnerships and possesses the capability for supporting services with different set of requirements to engage the third parties and for establishing innovative services and programmability of their network using open source software tools and interfaces. Based on this softwarization environment, 5G is enabled to support multi-tenancy and service-tailored connectivity on the top of shared physical network infrastructure. In this thesis we address the resource management in network slicing and the mobility management in network slicing for heterogeneous networks environment. Accordingly, this subsection is divided into two parts namely: the resources management in network slicing and the mobility management in heterogeneous network-based slicing.

1.1.1. Resource Management in Network Slicing

To date, many research efforts have been conducted aiming to provide better resource management models in mobile networks (e.g., [7, 8, 9, 10]). Some of these works proposed resource allocation mechanisms based on assigning a number of Physical Resource Blocks (PRBs) to each user's request in a cellular network. We can broadly classify a resource management mechanism into two levels: a low-level management model and a high-level management model. The advantage of applying a low-level model is that it is easy to implement because any requested resource gets resource

allocation in units (e.g., a user in cellular network could get 10 units of PRBs). By utilizing a low-level model, it provides accuracy of allocating resource to each resource demand in units. However, it is more difficult for high-level management entities (e.g., operators and service providers) to adopt a low-level management mechanism, because resources in the high-level management model are allocated in proportion (e.g., 30% of total available PRBs).

Looking at the research focus from industries and academia, we envisage that the future network will solely embrace network virtualization. The major factors that have resulted in rapid adoption of network virtualization are: cost-effective sharing of network resources, and high network utilization. In order to gain synergistic benefits of network virtualization, along with designing efficient network architectures, a research effort should focus on an effective resource management mechanism in a virtual network. Future virtualized networks need a new management mechanism that would provide accuracy of resource allocation and guarantee resource isolation. In order to accomplish these objectives, a novel resource management mechanism is required that will take into consideration both the low and high-level management models for resource allocation. On one hand, the major role of the low-level model would be providing PRB based resource allocation in numbers of units, thereby ensuring high accuracy in resource allocation. On the other hand, the high-level model should be capable of ensuring isolation among the dedicated resources.

At this stage, it is important to efficiently map the initial network slicing request onto the substrate of network resources. When the group of users or service providers specifies their service demands (e.g., logical end-to-end network topology, bandwidth, computing resources, storage resources, etc.) in the form of virtual nodes

interconnected by virtual links, then they can launch it into the network provider. Upon receiving the requests, the network provider collaborates with the InP to provision a scheme for establishing a network slice based on requested resources. The network provider and the InP should be wary of the resource demands fluctuating for the current network slice and periodically check it with respect to the re-optimization of substrate network resources usage.

1.1.2. Mobility Management in Heterogeneous Networks

When network development moves towards 5G network it seems to become increasingly heterogeneous. Therefore, a key feature of 5G networks can be the integration between various Radio Access Networks (RANs), providing a mobile device with a 5G-enabler (e.g., at mmWave frequencies) along with other network interfaces, such as 4G LTE network even with possibility of LTE-Unsilenced [11] and WiFi network, that turn out to be great in terms of increasing opportunity to introduce innovative connectivity services to enhance users Quality of Experience (QoE). In contrast, to determine the RAN to which a user should be associated is truly a big challenge for the network. Determining the optimal user connection can be a complex combination problem that can depended on considering many matrices simultaneously, such as the Signal to Interference and Noise Ratio (SIRN) at every user to every Base Station (BS), or the current load at each BS.

Increasingly the densification and heterogeneity of access networks are facing a big challenge in terms of support for mobility and always-on-connectivity, because it is difficult to measure the impact of mobility on network performance [10, 11]. Therefore, it is very important to reduce the number of user handovers between different BSs.

Therefore, it is important to give the network slicing capability of managing different RANs on the same logical network as a way to mitigate the impact of mobility (frequent handovers) and to enable a slice to control its users across different access networks. For instance, according to the Office of Communications of the U.K (Ofcom) [14] there are around 81% of mobile consumers using WiFi network at some point, therefore network operators consider a WiFi network an important player as a method for offloading mobile data traffic.

1.2. Research Aims and Objectives

1.2.1. Research Aims

Recently, the network slicing has been considered a vital player in the context of 5G and beyond networks, where it is an integration of virtual resources (e.g., Virtual Machines) wherein a number of Virtual Network Functions (VNF) are instantiated and logically linked together for establishing end-to-end virtual network. The main goal of this thesis is to facilitate, investigate and design network slicing architectures (namely NSRM and MMNS), which are capable of managing virtual network resources (in both RAN and CN), as well as managing user mobility in different RANs and maintaining network service continuity during mobility (seamless connectivity).

1.2.2. Research Objectives

The specific objectives for realising the goal of the thesis as stated above are:

To provide provisioning resource for a network slicing that satisfies resource requirements: satisfying the resource requirements of slices and users in respect of their Service Level Agreement (SLA) that in turn will lead to meeting the users' QoE and maximize the revenue of both Infrastructure Provider (InP) and a slice owner (e.g.,

Service Provider (SP)). For example, each network slice has allocated required PBRs in a RAN that satisfies users' service belonging to individual slice.

To provide resources isolation between slices: maintaining resources isolation between slices prevents the deterioration of the network performance. Isolation is the ability of restricting the impact of a slice on other slices in the same network, even if they share the same infrastructure (e.g., each network slice has dedicated VNFs).

To provide different traffic customization of each slice: each network slice has individual requirements depending on the service type and demands configuration. Therefore, each slice owner may prefer to shape its traffic scheduling flows in such a way of ensuring best network performance (e.g., QoS).

To develop Fair sharing distribution resources: distributing allocated resource in such a way that both slices and users fairly share the allocated resources, where each individual slice or user has to get at least the minimum required resources that satisfies the current service. Each InP or slice owner may apply a particular scheduling mechanism for balancing distribution resources in order to optimize network utilization.

To provide seamless IP-flow for session continuity in different RANs: it is needed to keep on-going connections for a user during mobility. The user mobile device can seamlessly exchange flow over different RANs while maintaining the session connections from service degradation, disruption and signalling overhead.

To develop a handover mechanism for keeping user mobility: it has a significant impact on a user's mobility management when there are changes to network attachment points during a movement between different access network(s).

Considering different parameters in the selection of access point mechanism during the handover procedure will enhance network connectivity (e.g., QoS and QoE) by reducing the handover latency and packet loss.

1.3. Contributions to Knowledge

The main contributions of this thesis are summarized as follows.

1. A novel Resource Management of Network Slicing (NSRM) Architecture

As one of the key contributions of the thesis, the NSRM has been developed for virtualizing (network slicing) a cellular network in order to maximize network resource utilization. This architecture facilitates slicing a virtual network into a number of slices each of which is configured based on the SLA of the service requirements of an operator.

2. A dynamic algorithm for the inter tier slices resource allocation

In addition, the thesis developed an algorithm that is based on the exponential smoothing model, for dynamically distributing bandwidth among different slices within an eNodeB to maximize resources utilization.

3. A distributed algorithm for intra tier resource allocation

This algorithm is based on a Max-Min model. It works inside a dedicated slice for managing a distribution function of slice resources between users, in such a way that ensures isolation of slice resources across flows and secures a fair share of minimum bandwidth among the users belonging to the same slice.

4. A novel architecture for mobility management in network slicing (MMNS)

Each network slice can manage its users across heterogeneous radio access technologies such as WiFi, LTE and 5G networks. In this architecture, each slice has different mobility demands and these demands are governed by a network slice configuration and service characteristics. Therefore, our mobility management architecture follows a modular approach where each slice has an individual module that handles mobility functions and enforces the policy of mobility management of the slice, as well as maintaining seamless flow connections for a user across different RANs.

5. A mechanism of selecting an access point for a handover procedure

This mechanism enables us to achieve a seamless connectivity in the heterogeneous environment and selects an appropriate AP allowing the cooperation between a mobile device and controller. This mechanism considers different parameters for selecting a most suitable AP. The parameters include user preferences (e.g., cost and location), services requirements (e.g., audio, video streaming and file sharing applications) and AP capacity in term of user density and throughput.

1.4. Thesis Organisation

The organization of the chapters in this thesis is shown in Figure 1.1. Chapter 2 discusses the thesis background, technologies and surveys related work. Chapter 3 presents a Resource Management of Network Slicing (NSRM) solution based LTE network, detailing how to accommodate network resources between different network slices, using inter and intra resource allocation of network slicing to ensure the isolation. Chapter 4 shows a proposed Mobility Management architecture in Network

Slicing (MMNS), where each slice can manage its users across heterogeneous radio access technologies such as WiFi, LTE and 5G networks. In this architecture, each slice has different mobility demands and these demands are governed by a network slice configuration and service characteristics. The chapter also presents the seamless connectivity model, which keeps user IP-flow sessions continuing across different RANs. Chapter 5 presents systematic design of a handover for supporting user mobility between heterogeneous access networks, detailing with operational steps and policy enforcement of implementing the handover. Chapter 6 presents experimental validation of the proposed frameworks. It also discusses the simulations setup and illustrates the results of a series of experiments conducted to validate the proposed solutions. Finally, chapter 7 presents conclusions and future direction of the thesis.

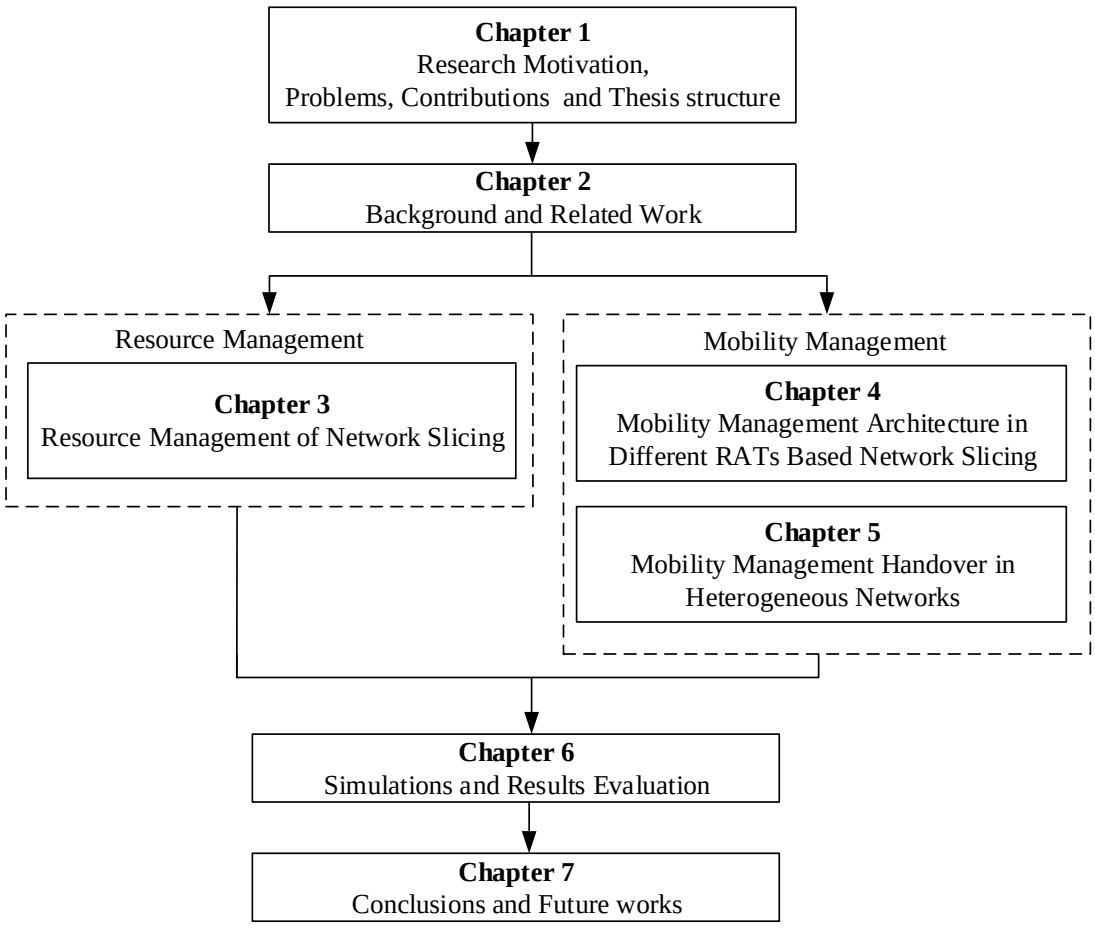


Figure 1.1: The thesis organization

Chapter 2: Background and Related Work

2.1. Introduction

This chapter introduces the background and related work on both network resource management and mobility management for heterogeneous networks, in the context of network slicing. It also discusses how this thesis relates to and differs from the state of the art. Section 2.2 provides an overview of heterogeneous wireless networks. Section 2.3 presents the network slicing concept with different enabling technologies. The network management and orchestration architecture are illustrated in section 2.4. Section 2.5 describes the related work of network slicing in RAN and core network, resource allocation and mobility management.

2.2. Heterogeneous Wireless Networks

In this section, we discuss different wireless network such as LTE, WiFi and 5G networks. In the LTE network, we explain various components in RAN and core network. Regarding WiFi, we describe different architecture from the cellular network view point. Finally, the different aspects of 5G will be discussed.

2.2.1. Long Term Evolution (LTE) Network

Figure 2.1 shows the network architecture of LTE that is comprised of three main components: The User Equipment (UE), The Evolved Terrestrial Radio Access Network (E-UTRAN) and The Evolved Packet Core (EPC).

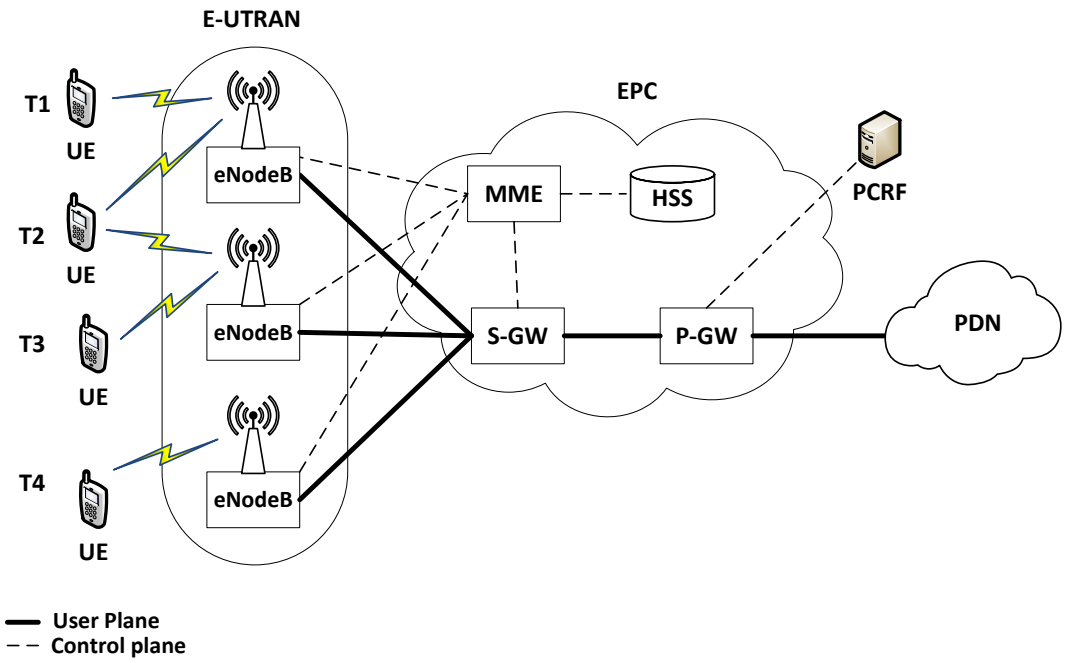


Figure 2.1: LTE Network

2.2.1.1. The User Equipment (UE)

UE is a user mobile device that is compatible with LTE radio specifications to connect with the LTE network via a base station. This equipment has the following important elements:

Mobile Terminal (MT): this module handles all physical communication functions via radio air interface (e.g., antenna).

Universal Integrated Circuit Card (UICC): this module runs an application called Universal Subscriber Identity Module (USIM). It is also known as the SIM card for the LTE network under different names of vendors.

The USIM stores user specific data and profile, such as information about the user's phone number, home network identity and security keys.

2.2.1.2. The Evolved Terrestrial Radio Access Network (E-UTRAN)

The E-UTRAN represents the radio access domain of LTE network. It includes one type of physical equipment that represents the base station, which it calls eNodeB. The E-UTRAN domain consist of either one eNodeB or more, if more than one eNodeB in the domain they link together across X2 interface. The main role of the X2 interface is managing the exchange of control messages between two neighbour eNodeBs in particular circumstances, such as in the user mobility management in the coverage area domain and in the case of load balance between two eNodeBs. Moreover, the E-UTRAN is linked between the user device and core network where the eNodeB connects with the UE via the LTE-UU interface and with the EPC side via the S1 interface to the S-GW as shown in the figure 2.2 [15].

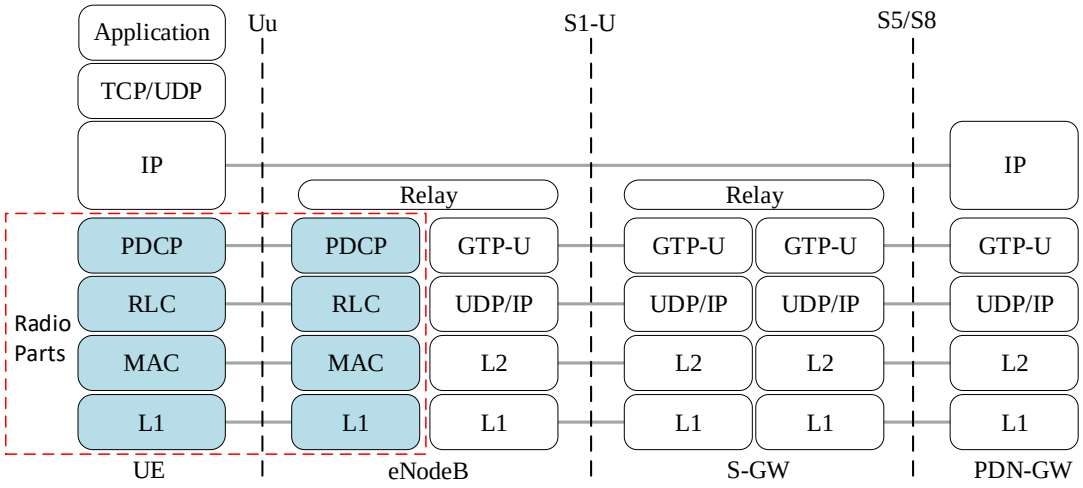


Figure 2.2: LTE protocol stack of user plane.

However, the E-UTRAN is composed of user plane and control plane. The user plane consists of a number of protocols that are used to deliver the actual date flow, whereas the control plane is used to configure user plane layers before actual data flow such as establish the user connections and bearers.

The figure illustrates the user plane protocols across different domains of the LTE network, including the E-UTRAN and various interfaces that link these domains with each other. As can be noticed from the figure, the eNodeB is linked between the radio and wire access network via different protocols. The main protocols of radio interfaces are briefly explained as follows:

Radio resource control (RRC): this layer has many tasks, such as broadcast of system information, RRC connection control which means that a user can not send any data without RRC controlling, state transition of user mode between idle and connected and vice versa, and measurement configuration and reporting during the handover procedure. A detailed explanation of the RRC functions can be found in [15].

Packet data convergence protocol (PDCP): this protocol is responsible for header compression/decompression of IP data packets, maintenance of sequence numbers where it assigns a sequence number of each packet before sending to the next layer (RLC), and security of the data over the air interface. More details about PDCP can found in [15].

Radio link Control (RLC): this protocol has functions for concatenation, segmentation and reassembly of RLC Service Data Units (SDUs). RLC at transmission enhances concatenation or segmentation these data packets dynamically according to the current bandwidth of air interface and RLC reassembly at in the receiving end of this data. A detailed explanation of the RLC functions can be found in [16].

Medium access control (MAC): is responsible of prioritisation among various data streams for a given user where data packets of different user streams have different priority based on their types. The MAC layer applies this priority when it is scheduling

the uplink and downlink flows for a user taking into consideration different QoS over the air interface. A detailed explanation of the MAC functions can be found in [17].

Physical Layer (PHY): PHY layer includes row data before modulation, therefore it has to go for encoding process before modulation and in the same context it has decoding/demodulation for the data. Moreover, the PHY layer has different techniques of measuring the air interface to know a channel quality; that is for many reasons such as mobility, interference and channel noise, etc.

2.2.1.3. The Evolved Packet Core (EPC)

The EPC is the core network and consist of four components that are explained as follows:

Mobility Management Entity (MME): this element controls most of the operations that occur in the EPC. We can say it is the brain of the operation in EPC. The major responsibility of MME is managing a tracking area location when the UE moves in different eNodeB coverage areas. MME interacts with other elements in EPC, such as Home Subscribe System (HSS), S-GW and P-GW [18]. The MME has a functionality to authenticate and authorize the UE. It interacts with the HSS to implement these operations because the HSS kind of databases store all data that are related to those two functionalities of MME. For example, to answer the question of authentication (e.g., the IMSI (International Mobile Subscriber Identity) of UE (the process of verifying)) and for authorization (e.g., roaming authorization). Among its duties, it also gives the key instructions to other node elements in EPC (SGW and P-GW). For example, MME gives the instruction directly to the S-GW and indirectly to the P-GW, when it is time to setup a bearer, the MME tells the S-GW to setup the bearer. The S-

GW will pass this indirect instruction on to the P-GW. These components can manage the data forward and backward flows from the mobile device to the IP flow network.

Serving GateWay (S-GW): It is the gateway which connects the interface between the EPC and E-UTARN. For each mobile device linked with the EPC, there is a single S-GW at a given point of time. S-GW focuses only on the user plane, it is responsible for forwarding the data packets from P-GW to eNodeB and to maintain the data session (e.g., the bearer and the mobile IP) in order to change and handover between the different eNodeBs locally. Therefore, it is sometimes called a local or mobility anchor. Moreover, when the mobile device moves from the current eNodeB to another one, the S-GW maintains the data session connectivity for the UE in the handover when switching between various eNodeBs [19]. For example, if the user works in a city (like Liverpool) he will be the Liverpool subscriber and he is connected to eNodeB LTE network close to his office. When he drives his car to go back home he will switch from one eNodeB to another; The S-GW will switch the connection of UE to the nearest eNodeB on his path to home. As a result, the S-GW is also located in Liverpool. S-GW maintains the data session from P-GW to eNodeB through the General Packet Radio Service (GPRS) Tunnelling Protocol (GTP).

Packets Data Network GateWay (P-GW): It is the gateway connecting to the EPC with external IP networks, such as Internet, IP Multimedia Subsystem (IMS), emails and special network services. P-GW is responsible for connecting the UE with IP networks by assigning an IP address (IPv4, IPv6) to the UE to connect to a specific network [20]. It works as an IP anchor to maintain the same IP address during mobility between 3GPP and non-3GPP services, it acts like a Home Agent (HA). In addition, the P-GW is responsible for enforcing Quality of Service (QoS) policy set by Policy

and Charging Rules Functions (PCRF) of QoS components in IMS. When a mobile device requests a bearer or when a bearer needs to setup an IMS call or video call, the P-GW and PCRF will interact together to make sure that the right policy has been enforced for that bearer.

Home Subscribe System (HSS): This component is a kind of database to store all the information related to the subscriber. HSS has two functions, the Home Location Register (HLR) and the Authentication Centre (AuC) that are already existing in the Global System for Mobile communications (GSM) and Universal Mobile Telecommunications System (UMTS) networks [21].

The HSS is responsible for storing and updating data related to user subscription such as:

- User addressing and identification numbers.
- User profile.
- Network authentication and authorization information such as path ciphering and integrity protection.

2.2.2. WiFi network

The wide deployment of WiFi encourages the telecom providers to pay attention to the roll of WiFi that can have an effect on helping their network to enhance user services. For instance, according to the Office of Communications of U.K (Ofcom) [22] there is around 81% of mobile consumers using WiFi network at some point, therefor network operators consider a WiFi network an important player as a method to offload mobile data traffic, reduce the cost for user services, expand the cellular coverage and mitigate interference of cellular networks in dense areas. According to

3GPP specifications [23], there are two methods of integrating non-3GPP networks (e.g., WiFi), namely un-trusted networks and trusted non-3GPP access networks, as illustrated in Figure 2.3. The following will explain the two methods with a dedicated 3GPP network (LTE).

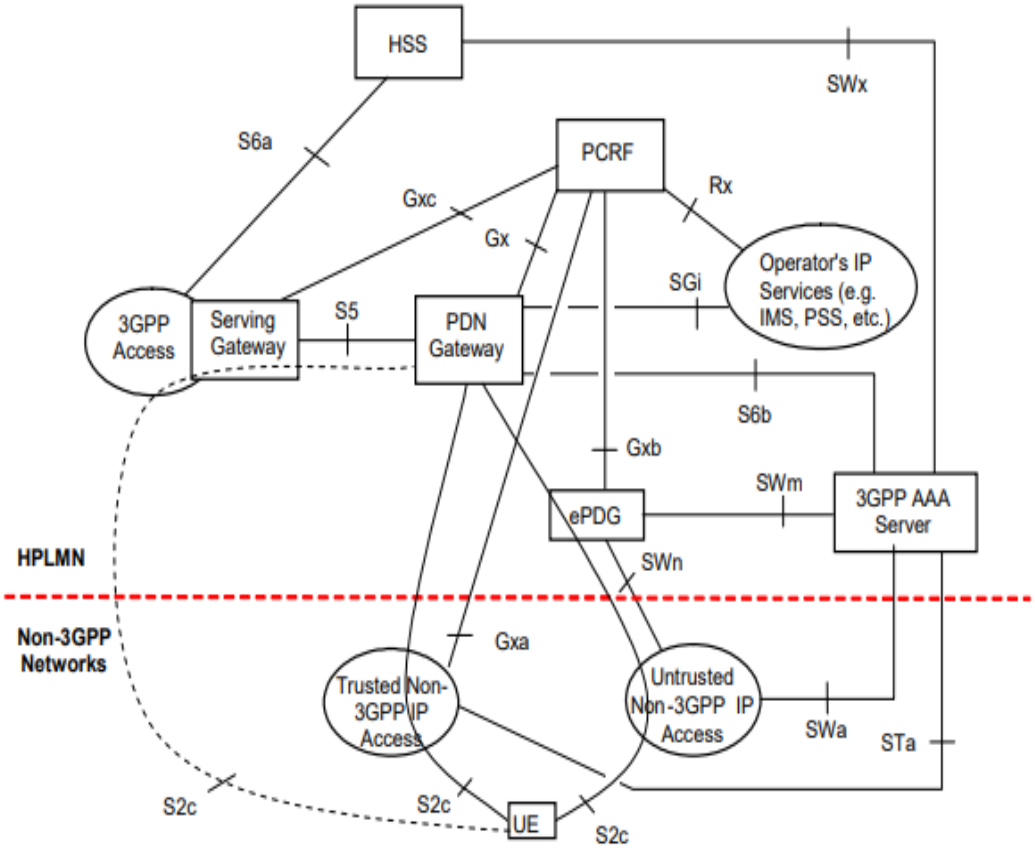


Figure 2.3: Un-trusted and trusted non-3GPP access networks [24].

LTE network architecture has been designed to collaborate with other networks (3GPP and non-3GPP networks) to support service with mobility. As mentioned in the 3GPP specification, the policy of the home operator decides on the type of non-3GPP access whether it can be trusted or un-trusted. When the operator decides that all the security matters are provided by the operator itself and satisfy all the security features, in this case it could be considered that the non-3GPP access is trusted. In contrast, if there is one or more security features not satisfying the access network, it will be identified as

un-trusted by the home operator. In such a case, the user device should use the IPsec tunnel protocol to establish a connection to enhance PGW (ePGW) node through invoking the Internet Key Exchange (IKEv2) with Extensible Authentication Protocol Method-Authentication and Key Agreement (EAP-AKA) for UE device authentication.

During the mobility the EPC relies on the PGW to work as an anchor node to keep the IP session continuous across LTE and WiFi networks. There are many solutions for transport the packets during this session. For example, all user packets connections are seamless meaning that all the packets keep their original IP during the handover; all user packets connections are non-seamless meaning that all the packets will get a new IP from WiFi network; a dedicated user data packet will be seamless; Individual IP flow service will be seamless during the handover.

2.2.3. Case Study (5G Network)

2.2.3.1. Services and Business Requirements

5G networks introduce innovative services with new requirements that are considered as a promising inspiration to the user experience. These services have different requirements including, enormous data traffic volumes and higher number of devices introducing new requirements to shape network platforms. The first trigger of 5G is anticipated by 2020 in order to meet the new business and customer requirements.

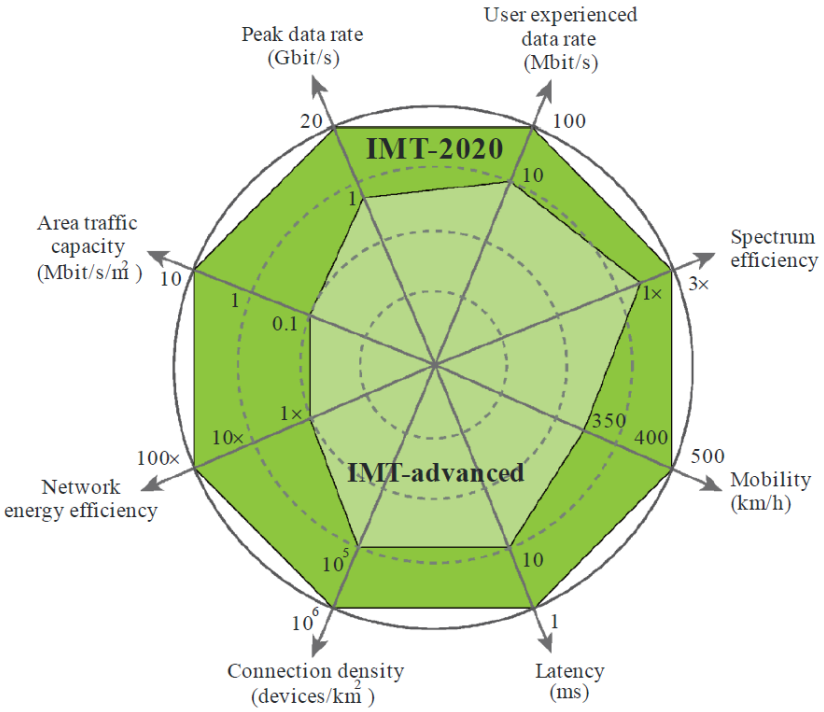


Figure 2.4: New services requirements in 5G system [25].

The International Mobile Telecommunication system–beyond 2020 (IMT-2020) identifies eight parameters as key capabilities of new services' requirements. Figure 2.4 shows the reference values of these parameters.

- Peak data rate per user/device is up to 20 Gbps under the ideal network conditions.
- 100 Mbps of user experienced data rate is achievable in a particular coverage area network to a mobile user/device.
- Low latency is expected to 1 ms over-the-air in some latency communication and high reliability scenarios.
- High speed Mobility with a defined QoS and seamless continuity service can be achieved up to 500 km/h.
- Connection density is expected to achieve the number of connected devices up to 10^6 per km^2.

- Energy efficiency can be achieved in respect of two aspects: (i) on the network side, where the energy efficiency is enhanced 100 times more than the current energy-networking; (ii) on the device side, where the energy lifetime for machines (e.g., sensors) is estimated to be greater than a decade.
- Spectrum efficiency refers to the average of data throughput of spectrum resources per unit, which can be expected to be three times higher than the spectrum of the 4G network.
- Area traffic capacity is expected to provide 10 Mbps/m^2 traffic capacities in the user density area (e.g., hot spots).

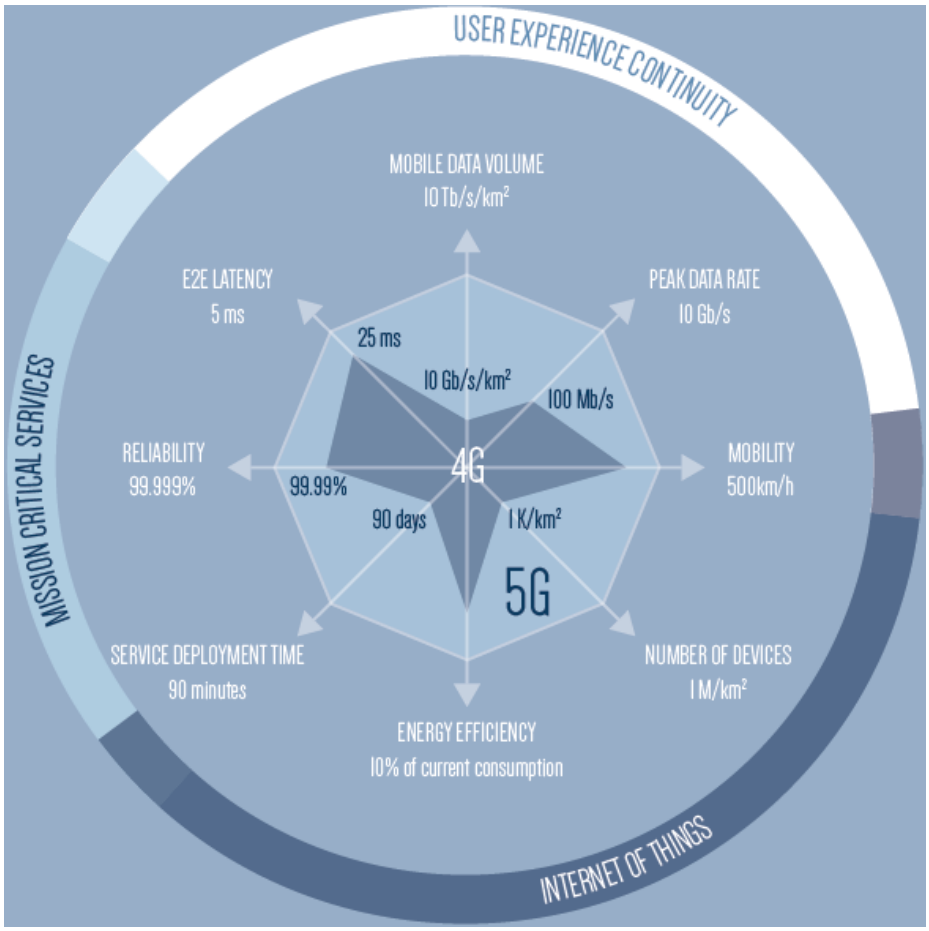


Figure 2.5: The key capabilities of new services' requirements in 5G [26].

Similarly, the 5G Infrastructure Public Private Partnership (5GPPP) provides some parameters as key enablers to identify the new services' requirements. Figure 2.5 shows a comparison between the capabilities of the 4G and 5G networks. Moreover, when looking at the figure, it represents the references values of these parameters, which approximately have the same above mentioned capabilities as in IMT-2020.

2.2.3.2. Business and Market-Trends

5G enables new innovative services and networking capabilities to facilitate a business ecosystem not only for consumers, but also for vertical industries. 5G is the technological answer, making it possible to adopt new business models and partnerships for enabling vertical markets and contributing to the fourth industrial revolution that affects different sectors [27]. Vertical markets can create new products and services, whereas, the network operators adopt new partnerships, for creating customised services to vertical industries. The different business principles are realized according to the virtualization and slicing of 5G architecture detailed as follows [28]:

Infrastructure Provider (InP): in general, it is responsible for managing the upgrades and maintenance of the physical devices at the network infrastructure layer. The network operators are in charge of running the InP in the current network. However, the partnership operator in the 5G network has the ability of take advantage of managing the networking and connectivity of network infrastructure in private and/or public networks (e.g., shopping centre or stadium).

Cloud Service Provider (CSP): it is a company (third party partner) that offers components which have the ability to provide computation, storage resources and cloud services. For example, Software as a Service (SaaS) such as Amazon, Platform

as a Service (PaaS) like Microsoft's Azure, Google's Kubernetes and Linux's OpenStack.

Virtual Network Operator (VNO): it is an operator that leases the resources and services from the InP. VNO is called a virtual provider because it provides the network services to the customer without owning the underlying network. Moreover, it is using the lease resources either to extend their network coverage in the areas that are facing leaks in the physical network services, or to increase the network capacity in the density regions (e.g., urban areas).

Service broker: it is the intermediate component that interacts with the physical network resources. It is mapping the resource requests of different service providers through abstracting information, such as CSP, VNO, application providers and verticals. After it collects the abstracting information, it allocates physical network resources to the mobile network operator's based on these abstraction information. This element can be a component of the infrastructure provider, mobile network operator or even independent third-party component.

Application providers: it is working on the top of a network operator. The network operator, can be either the same application operator or another network operator that owns the infrastructure services. 5G applications are characterized with high data rate consumption, which pushes the application operators to rethink about creating a new business partnership with network resource operators in order to satisfy the service requirements and enhance user QoE. The partnership model is identified as the application operators buying an independent network's resources to operate their services and offer free network services to their customers, the clients will pay just for the application services according to a pre-defined Service Level Agreement (SLA).

In such a manner, the application operator (e.g., Netflix) will insure the service requirements and satisfy the user experience.

Verticals: vertical industries (non-telecom industry) offer digital services, taking advantages of network and cloud resources from different providers. Most of these industries have a lack of knowledge regarding the complexity of the physical network, they care about delivering their services to the end user, such as healthcare and transportation.

Based on the aforementioned, the 5G network provides a different perception of the current network in business partnerships, where each stakeholder can establish its own network independently. 5G offers a unified platform to establish these networks under a complex virtual environment, as shown in Figure 2.6 that represents the 5G network architecture [29]. This environment empowers the operational capabilities of the 5G network, such as services as a programmable, software oriented capability set. The Network slicing concept is considered as an important key technology of 5G, where it enables 5G to perform different services belonging to various business scenarios on the same network infrastructure with efficient isolation and optimal resource utilisation. Additionally, network slicing can be creating network slices according to a business model, either on a permanent or on-demand basis, even in the some scenarios both.

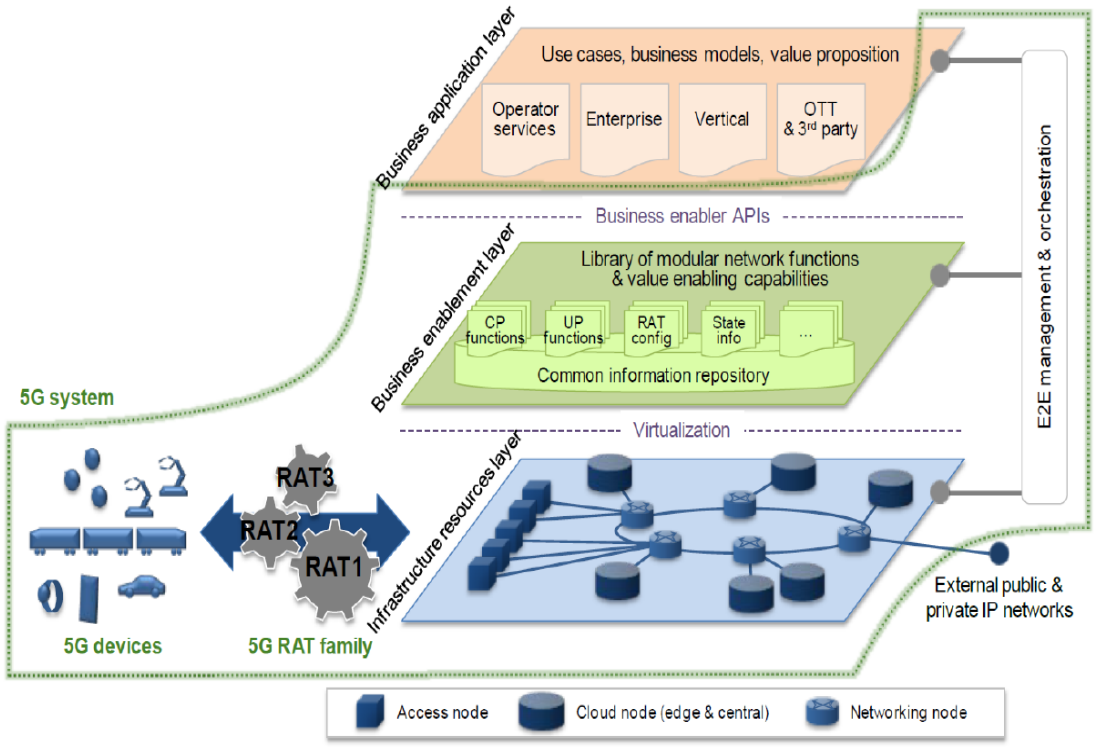


Figure 2.6: High level view of 5G network architecture [29].

2.3. Network Slicing Concept and Enabling Technologies

2.3.1. Network Slicing Concept

Network slicing as a concept was introduced for the first time by Next Generation Mobile Networks (NGMN) [30] within the context of 5G network. Network slicing enables multiple logical independent networks to operate on the top of a common network infrastructure. This logical concept provides an innovative integrated partnership between stakeholder ecosystems in both telecom and non-telecom industrial sectors, where the advantages of network slicing are utilized to establish independently a logical network for each stakeholder according to their service requirements. Such of these advantages are like, services as a programmable, multi-tenant network environment.

According to 3GPP standards [30], they define network slicing as an emerging technology that facilitates a physical network infrastructure to produce a logical platform, which allows the network operator to create many networks to satisfy different business service scenarios' requirements, in terms of functionality, performance and privacy (isolations).

Generally, network slicing consists of a chain of VNFs that represent the functional capability of a slice. VNFs differ from one slice to another according to the type of services and services' requirements. The most common network slicing requirements are explained below [31]:

Isolation: it is an essential feature of a network slice that confirms the performance guarantees and secures privacy (to enable the network to be economically open for multi-tenants). In other words, isolation means the ability to restrict the impact of one slice on other slices in the same network, even if they share the same infrastructure. That is to say, if there is any change of resource status in a slice (e.g. traffic load change), such a change should not influence the allocated resources of other slices. There are different degrees of network isolation such as infrastructure isolation, shared resources isolation (virtual resources) and isolation restricted to the policy guidance.

Customization: it confirms that resource management of each slice can be operated independently to meet the best individual service requirements. That is, the admission control policy of a slice can be different from the other slices.

Automation: this feature allows the third party (slice owner) to configure the network slice on an on-demand basis, which means it does not need to rely on a contractual SLA and manual involvement. This facilitates the slice owner to send requests

according to the SLA policy to update the slice reflecting the desired requirements, such as capacity, latency, jitter, etc., these change during the network slice lifecycle.

Elasticity: this is the fundamental feature in a network slice that enables resource allocation of the dedicated slice to change according to the SAL policy and fluctuating of services requirement belonging to the same slice; but with the restriction that these fluctuations do not affect the performance of other slices (e.g., relocating the VNFs of a slice, radio and network conditions)

Programmability: different APIs of dedicated network slice enable the slice owner to program the slice and controls all slice resources, which facilitates the capability of slice resources to accommodate on-demand services.

End-to-end: it is an essential property of network slicing to provide a logical network of service delivery from the service provider to the end-user that extends across different network domains, such as access network, core network transport network and network terminals. These domains link together to establish end-to-end network slicing and independently each domain has its own functions, resources and protocols. The network slicing facilitates all these domains and creates an end-to-end logical network that enables the third-parity to administer the network slice.

Hierarchical abstraction: network slicing provides different levels of virtual resources abstraction depending on the degree of the SLA and the granularity policy of resources provisioning. These levels are encapsulation, network slicing encapsulating one level of abstraction resources inside another which in turn maybe encapsulated with the other one, creating a hierarchy of abstraction resources in dedicated network slicing. For example, a virtual network operator leases a network

slice from an infrastructure provider, in turn, the virtual operator leases a part of the network slice resources to another service operator to run its services (e.g., enabling a utility provider to form its IoT slice).

2.3.2. Enabling technologies

2.3.2.1. Hypervisor and Container

The concept of virtualization refers to a method of representing actual things by creating a virtual construction which carries the same characteristics of the physical source, but with more flexibility. In the same context, the virtualization in the computer network (network virtualization) denotes a process of creating logical resources from the physical network resources (infrastructure hardware), such as router, switch, servers, physical link and terminal devices. Network virtualization needs an additional layer to map the link between the virtual resources and the physical resources and perform all network configurations. The Hypervisor is one of the enabler technologies of mapping virtual resources. It is a software package that is installed on top of computer hardware creating the virtualization layer and acting as a platform for creating different Virtual Machines (VMs), it manages the sharing of physical resources into virtual. There are two types of Hypervisor, the first type known as a Bare Metal Hypervisor and the second as a hosted Hypervisor. Bare Metal hypervisor, this type has its own operating system which is installed directly on the computer hardware so that it creates virtualization layer where the VMs work, such as Oracle OVM [32], XEN [33], and VMware ESX/ESXi [34]. The second type (hosted) hypervisor, in this type the hypervisor is a software application that is installed on the top of the hosting operating system, for instance Oracle Virtual Box [35], VMware Workstation player [36], and VMware fusion [37].

Container is a method of operating system virtualization that allows different applications running simultaneously in resource isolation processing [38]. The container works based on OS-level virtualization that partitions the operating system and creates multiple isolated virtual environments capable of running different VMs instance [39]. For example, different IT companies are using containers, such as Docker [40], Solaris Container [41], and Linux-Vserver [42].

However, the Hypervisor, VM and Container all are capable of running VNFs and chaining them together to perform the network slicing, where each VM represents a dedicated VNF.

2.3.2.2. Software-Defined Network (SDN)

SDN network has significant advantages by making the traditional network to be open and programmable [43]. This is due to the fact that the SDN comes up with a new concept of splitting up between the control plane and the data plane (user plane), which is unlike the traditional network where the control plane and user plane are residing on the same network device (node). In the SDN the control plane moves to a central device called an SDN controller that is responsible for handling all control messages to guide the user plane packets from the source to destination and vice versa. The SDN architecture consists of three planes with number entities distributed on these planes according to their function as shown in Figure 2.7. In general, there are three main views of SDN architecture but all of them have the same components: the first view is based on the OpenFlow protocol, the second via Application Programming Interfaces (API) and the last is according to overlay network. The Open Networking Foundation (ONF) is the organization in charge of the OpenFlow protocols where all the network infrastructure operates based on this protocol [44]. The CISCO organization

introduces the API modules, where the developers can run SDN network devices using the API interfaces for instance Command Line Interface (CLI) and Simple Network Management Protocol (SNMP), so the network can be programmed using extended APIs. Moreover, the VMware [45] is using SDN via overlays. For example, the Virtual Extensible LAN (VXLAN) tunnels are used across the network infrastructure.

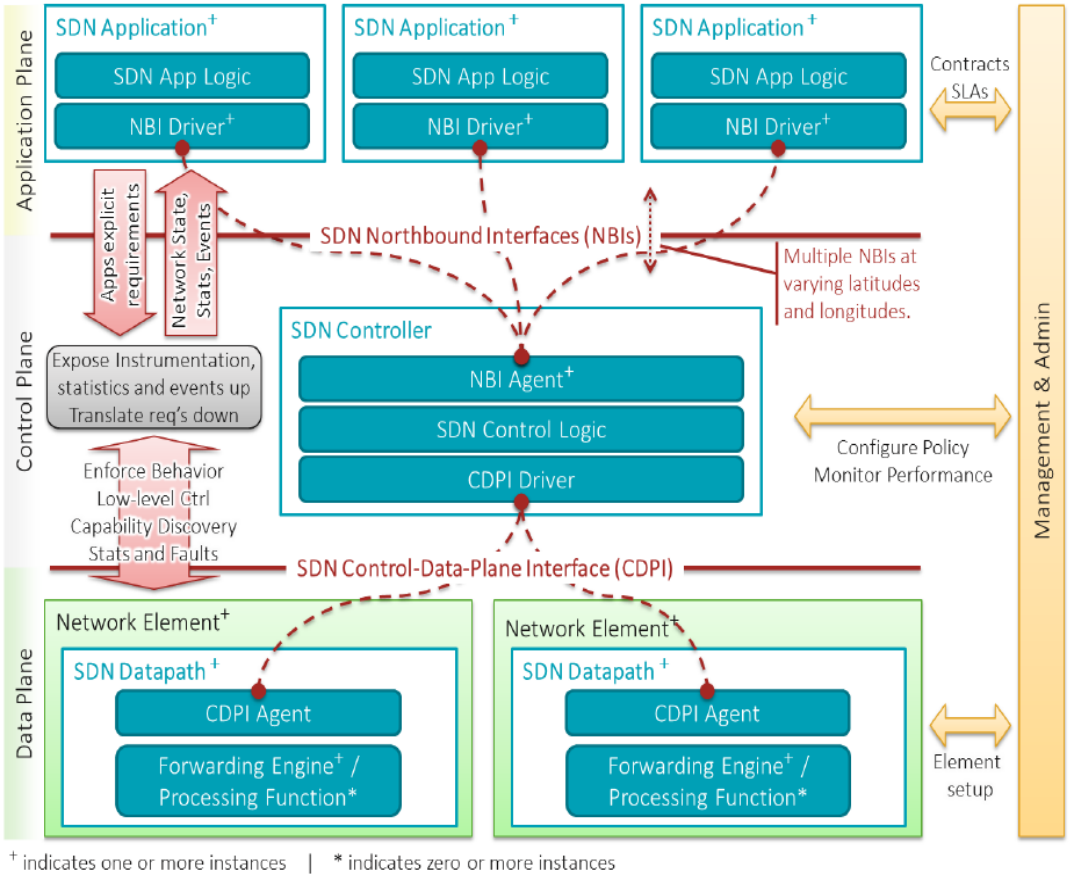


Figure 2.7: Overview of Software-Defined Networking Architecture [46]

The main essential elements in the figure of SDN architecture are explained as follows:

SDN Control-Data-Plane Interface (CDPI): CDPI is also called Southbound Interface (SBI), which is in charge of handling the control messaging between an SDN controller and an SDN data path. SBI provides a programmable platform to control all forwarding operations, statistics reporting and event notification.

Northbound Interfaces (NBIs): NBIs are interfaces between the SDN controller and the SDN application layer. These interfaces provide a command line instruction to abstract the network requirements and behaviours between those SDN entities. NBIs add more value to SDN in a way such that they are unrestricted to be implemented in any vendor platform (vendor-neutral and interoperable way).

Application Plane: The set of applications that hold the network behaviours and requirements, where they explicitly deliver these network demands to the SDN controller via the NBI.

An SDN application has one application logic and one or many NBI engine drivers, which ultimately translate the application instructions to NBI in order to send to the SDN controller.

SDN controller: It is a logical centralized entity that represents the network brain, which it is responsible for (i) delivery of the requirements of the application layer to the SDN data paths and (ii) providing the application layer with the global view about the network status by statistics and events. Additionally, it enables the network applications to program the SDN data paths via SBIs. However, the SDN controller is responsible for establishing flows in the network. There are two methods for establishing flows that the SDN controller used, namely proactive and reactive. Moreover, the two key performance metrics associated with SDN controller establishing flows are the flow time setup and the number of flows setup per second.

2.3.2.3. Network Function Virtualization (NFV)

NFV is a conceptual network architecture utilizing IT technologies to programmable dedicated hardware functions and virtually running them on commercial off-the-shelf

(COTS) servers and building blocks of Virtual Network Functions (VNFs) that chain together to establish communication services [47]. From the telecom operator's viewpoint, there are many components in the network responsible for maintenance (firewall, router, load balancer, media server, web server and switch), and they have difficulties in terms of inflexible, costly, high power, truck rolls. The number of truck rolls may be required, which is referred to a number of people may be taking a responsibility of installation and maintenance all these complex environments. So that, they need to reduce these complex environments and one way of doing this is by using a virtual environment with standard NFV. Most operators and enterprise users would prefer to use a virtual environment to perform a network with facilities such as less complex, very flexible, reduce power consumption, lower CapEx and OpEx, the ability to test new applications and tools with a lower risk, reduced time-to-market (TTM), and open market to many other software suppliers.

The high level of NFV framework architecture consists of four components (NFV infrastructure, Virtual Network Functions, OSS/BSS layer and management and network orchestration), as show in Figure 2.8 [48]. The following explains the four components in detail.

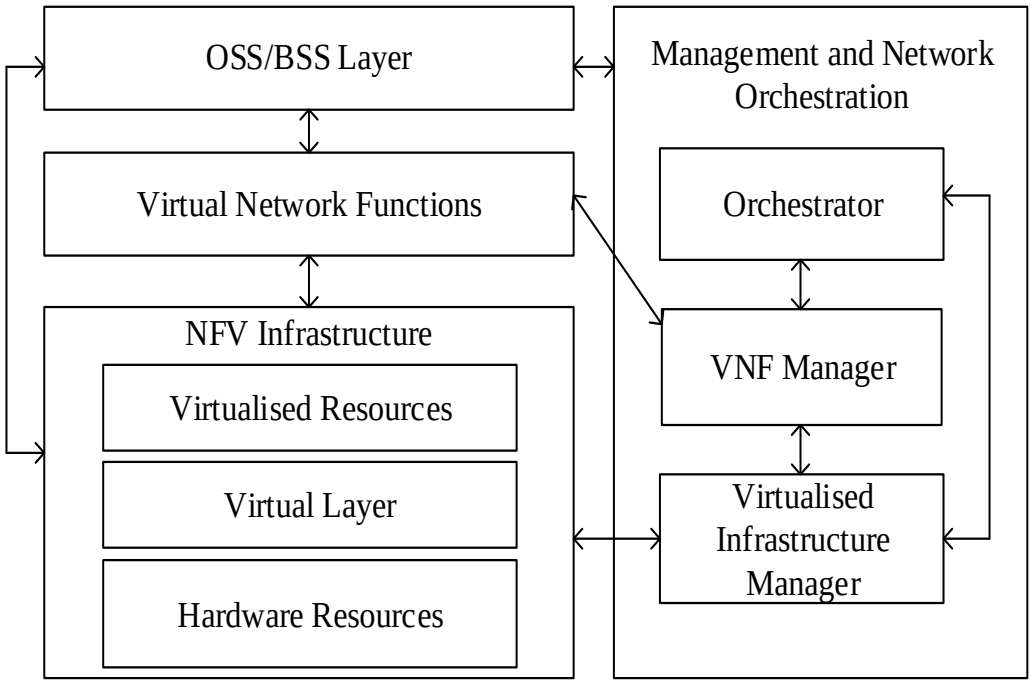


Figure 2.8: The high level of NFV framework

NFV Infrastructure (NFVI)

It is a kind of cloud data centre containing hardware and virtual resources that build on top of the NFV environment, these include (servers, switches, virtual machines, and virtual switches). The VNF contains three elements as follows.

Hardware resources: This element includes computing resources (such as server, RAM), storage resources (such as disk storage) and network resources (such as switches, firewalls, routers).

Virtualization layer: It abstracts the hardware resources and decouples the hardware from the software. This enables the software progress independently from the hardware. Many open source and repertory can be use to implemente the virtual layer such as KVM, QEMU, VMWare and OpenStack.

Virtualize resources: This element includes, virtual compute, virtual storage and virtual network.

Virtual Network Functions (VNFs)

The VNF layer is based on building blocks in the NFV architecture. These blocks are software implementation of network functions. VNF can be connected or combined together as building blocks to offer a full scale network communication service, this is known as service chaining. For example, (v-IMS, v-Firewall and v-router).

Management and Network Orchestration

This unit is known as MANO, which it has three parts as explained below.

Virtual infrastructure manager: It controls a number of managed components, such as the interaction of VNF with (NFVI compute, storage and network resources), it also has necessary deployment and monitoring tools for the virtualization layer.

VNF manager: It manages the lifecycle of VNF instances. It is responsible for initializing, updating, enquiring, scaling and terminating the VNF instances.

Orchestrator: It is managing the lifecycle of network services, which include instantiation, policy management, performance management and Key Performance Indicator (KPI) monitoring.

OSS/BSS layer

This is the last unit of NFV architecture, which manages the Operations Support System (OSS) and Business Support System (BSS) in the NFV framework. Each of these components deal with a number of network functionalities, where the OSS

responsible for network management, fault management, configuration management, service management and element management. Whereas, the BSS is responsible for customer management, operations management, order management, billing and revenue management.

2.4. Network Management and Orchestration Architecture

2.4.1. Network orchestration architecture

Network slicing is a very complex environment which includes different VNFs and Physical Network Functions (PNFs), multiple network domains and technologies, that makes it very complicated to control these elements together and simultaneously. Therefore, 5GPPP introduces a conceptual architecture of network slicing orchestration to facilitate the automated arrangement, coordination, and management of complex systems, middleware and services, as illustrated in Figure 2.9. This architecture consists of:

End-to-end service management: network slicing represents a fundamental means for service provisioning to accommodate various vertical sectors on the top of a unified network infrastructure view. Each network slice represents a dedicated service which satisfies the service requirements. The end-to-end service management unit receives service requests from different verticals and according to the requirements of each services, this unit runs different functions, (e.g., slice brokering, policy provisioning, considering the desired SLA, resource customization, service Mapping and considering the desired SLA), to establish a slice for each service. Moreover, this unit is also responsible for managing multi-domain slicing and creates a network service graph on the top of the abstraction virtual resource orchestrator.

Virtual resource orchestration: this unit is in charge of performing all MANO operations, such as the life-cycle management of VNF instances, as well as being responsible for mapping service requests of the instantiation of the virtual network service graph, either virtual and/or physical.

Network resource programmable controller: is responsible for facilitating the programmability resource separating of the control plane and data plane. It can enable the verticals and the third party to program the infrastructure's PNF and allows them to directly control VNFs and allocated slice resources. The programmable controller performs resource coordination (e.g., spectrum management) allowing different tenants to share different resources (wholly or individually).

Life-cycle Orchestration: it performs the principles of network management (e.g., administration, dependencies across service instances). Moreover, it monitors policy provisioning of all current services running according to their SLA contract.

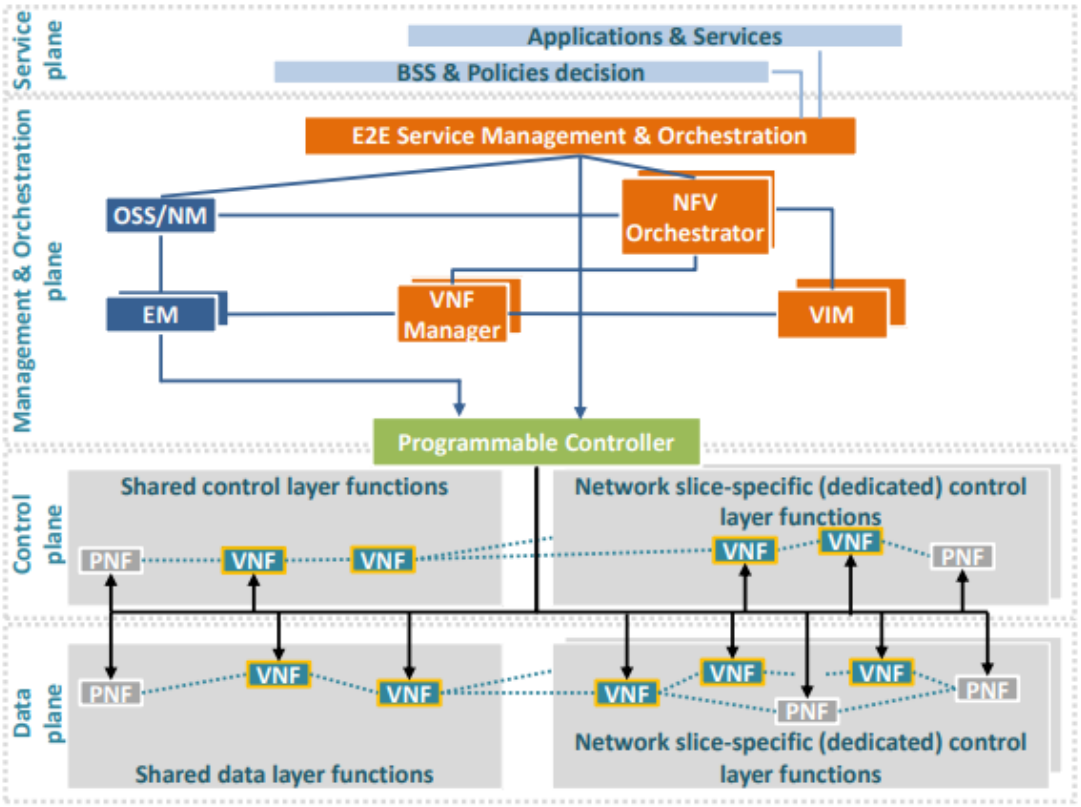


Figure 2.9: Network orchestration architecture [49].

As can be noticed, this architecture builds on the concepts of SDN and NFV that turn to provide flexible separation and programing for the control and data planes across all network segments. Additionally, many proposed solutions relay on the functionalities of this orchestration architecture. For example, the authors in [50] and [51] proposed a network slicing architecture describing the RAN and the distributed core network slicing to support the concepts of multi-tenancy and multi-services. Based on the LTE network, the author in [52] introduced a network slicing architecture, which extensively studied different technologies to demonstrate and enable the orchestration architecture of network slicing in the LTE network.

2.4.2. Network Slicing Life-Cycle Management

Network slicing life-cycle refers to the Network Slice Instance (NSI) that is responsible for managing the entity in the operator's network. Note that, according to the 3GPPP specification [53] the lifecycle of the service instances may not be necessary to be active during the run-time of NSI. The NSI life-cycle provides the following phases, as illustratd in Figure 2.10.

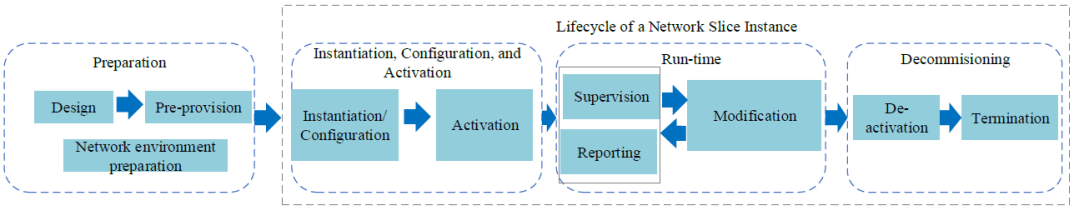


Figure 2.10: NSI Lifecycle management [54].

Preparation phase: this phase is not included within the lifecycle of NSI as shown the figure, which is responsible for preparing configurations of the network environment for supporting the lifecycle of NSIs, via the creation and confirmation of network slice template(s).

Instantiation, Configuration and Activation phase: it consists of two sub-phases, the Instantiation/Configuration sub-phase includes all resources that have been created or configured, either shared and/or dedicated, i.e., to a state where the NSI is ready for operation. Whereas, the activation sub-phase handles any actions putting the NSI in the active mode (e.g., diverting traffic to it, provisioning databases if it configured to the network slice).

Run-time phase: it is handling traffic to support different services communication. It also provides observation, reporting and activities regarding any upgrade and/or reconfiguration to the services. Reconfiguration activities may involve a number of

workflows related to run-time tasks, such as, instance change capacity, NSI scaling, change NSI topology, connection and disconnection of network function with NSI.

Decommissioning phase: this phase is responsible for deactivation of NSI and recovery of the associated resources with the NSI. After this termination process the NSI dos not exist anymore with the network function.

2.5. Related work

2.5.1. Network slicing in RAN and Core network

Network slicing is a structure of virtual network architecture that allows sharing a common physical infrastructure between different virtual networks. It enables a cellular system to share network physical resources residing in Core Network (CN) and Radio Access Network (RAN) among the virtual networks [55]. Figure 2.11 demonstrates a generic conceptual diagram of a network slice. Generally, cellular networks are composed of two different segments: RAN and CN. However, in the case of network slicing, we need an additional logical functional entity (i.e., Slice pairing function) which facilitates resource mapping between RAN and CN slices, as depicted in this figure. Each network slice is logically composed of one or more Network Functions (NFs) of CN and RAN. Note that, an NF can be occupied by a single slice or shared across multiple slices.

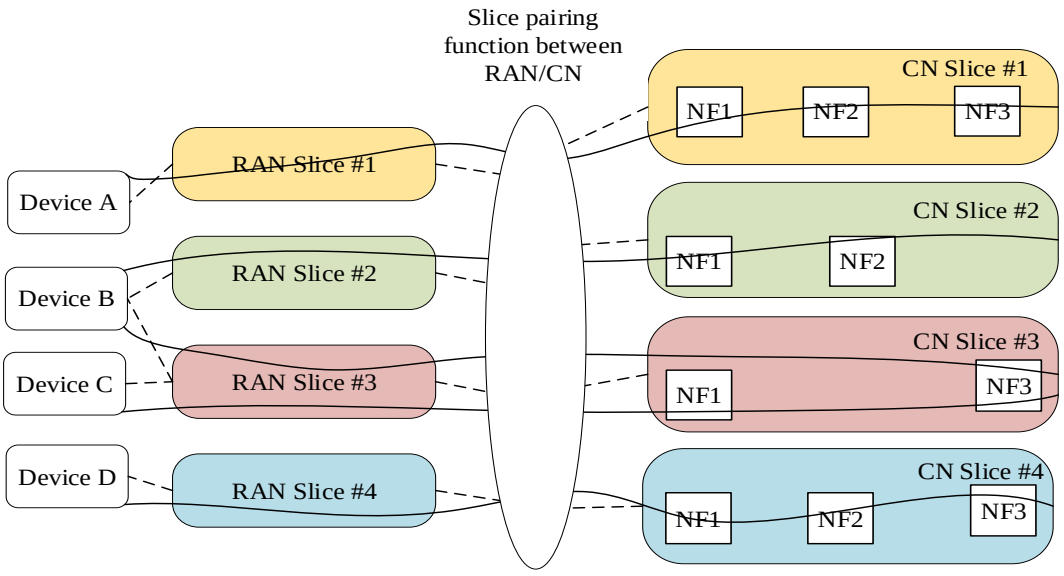


Figure 2.11: Conceptual diagram of network slicing.

Regarding the RAN/CN slicing, these domain need flexibility, customization and efficient resource sharing. These features of RAN contribute effectively in managing the scarce and limited frequency spectrum resources. Moreover, RAN/CN in network slicing are required to meet certain requirements, for instance, i) the resource management mechanisms should be dynamic, programmable and utilizing the open APIs in order to perform the different Key Performance Indicators (KPIs) for each slice; ii) end-to-end network slicing required a guarantee of resource isolation and sharing in order to enable logical self-contained network resources management; iii) each network slicing has different functional requirements in which the control plane and user plane are optionally selecting a dedicated VNF in order to ensure an optimal performance.

2.5.2. Resource Management

2.5.2.1. Virtual Resources Allocation in Cellular Networks

We have witnessed many research efforts on wired network virtualization; for example, wired network virtualization for distributed cloud data centre in order to maintain desired Service Level Agreement (SLA) [56], [57], [58]. The wired network virtualization is accomplished at different levels of a network such as processor, memory, ports connection and physical link layer. Unlike wired network virtualization, a wireless network requires virtualization in both the CN and RAN. Note that, the concept of wired network visualization could be applied on the CN. However, accomplishing virtualization in the RAN part is relatively challenging due to two important reasons: i) a radio link connection is affected by stochastic fluctuation of wireless channel quality, and ii) the wireless networking protocols are completely different from the wired network [59].

In cellular networks, a user may have many flows (user bearers) associated with different applications running on the user's mobile device. User bearers may share network resources with other bearers of different users through a virtual layer, which is mapped with physical network resources (infrastructure) [60], [61]. In [62], the authors propose a virtual cellular network architecture based on SDN. This architecture facilitates resource virtualization across the CN and RAN for all the packet flows in order to maximize network resources utilization. In their proposals, the authors apply the concept of Virtual Bearer (VB), which has been popularly used in wired networks. The concept of a VB is similar to the PRBs in the LTE architecture. However, there are two basic differences between them. First, they differ in time scale. In the case of a PRB, the length of a slot is fixed at 0.5 ms in LTE. On the other hand,

in a VB, the length of a slot may be negotiated between the service provider and the network operator depending on requirement(s). Second, in terms of ownership, a VB is owned by a service provider who lacks the knowledge about the wireless resources allocation (a service provider has concerns on meeting QoS requirements of the end users). Whereas, in the case of PRBs, they are owned by a physical Infrastructure Provider (InP).

In the next section, briefly, we present some of the existing research efforts in resource management in respect of network slicing.

2.5.2.2. Research efforts in resource slicing

A large and growing body of literature has investigated architectures for cellular networks slicing. M. Yang et al. in [10] propose a Karnaugh-Map algorithm in order to facilitate multiple user access in a virtualized embedded wireless network. This algorithm allows the network to handle real time resource requests. In this work, the authors did not provide an explanation how their proposed mechanism can be implemented in real hardware, such as in an LTE scheduler.

The authors in [63] extend the work proposed in [10] by considering a case of a dynamic embedded system that rearranges the requests that have already been rejected due to the static nature of the network topology. One major drawback of this mechanism is that its calculation of each scheduling time is too complicated.

The solution proposed in [8] aims to slice the resources of an LTE eNodeB into several virtual networks (slices) so as to allocate each of the slices to different Service Providers (SPs). Each SP has a number of users with different Service Level Agreements (SLAs). The scheduler in an eNodeB assigns a PRB to a user based on

the SLA between the user and the SP. For instance, the eNodeB scheduler guarantees that the minimum PRBs that should be allocated to a user. However, it is beyond the scope of this work to ensure isolation among the slices explicitly. This could result in not ensuring the SLAs of all the users. This in turn will result in degrading QoE of some users.

Another relevant work by R. Kokku et al. [9] introduce Network Virtualization Substrate (NVS). The architecture and algorithm of this proposal are designed considering a WiMAX network architecture. The proposal devises a slice scheduler (a slice pairing function), which allows simultaneous coexistence of two kinds of resource allocation mechanisms: resource-based and bandwidth-based reservation mechanisms. Yu et al. [9] highlight that, the flow isolation in WiMAX could be challenging. This is due to the fact that, according to the WiMAX standard, if a flow of a user requires more bandwidth than the initially allocated amount, the scheduler could allow the flow to occupy bandwidth of other flows belonging to the same user. Therefore, in order to ensure flow isolation, the authors propose to modify MAC of WiMAX in their solution. This solution introduced in [9] could be adopted to LTE with some modifications.

A heuristic-based admission control mechanism is proposed in [64]. The proposed idea mainly focuses on prioritization of the slices and users. A RAN scheduler takes into account a user's satisfaction while scheduling downlink transmission, resulting in improving overall QoE of users. Authors in [64] evaluate their solution based on a mathematical model.

The research efforts discussed above are promising. However, they all have one weakness or another. Unlike the existing proposals, the solution we introduce in this

paper is not computationally intensive (i.e., the solution does not require a long time to estimate resources required in each TTI). Additionally, in our solution, the user bandwidth request is met with regard to fair sharing of resources among users belonging to the same slice. It is worth highlighting that our proposed work is capable of optimizing resource allocation in case a slice needs an extra bandwidth in each TTI scheduling time. Finally, it must be noted that most of the existing solutions are evaluated based on mathematical analysis. Unlike the existing solutions, we use the OPNET Modeler in order to demonstrate the effectiveness of our solution in realistic scenarios.

2.5.3. Mobility management

2.5.3.1. Different protocols of Mobility management

Mobility management has a significant impact on a user's service continuity when it changes network attachment points during the handover. Different mobility management mechanisms are applied in different networking layers [65], Table 2.1 illustrates basic functions of mobility management in different networking layers [66]. In the physical layer, the mobility management is responsible for managing attach and detach functions of mobile node to different Access Points (APs) during the handover procedure [67]. In the network layer, the mobility administration works with the IP network where the mobile node changes its IP-subnetwork during the movement. For example, when the cellular IP network is used mobility based on routing protocols (Mobile IP [68] and Proxy Mobile IP [69]). In the transport layer, the mobility support in this layer deals with session continuity of TCP/UDP connections (e.g., for example used in Mobile Stream Control Transmission Protocol (M-SCTP) [70]. In the application layer, the mobility support mechanisms are based on the application

specification where the mobility management between two APs is managed according to specific application type (e.g., The Session Initiation Protocol (SIP)) [71]. However, the mobility management at the network layer is considered as a popular one that can support mobility management to all types of applications. The Internet Engineering Task Force (IETF) organization classifies the IP based mobility into two types, namely host based mobility management protocols and network-based mobility management protocols. For examples, most standard mobility protocols working at this layer are, (Hierarchical Mobile IPv6 (HMIPv6) for host based [72]) and (PMIPv6 for network based [73]).

Table 2.1: Mobility management in different networking layers.

Protocol layer	Basic functions in mobility management
Physical layer	• Provides mobility management-related physical signal detection and measurement, which can be used for function and performance optimization
Data link layer	• Provides terminal mobility within an IP subnet • Provides necessary information about link status and L2 (Layer 2) handover starting/finishing event notification, which can be used for function and performance optimization
Network layer	• Provides mobility independent of the lower-layer protocols and physical transmission media, and transparent to the upper layers • Mainly supports terminal mobility and network mobility • Provides L3 (Layer 3) handover starting/finishing event notification to the upper layers for handover performance optimization
Transport layer	• Provides end-to-end mobility support • Support reachability (IP tracking)
Application layer	• Provides various types of mobility support, especially for high-level mobility (personal mobility and service mobility)

HMIPv6 is host-based mobility protocol, when a Mobile Node (MN) moves between different APs as a first step, it registers to Mobility Anchor Point (MAP) during mobility management that it works as a Home Agent (HA) to the MN. Then a bi-directional IP-tunnelling is established between the MAP and the MN for exchanging packets. However, if the MN changes its location and leaves the current MAP domain, it will assign to a new MAP in the form of a hierarchical-tree. In this manner, the HMIPv6 reduces the handover overhead and optimal utilizing network resource.

PMIPv6 is network-based mobility, in such an approach the MN does not needs to signal direct to the local mobility anchor, where all this signalling is done by the network. The MN registration is done in a Local Mobility Anchor (LMA) by the Mobility Access Gateway (MAG) that is managed by the network. When the MN moves and changes the current network, it is detected by a new MAG that belongs to a new network. The new MAG in turn will send signalling messages to the LMA for updating the location of the MN and the LMA establishes bi-directional tunnelling to the MAG. Notice, the tunnelling between the LMA and MAG (the MN is not involved). This protocol provides better network performance in handover delay and less signalling cost.

2.5.3.2. Different Researches in Mobility Management

5G network has attracted many research interests in academia especially when the concept of network slicing is considered. The authors in [74] introduced mobility management mechanism considering low latency services in network slice. The authors optimised the selection of the mobility anchor during the attachment procedure between the edge nodes. However, this mechanism is capable of providing an efficient solution to enhance mobility management by considering handover latency as a key

concept, but it did not take into consideration the case of heterogeneous access network, which is considered one of the challenges of emerging wireless networks. In [75], the authors proposed a unified approach to mobility and routing management that offers Connectivity Management as a Service (CMaaS). This approach was designed based on SDN network architecture with hierarchical network control capabilities to allow different levels of network performance. The CMaaS enabled the service providers to work on the top of application services to manage their customers at different level of prices. In [76], the handover mechanism of the LTE network was redesigned to trigger a decision scheme based on the grey system theory. This handover mechanism can be applied to the railway communication system to provide less co-channel interference for passengers in carriages.

The authors in [77] proposed a hybrid computation offloading scheme for managing the increasing of traffic demands in 5G dense area networks. The proposed scheme takes into consideration the impact of the user mobility and the network caching in distributed Small cell Base Stations (SBS). The effectiveness of this solution, is that it provides efficient energy cost and better Quality of Experience (QoE). The solution in [78] is based on the Individual Mobility Model (IMM) instead of the traditional Random Waypoint (RWP), in order to evaluate the results performance of user mobility. The main consideration of this solution is to investigate the impact of human tendency and clustering behaviours on the performance of user mobility in 5G small cell networks. A novel resource-based mobility management mechanism is proposed in [79] for video users in 5G networks. This mechanism proposed N-step algorithm for selecting optimal routes between serving nodes and utilized the Homogeneous Discrete Markov model for user mobility patterns. The handover approach is energy-

based and the results show a reduction in the handover latency compared with the existing solutions.

Most of the existing research efforts are promising. They have the ability to provide efficient control and mobility management solutions, but these approaches are not focusing on mobility management in the concept of network slicing for heterogeneous wireless access networks. To the best of our knowledge, our work is a unique effort that addresses the mobility management in heterogeneous access networks. Additionally, section 1.3 describes our contributions to support mobility solution in network slicing, whereas we consider 5G networks as a network environment.

2.6. Chapter Summary

It is of great interest to both academia and industry investigating resource management and mobility management of emerging wireless networks (e.g., 5G networks) to satisfy the Quality of Service (QoS) requirements and enhance user experience with minimal resource cost, as well as opens a new market to introduce innovative services based on user demands. Network slicing has been receiving extensive attention from the researchers as an enabling technology of network virtualization which has paved the way for sharing resources, guarantee resource Isolation, SLA-aware, self-adaptive deployment network functions and enhancing different network services such as mobility management.

In this chapter, we summarised the efforts made on this front by presenting different research works regarding the resource allocation and mobility management in heterogeneous wireless networks. Our narrative starts with a background on heterogeneous wireless networks, where the architecture of the LTE, WiFi and 5G are

discussed in details. We then explained the concept of network slicing highlighting the roles of SDN and NFV, and identified the enabling technologies of network slicing. Also, we discussed the existing works in details for both resource allocation and mobility management, and showed the strengths and weaknesses of different methods. In the following four chapters, we present our research contributions in this area.

Chapter 3: Resource Management of Network Slicing

3.1. Introduction

This chapter explains the Resource Management of Network Slicing (NSRM) solution base LTE network. NSRM presents three main contributions: (i) a novel architecture framework for virtualizing (network slicing) the LTE network in order to maximize network resources utilization; (ii) a novel algorithm which is capable of dynamically distributing bandwidth among different slices within an eNodeB to maximize resources utilization; and (iii) a Max-Min model that ensures isolation of slice resources across flows and secures a fair share of minimum bandwidth among users. The prime objectives of NSRM are twofold: (i) satisfying the requirements of slices in order to meet the users' QoE, which in turn will lead to maximizing the revenue of both InP and a slice owner (e.g., SP); and (ii) meeting QoS requirements for all the flows belonging to the same slice.

3.2. Medium Access Control (MAC) in LTE network

This sub-section describes the two types of LTE frame structure, namely Frequency Division Duplex (FDD) and Time Division Duplex (TDD) . Then, we introduce some of the existing research efforts in virtualization of network resources in cellular networks.

3.2.1. LTE Frame Structure

MAC is a layer 2 protocol-stack of an LTE air interface, which processes the uplink and downlink flows [80]. LTE applies Orthogonal Frequency Division Multiple

Access (OFDMA) and Single Carrier-Frequency Division Multiple Access (SC-FDMA) for downlink and uplink communications, respectively. OFDMA divides the available spectrum into sub-carriers and allocates these sub-carriers to each user in the coverage area. The task of assigning resources to each user is referred to as scheduling. This scheduling works when each Transmission Time Interval (TTI) triggers an assignment decision made for each user on how many resource blocks should be allocated. Mainly, there are two types of frame transmission mode in a TTI: the Frequency Division Duplex (FDD) and Time Division Duplex (TDD) [81].

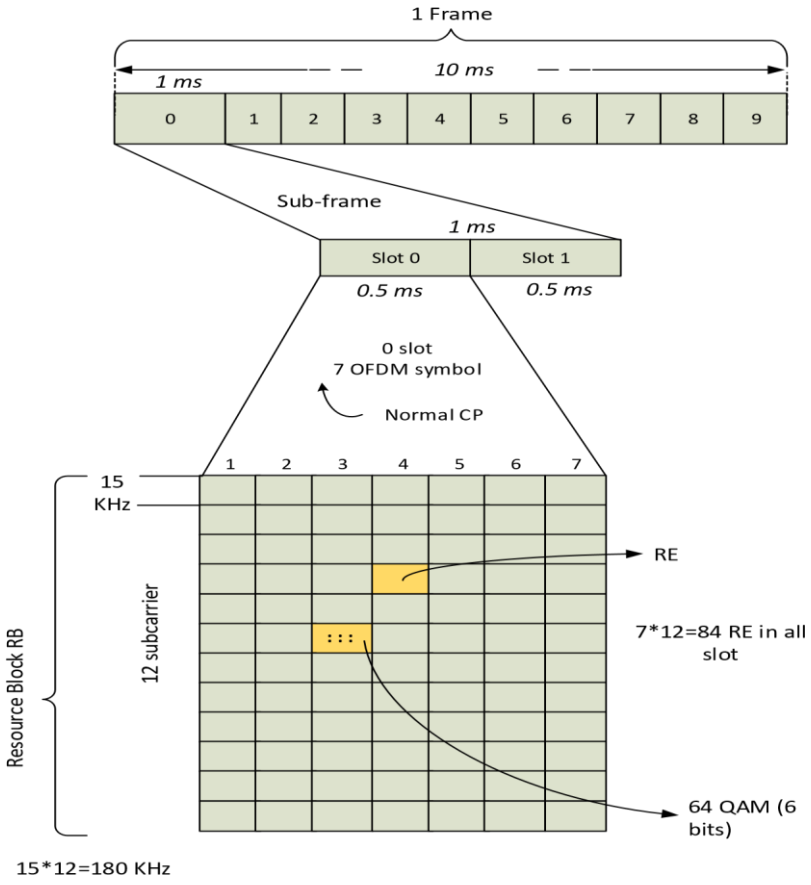


Figure 3.1: LTE resources allocation frame.

Generally, an FDD frame has a total length of 10 ms, whereas in TDD each frame of 10 ms in length is structured into two half-frames, each of them is 5 ms in length. An FDD frame is divided into 10 sub-frames in length of 1 ms for each one. Each 1 ms

sub-frame is divided into two slots where each has length of 0.5 ms. Each individual slot carries 7 OFDMA symbols which is defined as a Resource Element (RE). Moreover, there are 12 sub-carriers in each slot and one sub-carrier is the equivalent of 15 KHz [15]. Let us consider a specific example to explain the resource scheduling of a base station with transmission bandwidth of 20 MHz, as illustrated in Figure 3.1. In LTE, 12 sub-carriers in each slot form a PRB, which is the smallest unit that could be assigned to a user. Each sub-carrier has 15 KHz bandwidth; therefore, a PRB is equivalent to 180 KHz (12_15 KHz). We can quantify the total PRBs (B) of a base station as follows [14]:

$$B = (1000 \times T)/180\ KHz \tag{3.1}$$

where T is the frequency band of a base station in MHz units. The figure portray how total bandwidth of a base station is calculated. If the base station transmits at 20 MHz, first to obtain the number of PRBs, we divide 20 MHz by 180 KHz, which is 100 PRBs. Now, since each PRB has 12 sub-carriers, if we multiply that by 7 OFDMA, we will obtain 84 REs per slot. Thus, each sub-frame has 168 REs (each sub-frame contains two slots). In addition, in case the LTE base station uses a 64 Quadrature Amplitude Modulation (QAM), each symbol length is 6 bits (i.e., each RE is the equivalent of 6 bits). Finally using (1), the total quantified bandwidth is 100,800 bps (i.e., 100.8 Mbps). In reality, according to the LTE specification around 25% of the total bandwidth is consumed due to overhead (signalling associated control messages) [82].

3.2.2. LTE Traffic Scheduling

The LTE standard classifies network services into nine classes, such that four of them are handled as Guarantee Bit Rate (GBR) services, whereas the other five classes are handled as None Guarantee Bit Rate (NGBR) services [83]. The LTE scheduler uses these classes to prioritize flow services. An operator sets a scheduling scheme for its eNodeBs. A scheduling scheme should take into consideration different QoS with the LTE service class attributes and it has a very strict priority of flow services. Therefore, due to this priority, it would result in either starving of NGBR (best effort) class or in some cases the GBR themselves would face lack of resources because of less suitable channel conditions [84].

3.3. NSRM System Architectural Model

Figure 3.2 illustrates the network slicing concept along with the physical entities in RAN and CN of an LTE network (the reader is referred to [85] for more details about these components). The physical entities shown in this figure take part in forming all the logical entities of the network slices.

At this point, we need to highlight that in our solution, a slice owner is responsible for scheduling slice resources. It allocates the required resources for each user's flows according to a predefined SLA. The following subsection presents the NSRM system model.

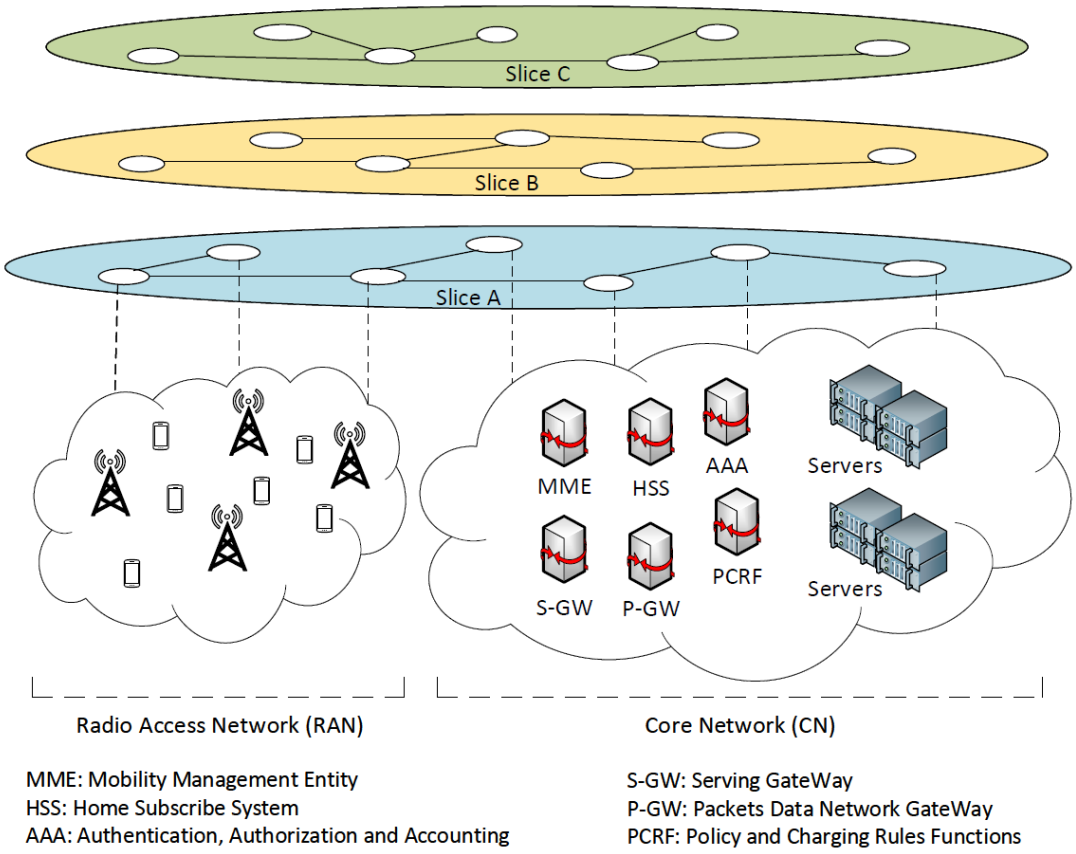


Figure 3.2: LTE physical resources with network slices.

In our NSRM architecture, the network slicing is actualized based on SDN and NFV. The conceptual architecture of the NSRM for the network slicing based LTE network is depicted in Figure 3.3. This architecture is broadly segmented into three layers: Slice layer, LTE Slice Controller Manager (LSCM) layer, and Slicer layer. Moreover, the architecture facilitates slicing a virtual network into a number of slices each of which is configured based on the service requirement of an operator.

To present our system model, we consider that in an LTE network there are three slices (slice A, slice B and slice C), as shown in Figure 4. We consider that each slice belongs to an operator and it is managed by its controller (Slice pairing function). The controller is in charge of maximizing utilization of the slice resources (all the virtual resources).

Generally, a user may have one or more flows. These flows might belong to the same slice or different slices [86]. In the case when the flows belong to the same slice, in our proposal, the controller needs to manage intra slice resources in order to allocate required resources to each flow. Besides, it should ensure the isolation between the flows in a slice. To make sure that each of the slices can have predefined allocation, we need to have inter slices isolation. In our proposal the Slicer layer is responsible for inter slices isolation (we provide more details in the subsequent part of this section).

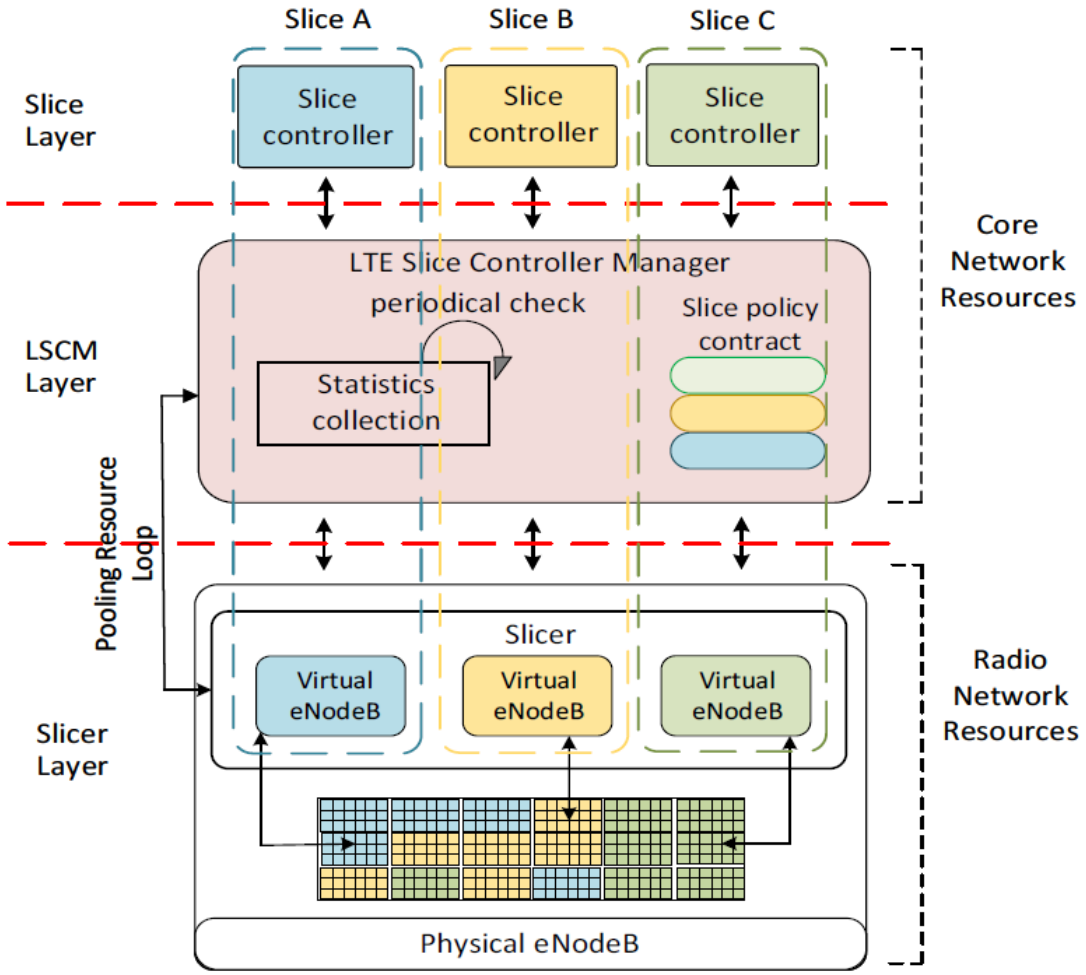


Figure 3.3: Conceptual LTE network slicing architecture.

As mentioned in [9], slice resource isolation can be classified into three general categories depending on: (i) group of users with the same type of application, (ii) end-to-end networking (different end-to-end flows), and (iii) resources allocated across different slices (the amount of allocated resource is predefined according to a policy). In our work, we consider that type (i) and (ii) fall under intra slice isolation. Whereas, the type (iii) requires inter slice isolation.

We assume that a policy administrator (see Figure 3.4) negotiates with an SP and settles the contract. Besides, it configures the LSCM layer in order to meet the slice requirements defined in the contract.

The elements of Slice layer, LSCM layer and Slicer layer are presented in Figure 3.4. In this figure, these elements are logically interconnected to illustrate the main functionalities of the proposed logical framework architecture. Next, we provide a detailed explanation on how these elements are worked under each layer.

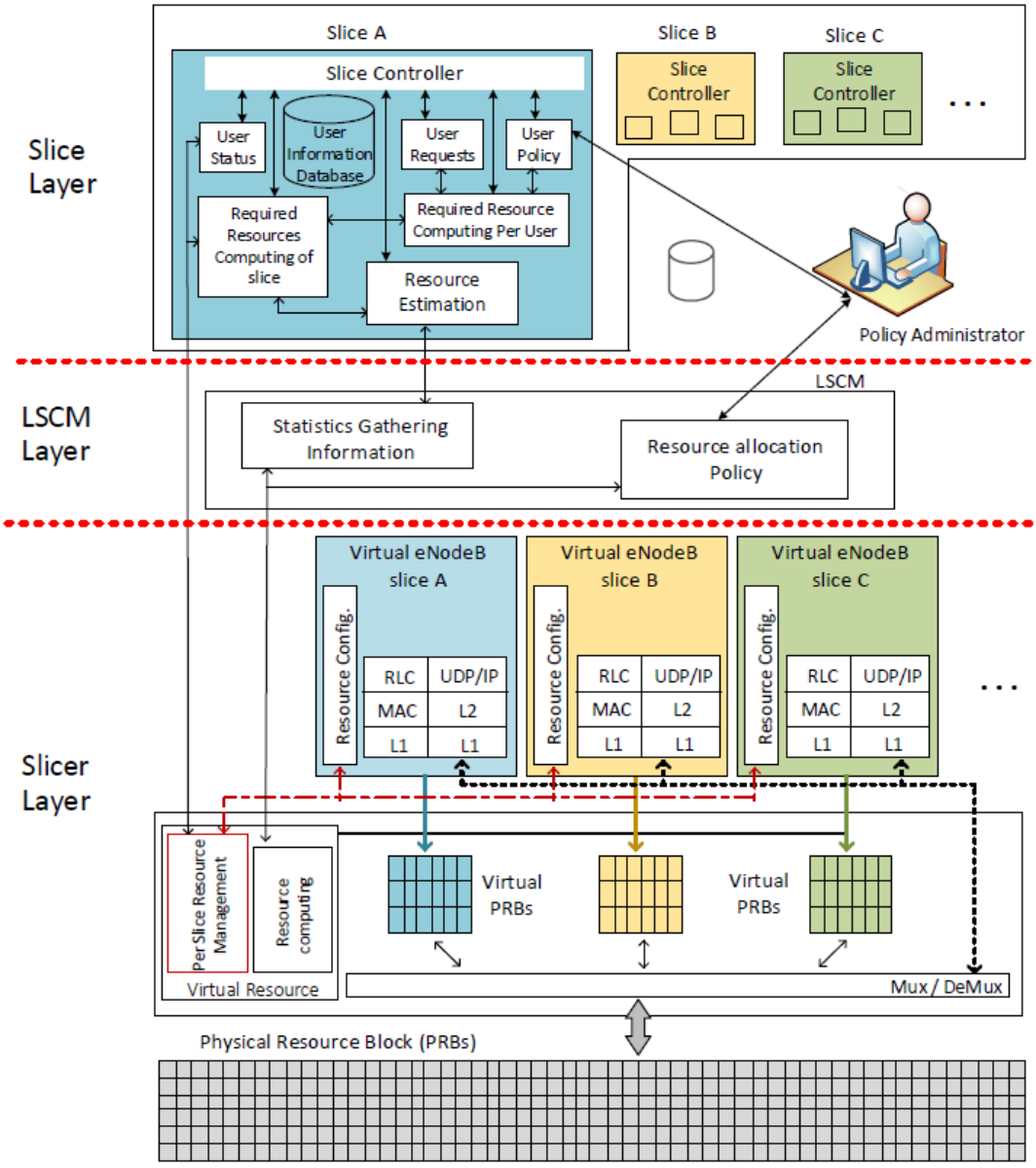


Figure 3.4: Logical interconnection of three-layer elements.

3.3.1. Slice Layer

As we mentioned earlier, each slice in this layer is owned by a slice owner and a slice controller is in charge of managing resources of a slice, as we can notice from Figure 3.4. The controller coordinates the interaction between slice elements and stores all slice information, such as users' information and resource requirements, in the User

Information Database (UID), as depicted in the figure. The following are the main elements of the slice layer.

- User Requests (UR): this element holds user requests. When a user wants to have a service from a slice, first, it needs to invoke the associated UR element of the slice. At this initial stage, the user sends information about type of service and amount of required resources to meet the requirement of the service. Upon receiving this information, the UR stores the information in the UID. The slice controller retrieves a user requirement from the UID whenever required.

- User Policy (UP): this element handles a policy for each user (i.e., each user is associated with a policy). The policy is defined by the policy administrator. The slice controller uses the policy defined for a user while processing any requests from the user.

- Resource Computing Per User (RCPU): RCPU computes the resource requirement in order to satisfy the request of a user. The slice controller of a slice uses RCPU to know the exact number of slice resources required to meet a user's request. The RCPU retrieves a user's information from the UR and UP before computing the resource requirement for the user.

- User Status: A user could be in an active or idle mode at a given time [87]. This element periodically tracks the status of a user (i.e., active or idle). This facilitates the controller to release the allocated resources of a user if the user is found idle at a given time. This approach will maximize utilization of the slice resources.

- Slice Resource Tracker (SRT): This element has the global view of the slice resources. It periodically observes overall resource utilization of a slice and notifies the slice controller.

- Resource Estimation (RE): This element is responsible for estimating the future expected amount of resources that would be required based on users' demand within the slice.

3.3.2. LSCM Layer

In our architecture, the LSCM layer manages the LTE core network (it facilitates communication among the CN entities). Additionally, the LSCM has a global view of network resources requirements. It dynamically monitors network resources' status through statistics of required resources and policies of assigning these resources. The following are the main two elements of this layer.

- Statistics Gathering Information (SGI): the task of SGI is to obtain statistics of the resource required for each slice. Periodically, the SGI collects and stores an estimated resource for each slice through the RE element. Therefore, it has historical statistics of resources for each slice. Based on these statistics, the mean value of required resource is measured in order to realize the exact resource requirement of a slice.

- Resource Allocation Policy (RAP): RAP element holds all the policies between the SP and InP. The policy administrator places these polices in RAP. This will allow the Slicer to get policy associated information before allocating resources to each slice (see Figure 3.4).

Mainly, there are two different categories of slice allocation depending on the type of contract (SLA): Guarantee bandwidth and Best effort [88], [89], [90]. We explain them briefly below:

Guarantee bandwidth is categorized into two subcategories, as explained below:

- Fixed Guarantee (FG): in this type of contract, the SP will request the Slicer to allocate a fixed amount of bandwidth all the time (this bandwidth may or may not be 100% utilized).

- Dynamic Guarantee (DG): In this case, the bandwidth allocated to an SP is dynamically changed. The Slicer guarantees bandwidth allocation with the change of an SP's bandwidth requirement. The SP will pay the InP depending on the usages.

Similarly, best effort bandwidth is classified into two subcategories, as presented below:

- Best effort (BE) with no guarantee: This type of bandwidth request has less priority than DG and FG. That is, in the absence of high priority bandwidth requests (ie. DG, FG), BE bandwidth request is accepted if the network has available bandwidth.

- BE with Minimum Guarantee (BEMG): In this type of contract, an SP can mention the lower and upper limit of its bandwidth requirement. The Slicer would ensure the lower limit of bandwidth request and the upper limit of a request will be satisfied in the presence of abundant bandwidth.

3.3.3. Slicer Layer

As shown in Figure 3.4, we introduce a virtual layer (called Slicer layer) on the top of an eNodeB physical resources. The Slicer concept introduced here is similar to the FlowVisor concept, which is designed for wired network virtualization [91].

The Slicer is responsible for virtualizing the eNodeB into a number of virtual eNodeBs where each of these eNodeB represents a network slice. It schedules eNodeB physical resources among slices instances. That is, the Slicer allocates bandwidth resources (PRBs) for each slice using a bandwidth allocation algorithm after taking into account predefined contracts between SP (slice owner) and InP. Note that it is challenging for the Slicer to allocate PRBs to the slices in a fair manner. To obviate this, in this thesis, we come up with an algorithm, which is referred to as a simple exponential smoothing model, to measure the number of PRBs required for each slice (Section 3.4.1.3 presents this model in detailed). The following are the main elements of the Slicer layer:

- Virtual Resources (VRs): the task of VRs is to create a logical platform and divide this platform into different logical instances, where each logical instance represents a slice. Moreover, VRs have two components running the functionality of this platform (see Figure 3.4):
    - Per Slice Resource Management (PSRM): PSRM controls a configuration of slice resources between users of a slice. Additionally, PSRM with the slice controller are enabling the distribution of slice resources among the users of the slice in a fair manner utilizing the concept of Max-Min model.

- Resource Computing (RC): RC is responsible for computing the estimated resource of each slice. RC utilizes the exponential smoothing model to calculate required physical resources in PRBs for each slice in every Round Trip Time (RTT). Moreover, SGI and RAP of LSCM layer are providing the RC with required statistics and policy rules to complement a process of slices resource allocation.

- Multiplexing/DeMuliplexing (Mux/DeMux): it is responsible for managing multiple data streams coming from/to different slices over eNodeB channel. Moreover, the Slicer uses this element in order to facilitate mapping between virtual and physical resources (see Slicer layer in Fig 3.4).

3.4. NSRM Solution

In this subsection, we present our NSRM solution. Before we delineate the proposed solution, we present mathematical models which assist the algorithms introduced in NSRM for making a decision in network resources allocation. We devise two mathematical models: the exponential smoothing model and the Max-Min model. The first model has the objective to quantify resource allocation among slices. The second model is formulated with the objective of fair resource allocation among the users in a slice.

Next, the NSRM presents two algorithms: (i) Resource estimation algorithm, which uses the estimation model we derive in this section and (ii) Fair resource sharing algorithm that uses the Max-Min model.

3.4.1. Mathematical Models for Estimating Resource Allocation of Network Slices

The resource allocation for slices using exponential smoothing model is presented. In addition, we provide a solution based on user's fairness and isolation using Max-Min model.

3.4.1.1. LTE Network Virtualization

In LTE, the RAN consists of a number of Base stations (BSs). Let $X = \{x_1, x_2, .., x_n\}$ denote as a set of BSs. For each x there is a set of slices $V = \{v_1, v_2, .., v_n\}$ with a set of users $U = \{u_1, u_2, .., u_n\}$ for each v. In BS, the spectrum bandwidth allocated to x is B_x (as described in Section 3.2.1). By using Shannon bound, we can define the spectrum bandwidth efficiently for user $u_i$associated with BS x as shown in the equation (3.2) [92].

$$\eta_{u_i\, x} = \log_2\left(1 + \frac{S}{N}\right) \tag{3.2}$$

where $S\ and\ N$ represent the average signal and noise power, respectively.

Let $L_{(u_i\, x)}$ be a pointer that indicates the user u_i is associated with BS x or not, where if the $L_{(u_i\, x)} = 1$ means u_i is connected to BS x; Otherwise $L_{(u_i\, x)} = 0$ means it is not. $Y_{(u_i\, x)}$ represents the percentage of radio resources allocated to user u_i by BS x, where $Y_{(u_i\, x)} \in [0,1]$ and notes that:

$$\sum\nolimits_{x_i \in X,\ v_i \in V,\ u_i \in U} Y_{u_i\, x} \leq 1 \tag{3.3}$$

So that, the instantaneous user u_i data rate is defined by:

$$R_{u_i\,x} = \sum_{x \in X} L_{u_i\,x}\, B_X\, Y_{u_i\,x}\, \eta_{u_i\,x} \tag{3.4}$$

3.4.1.2. Resources Slicing

Usually the PRB is assigned to a bearer as a pair of sub-frames in the time domains (as described in section 3.2.1). Thus, we consider one VB (Virtual Bearer) to be equal to a pair of PRBs sub–frames representing the resource of a slice in Slicer. Let δ_{u_i} represent the total number of VBs that the slicer actually assigns to a user bearer u_i over some observation period ΔT. Therefore, the total user bearer data rate ρ_{u_i} over this period is given as illustrated in equation (3.5):

$$\rho_{u_i} = \frac{\delta_{u_i}}{\Delta T} \tag{3.5}$$

Thus, we can formulate the actual data rate load $Q_{u_i x}$ of a user bearer in slice on base station from equations (3.4) and (3.5):

$$Q_{u_i x} = \rho_{u_i} R_{u_i\,x} \tag{3.6}$$

The slice has to allocate and prepare required resources by the Slicer to satisfy a user data rate each time trip as shown in equation (3.7):

$$(\rho_{u_i})_{t+1} \geq (\rho_{u_i})_t \tag{3.7}$$

At least the minimum $(\rho_{u_i})_t$ is required in the next t time trip $(\rho_{u_i})_{t+1}$ of scheduling allocation to satisfy the requirements of a user data rate. Notice that, sometime user data rate in $(\rho_{u_i})_{t+1}$is greater than the user data rate in $(\rho_{u_i})_t$ to satisfy the user demands (described in Section 3.2.1.4).

The overall slice bandwidth capacity over base station x is:

$$v_B = \sum_{u_i \in V} \rho_{u_i} \tag{3.8}$$

Therefore, the total slices bandwidth in the base station x is:

$$V_B = \sum_{v_i \in V} v_{B_i}, \quad where\ V_B \leq B_x \tag{3.9}$$

3.4.1.3. Slicer's Resource Allocation Using Exponential Smoothing Model

PRBs in a BS needs to be allocated and shared between slices based on resource requirements of each slice (as shown in Figure 3.4). Thus, each slice should provide an estimated value of the required resources and periodically send them to the Slicer. In order to achieve this, a slice controller needs to calculate required bandwidth of the slice periodically as shown in (3.8). The LTE Slice Controller Manager (LSCM) collects all estimated bandwidth values from slices and sends them to the Slicer. The Slicer uses these values to allocate PRBs of each slice efficiently. To enable this, we utilize the simple exponential smoothing model as shown in equation (3.10).

$$\lambda_{t+1} = \alpha \times (v_B)_t + (1 - \alpha) \times \lambda_t \tag{3.10}$$

where λ_{t+1} indicates the estimate of PRBs for each slice during the $(t+1)$ interval time. The λ_{t+1} describes slice status where it either requires additional PRBs or the slice needs to release some PRBs. λ_t refers to the current estimate amount of PRBs during $TTI(t)$ interval. t is the Slicer interval, which consists of a number of $TTIs$. α is a smoothing constant, which it serves as a weighting factor. It indicates how many intervals are taken in consideration of an average function. According to α we have reformulated equation (3.10) as follows:

$$\lambda_{t+1} = \alpha(v_B)_t + \alpha(1-\alpha)(v_B)_{t-1} + \alpha(1-\alpha)^2(v_B)_{t-2} + \cdots + \alpha(1 - \alpha)^{t-1}(v_B)_1 + (1-\alpha)^t\lambda_1 \quad (3.11)$$

where λ_1 represents a simple average of the $\sum_{t=1}^{n}(v_B)_t$, and α has a value between (0 and 1) where $(0 < \alpha < 1)$. In equation (3.11), too large value of t would result in making value of $(1-\alpha)^t$ close to zero.

Generally, λ_{t+1} have either positive or negative values when compared with λ_t. In the case of positive value, the slice needs more PRBs, whereas in the case of negative value, the slice operator is satisfied with the current state of allocation PRBs. The Slicer utilizes these values to calculate and allocate PRBs to each slice (virtual network). Moreover, this type of calculation is especially useful for network slicing within a contract from type DG, BE or BEMG. The DG contract represents the actual allocated bandwidth to slice operator for serving users' requirements, and the maximum bandwidth by the terms of contract. In respect of BE and BEMG contracts, the slicer determines the minimum requirements of type BEMG slice operator and the remaining PRBs will be assigned to type BE slice contract.

The isolation between slices is based on the fairness factor as calculated in the following equation (12):

$$FF_v = (\lambda_{t+1})_v / \omega \tag{3.12}$$

FF_v is the fairness factor of slice v; $(\lambda_{t+1})_v$ is the estimation of PRBs for slice v; ω is a total PRBs over all BE slices. The ω is computed in the following equation (3.13).

$$\omega = \sum_{v=1}^{\forall BE \text{ x } slices} (\lambda_{t+1})_v \tag{3.13}$$

The total number of PRBs (φ) allocated for each BE slice v is described as illustrated in equation (3.14).

$$\varphi_v = int(FF_v * \Upsilon) \tag{3.14}$$

where the Υ is the remaining PRBs after allocating guaranteed bandwidth to slices.

3.4.1.4. Max-Min Model for Users Fairness and Isolation in Slice

Generally, the scheduling mechanism should be fair and it should isolate the bandwidth between users in the same slice. To realise this, we use the Max-Min fairness model. The Max-min fairness means maximizing the minimum fair share of the bandwidth for each user within a certain slice. Three principal steps have to be considered in Max-Min mechanism:

1- Resources allocation is in increasing order of their demands

2- No user gets a share larger than its demands.

3- Users with unsatisfied demands get equal shares.

Let U_P be a set of users U with their bandwidth demands p in v such that these users are arranged in ascending order, which we formally define as follows:

$$U_P = \{\rho_1, \rho_2, \dots, \rho_N\} \quad \text{such that} \quad \rho_1 < \rho_2 \ \dots < \rho_N \tag{3.15}$$

To equally share a slice's resources (bandwidth) between users let's consider u_E is the bandwidth share of individual user u in slice v. u_E gets as follows:

$$u_E = v_b/N \tag{3.16}$$

where v_b is the total bandwidth of a slice v and N is the number of users in v. Therefore, the user will be protected by allocating the same bandwidth as other users. Not only that, allocated bandwidth represents the minimum satisfied requirement of a user service in slice v.

In some cases, the user's demands ρ are greater than the allocated bandwidth u_E, which means that the user is unsatisfied. In such a case, for all unsatisfied users, they will get the same (equally) extra bandwidth from the slice controller if it is available. In the slice, not all the users are unsatisfied. Some of them have more bandwidth than they actually need. Therefore, we can calculate the excess bandwidth and equally distribute it between unsatisfied users. Thus, assume that z represents the excess bandwidth for individual user u, we compute the value of z as illustrated in equation (3.17):

$$z = u_E - \rho \tag{3.17}$$

Now, for each unsatisfied user in slice v they will get z/x bandwidth, if we assume that x represents the number of unsatisfied users in v. The slice operator repeats this process by the slice controller each time if excess bandwidth is available. As a result, no users will get more allocated resources (bandwidth) than they need.

3.4.2. NSRM Algorithms

From the previous discussion on how the estimated resource model and the Max-Min model influence the resource allocation, we conclude that both models work in different tiers (intra, inter). The inter resource allocation is where the estimated resources are allocated among different slices, while the intra resource allocation is where the resources of a slice are allocated between different users in the slice. Therefore, we propose two algorithms for resource allocation namely, NSRM inter tier of resource allocation (Algorithm 3.1) and NSRM intra tier of resource allocation (Algorithm 3.2). Both algorithms are implemented partially or totally into the Slicer.

As mentioned, Algorithm 3.1 allocates resources between different slices. For that, it needs the required resource of each slice v_B and the total PRBs of an eNodeB Bx. The algorithm invokes the GET-PRBs function to calculate the estimated resources of each slice according to Eq. (3.11). Then, it finds a value of the total estimated resources of all slices. This algorithm checks whether the total value of slices is less than or equal to the total PRBs of the eNodeB. If so, the algorithm assigns a required resource to each slice, otherwise, all the slices continue with the same currently allocated resources until more resources are available in the Slicer. That is, sometimes the estimated forecasting of resource allocation of a slice is less than the current resource allocation. In such a case, Algorithm 3.1 will release the surplus resources to allocate to other slices that are unsatisfied with a current resource allocation.

Algorithm 3.1: NSRM Inter tier resource allocation

INPUT: $V,\ B_x$ /*set of slices in a base station*/

OUTPUT: $(\lambda_{t+1})_v$ /*PRBs for each slice within the base station*/

For all $v = 1$ ***to*** V **do**

$\omega_V = \omega_{V-1} + call\ (GET - PRBs)_v$ /* invoke GET-PRBs to get PRBs for a slice v */

end for

If $\omega_V \leq B_x$ **then**

$v = 0$

For all $v = 1$ ***to*** V **do**

$(\lambda_{t+1})_v / \omega_V$

End for

Else

$(\lambda_t)_v / \omega_V$

End if

If $(\lambda_t)_v > (\lambda_{t+1})_v$ **then**

Release $PRBs = (\lambda_t)_v - (\lambda_{t+1})_v$

End if

GET-PRBs: Sub-algorithm to assign PRBs to a slice

INPUT: α, v_B, λ_{1_v}

OUTPUT: return value of $(\lambda_{t+1})_v$ for the calling function.

/* Using **Eq. (3.11)** to calculate $(\lambda_{t+1})_v$ */

/* Where v_B can calculate in **Eq. (3.8)** */

In Slicer, the Algorithm 3.2 is responsible for intra tier resources allocation between users within the same slice. This algorithm needs the number of users' N in the slice along with their resource demands U_p and the overall resources allocated to the slice $(\lambda_{t+1})_v$ from the Slicer (Algorithm 3.1).

First of all, in this algorithm all N users get an equal share of resource u_E. Then, the algorithm 2 checks whether a user demand ρ_i is greater than $(u_E)_i$ or not. If ρ_i is greater than $(u_E)_i$ (i.e., the assigned resource for a user is unsatisfied), the algorithm will add the user to a list of unsatisfied users. This process will continue until all users are checked. Moreover, the algorithm will check if there is any user whose $(u_E)_i$ is greater than ρ_i. If so, Algorithm 3.2 will distribute equally the surplus resources from the user among all users in the unsatisfied list. This process continues until finishing all the users in the slice. As a result, all the users will meet their demand for resource allocation as much as possible and there is no user who will get more than what it needs.

Algorithm 3.2: NSRM Intra tier resource allocation

INPUT: $(\lambda_{t+1})_v$, N, U_p /* U_p set of users demand p in a slice */

OUTPUT: u_E /* the bandwidth for each user in a slice */

$v_B = (\lambda_{t+1})_v$ /* resource allocation for a slice v by the slicer */

$u_E \; v_b/N$

$x = 0$

For all $i = 1$ ***to*** N **do**

If $P_i \leq u_{E_i}$ **then**

$U_E[x] = u_{E_i}$ /* all unsatisfied users will store in U_E set */

$X = X + 1$

End if

End for

$i = 0$

while $u_{E_i} > P_i$ **do**

$z = u_{E_i} - P_i$

z/x /* for all the users in U_E get z/x share resources */

$i = i + 1$

End while

3.5. Chapter Summary

In this chapter, a network slicing architectural solution for resource allocation in LTE networks has been presented. The proposed solution is based on the simple exponential smoothing model that takes into consideration the estimated bandwidth that each slice needs periodically. Then, we propose Max-Min fairness solution for isolating and fair sharing of a distributed bandwidth between users. Moreover, the two propose algorithms for inter and intra slice resource allocation have been explained in this chapter. In the next chapter, we propose mobility management architecture for managing user mobility between different access networks, where the propose solution is based on the concept of network slicing.

Chapter 4: Mobility Management Architecture in Different RATs Based Network Slicing

4.1. Introduction

In this chapter, we propose a Mobility Management architecture in Network Slicing (MMNS) where each slice can manage its users across heterogeneous radio access technologies such as WiFi, LTE and 5G networks. In this architecture, each slice has different mobility demands and these demands are governed by a network slice configuration and service characteristics. Therefore, our mobility management architecture follows a modular approach where each slice has an individual module that handles the mobility functions and enforces the policy of mobility management of a slice.

Several benefits of applying our proposed architecture are: i) Sharing network resources between different network slices; ii) creating a logical platform to unify the resources of different radio access technologies which allows all slices to share the resources; iii) satisfying slice mobility requirements by enforcing a policy of slice mobility taking into account the network slice configuration and service requirements.

4.2. Network virtualization

4.2.1. LTE Network Virtualization

This section describes how we can virtualize the function of the EPC elements that were mentioned in section 2.2.1.1. Let us take three basic elements (MME, S-GW and P-GW) of EPC and put them in the same physical hardware platforms and logically

softwarize them. Therefore, in our context, the EPC Function Virtualization (EFV) is a process of virtualizing the network function (VNF). As shown in Figure 4.1, the MME-VNF, SGW-VNF and P-GW-VNF all of them sit in the same physical server [93], [94]. The hypervisor places the rules of which device should be placed logically in the platform. Moreover, the hypervisor in our context has a virtual switch (VSwitch) with which it can handle the traffic between different logical ports of VNFs and physical ports of the hardware server. For example, in the Figure, the interface (S11) connects between the MMEVNF and S-GW-VNF through logical ports by VSwitch, Also, the S1-MME interface connects MME with the eNodeB through the MME-VNF logical port and server physical port by VSwitch. In the same context, other interfaces represent a logical connection of different elements in the LTE network. In addition, if the P-GW intends to forward a message outside the core LTE network via the SGI interface. The logical port of PGW-VNF sends the message through the SGI interface to the VSwitch Controller, and the latter recognizes the direction of the message outside the core network, then it forwards the message to the server physical port to send outside LTE via SGI interface.

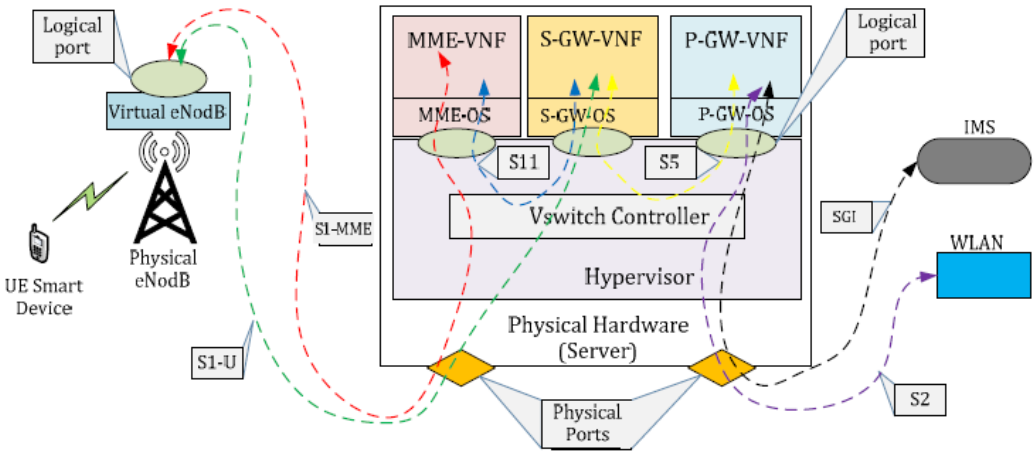


Figure 4.1: SDN and Virtualize Core LTE network

4.2.2. WiFi Network Virtualization

4.2.2.1. Virtual WiFi AP migration

The traditional way of designing WLAN is to follow up a micro cell architecture. The Figure 3 shows how the micro cell works. As depicted in the figure, each AP has its own coverage region and its own BSSID. When the client tries to start the establishment of a new connection to an AP, it sends a probe request message to see which BSSIDs are available (APs) so that it can decide the appropriate one to connect with [95]. For instance, let us assume that there are two APs and each one has its own unique BSSID (BSSID1 and BSSID2). As shown in Figure 4.2, when the client enters into AP1 coverage area at T1 it sends a probe request message and there is just AP1 with BSSID1, the AP1 can hear the message and responds to the client allowing the client to connect with it. As the client moves to T2, it starts to see the AP2, and at T3 the client notices that the radio signal strength (RSS) for BSSID1 becomes weak, therefore the client will make its own decision to figure out whether to continue with the current BSSID (in this case BSSID1) or to look for another one. It then starts to send a probe request message to the available APs and both BSSIDs for AP1 and AP2 will hear the probe message and respond. At this point, the client will choose which AP is appropriate to connect with (here AP2 is chosen).

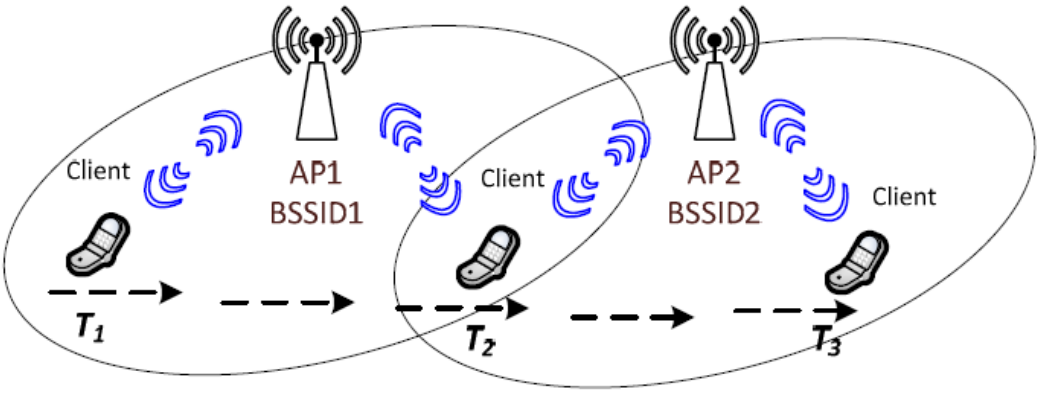


Figure 4.2: Traditional WiFi Architecture

From this scenario, we can notice a couple of things. Firstly, when an AP advertises its presence by BSSID, the responsibility of the client is to make a decision on whether to join the AP or not. Secondly, as a client moves, the decision of where the handover occurs is the client's choice. At the key point here, we want to take the decision of initiating a network connectivity away from a client because one client can affect the behaviour of other clients in the network. To take the decision away from a client, both APs should have the same BSSID from the client perspectives. As shown in Figure 4.3, when both APs advertise the same BSSID, it does not matter whether the client's position is at T1, T2 or T3 because it will hear just BSSID1. Furthermore, when the client sends a probe message to connect to an AP, it may hear the response from one AP or multiple APs, all of them having the same BSSID1 from the client view point (in such a case it takes a decision away from the client). In addition, as the client moves, it is up to the infrastructure to figure out which AP is in a better position to serve the client (AP1 or AP2) and from the client side there is no handover.

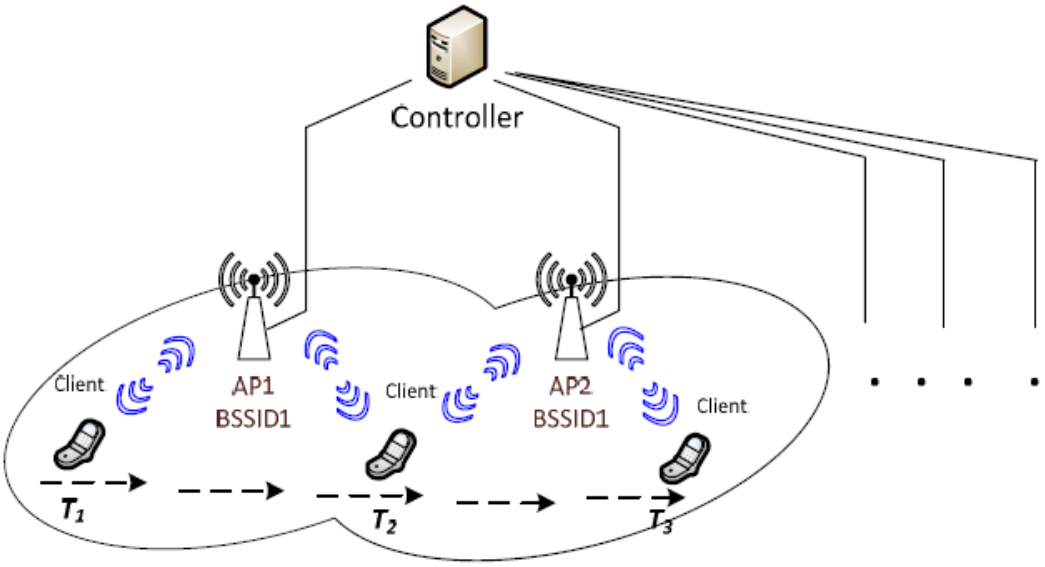


Figure 4.3: Virtual WiFi-APs Architecture

Let us explain the handover from a client perspective. As shown in the Figure 4.3, both APs are connecting to the controller and periodically APs compose a message digest of all devices (e.g., frame rate, number of transmissions, RSS, etc.) that receive the BSSID to the controller, so the controller has a global view of network status. Therefore, the controller can manage all the topology and assign which AP has a better link for the client to connect. In this case, the client does not experience any handover process because all the APs have the same BSSID resulting in what is called virtual AP.

4.2.2.2. Clients (UE) Virtual Port

Each device has its own personalized BSSID, if there are two devices within the same AP, each one has a unique BSSID [96]. Let us assume that there are two clients (UE1 and UE2) assigned to the same AP1 but each one has a different BSSID (we assume that BSSID1 for UE1 and BSSID2 for UE2). As the UE1 moves over from AP1 to AP2, at some point the controller decides that the AP2 is better to serve UE1; at that

point, depending on the topology design, the controller will send the BSSID1 from AP1 over to AP2. This process will continue in the same context as long as the client is migrating from one AP to another. Note that the BSSID associated with a client has all corresponding information related to a client (e.g., all the packets, all the sequence numbers, all the corresponding security state, etc.). The benefit of assigning a unique BSSID to each client (UE) is that the infrastructure has an ability to distinguish the service between APs for an individual client. The migration from virtual AP to the virtual UE port technique can create a switch like abstraction when each UE device effectively gets its own virtual port that allows the controller to handle a network topology per-device control in terms of channel access and security parameters.

4.3. Network Slicing in LTE and WIFI

4.3.1. Slice Assigning in LTE

When the service operator asks the LTE Slice Controller Manager (LSCM) and Slice Allocation (SA) to assign a slice for a service (S), as shown in Figure 4.4. There are three possible scenarios for assigning a slice to S. The first scenario is when the LSCM assigns the current slice to S. The second one is when the LSCM decides to expand the current slice to meet the S requirements such as video streaming. Lastly, this scenario is when the LSCM decides to create a new slice based on the new S technical and QoS requirements such as the remote monitor surgery service [97], [98].

Assigning a slice to a service S depends on the technical requirements t_s (e.g., mobility management, tunnelling, etc.) and QoS q_s (e.g., the maximum latency, minimum bandwidth). When the service operator requests to assign a slice to a certain service, it sends S requirements of the slice to the LSCM and SA. Where the LSCM will decide to assign a slice for S according to the following equations (4.1) and (4.2).

$$d_t\,(n) = t_n - t_s \tag{4.1}$$

$$d_q(n) = q_n - q_s \tag{4.2}$$

Where $d_t\,(n)$ and $d_q(n)$ represent the difference of requirements of the required slice (t_s, q_s) and the current slice (t_n, q_n). If one or both parameters have a negative value that means the current slice does not meet the technical or QoS requirements for S. In the case of expanding the current slice or creating a new slice, the LSCM's decision will be according to equation (4.3).

$$d_{ctech}(n) = (C_{en} + C_{oen} + l_{bn}) - (C_c + C_o + l_b) \tag{4.3}$$

For any slice n the LSCM calculates $d_{ctech}(n)$, which is the difference between the cost of expanding the current slice and creating a new slice-based service. C_{en} is the cost of expanding the current slice, C_{oen} denotes the effective operating cost of the current slice after expanding, and l_{bn} represents the cost of losing bandwidth for expanding the current slice. C_c is the cost of creating a slice based service, C_o denotes the cost of operation to create a new slice, and l_b represents the cost of losing bandwidth needed to create a new slice. If the value of $d_{ctech}(n)$ is negative that means the cost of expanding the current slice is less than the cost of creating a new slice, therefore the decision of LSCM to assign a slice will have the lowest value (in this case, the expansion of the current slice) and vice versa.

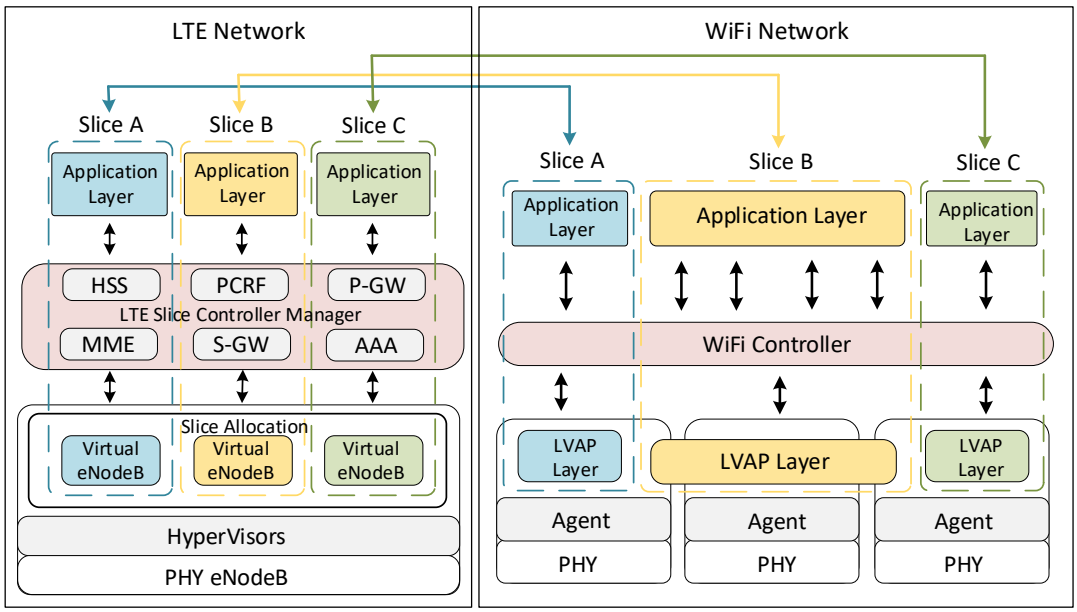


Figure 4.4: LTE-WiFi Slicing Networks

4.3.2. Slicing WiFi Network

When the UE seeks an AP, the WiFi controller will assign a new Light Virtual Access Point (LVAP). The LVAP abstracts the controls logical association and isolation of clients by assigning a unique BSSID to each client in order to connect to the AP (virtual AP) as described previously.

Each LVAP is allocated to a client by the WiFi controller. Its content contains information that enables the client to logically connect to, and isolate from, others in the same coverage area. An individual LVAP client contains a unique BSSID, one or more SSID, client MAC address, IP address, a set of open flow rules to manage the switch flow tables.

As described in the client virtual port, the benefit of the unique BSSID in LVAP is that the controller can distinguish a certain UE when moving between different APs.

This allows handling the handover of client between varying APs. From a client perspective, there is no handover because APs always have the same BSSID.

For slicing the WiFi network, the LVAP will assign a specific slice by defining a set of SSIDs, because these SSIDs are related to the specific slice in the LTE, as shown in Figure 4.4. When a UE is assigned to one of these SSIDs, it is automatically assigned to a certain slice.

4.4. Network Function in Network Slicing

Different network slices work on top of shared infrastructure, which is constructed of common hardware resources such as network functions virtualization infrastructure (NFVI). Also, it could work on the dedicated hardware such as network entities in the RAN. Each network slice is realized by a number of network functions NFs, which are either physical or virtual, based on the slice functionality. These network functions are controlled by SDN where the network can be classified into control plane (CP) and user plane (UP).

Despite the NFV and SDN concepts being completely different, they are highly complementary to each other. NFV can work as a virtual SDN controller (network function) to run on the cloud. This allows the SDN controllers to move to the optimal locations in the cloud. On the other hand, SDN provides logical connectivity between virtual network functions (VNFs) to optimize network traffic engineering [99].

End to end slices are sharing resources of the CN and RAN. For example, in the RAN domain, the shared NFs include monolithic and distributed base stations. In the CN, they share different virtual network functions (VNF), instances include mobility management and home subscriber server (HSS). According to 3GPP standards [100],

there are three solution groups of common functionality of the network slice as illustrated in Figure 4.5. Group A is depicted by deploying a common RAN and independent CN slices such that each network slice handles a user, and its mobility management, sessions and subscription. Group B assumes that all network slices are on a common RAN where mobility and subscription are shared between slices, while other functionality handles the network slice. Finally, group C assumes a fully shared RAN and a common CN control plane, but CN user plane is under a dedicated slice's control.

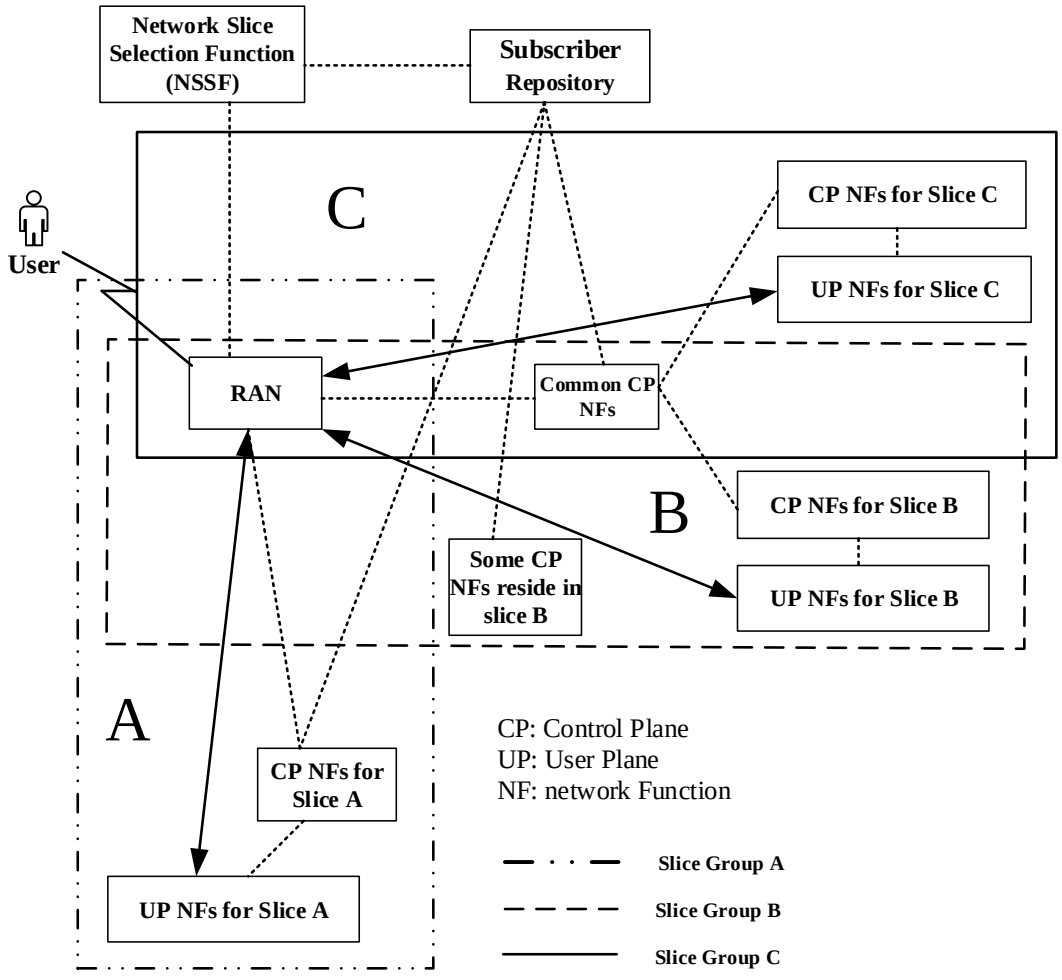


Figure 4.5: Network slice architecture with different groups of network functions (NFs).

4.5. Mobility Management Architecture

Today's network operators are facing many issues such as increasing mobile data traffic volume, congestion in users' dense area and the need for expanding the current network coverage area. Therefore, it is becoming important to find suitable solutions to overcome these issues. The offloading of data solutions appears as a promising solution to solve these networking issues. There are many mechanisms for offloading mobile data traffic such as capping user data, device-to-device data offloading and using complementary network to offload mobile data (e.g., WiFi network). According to [101], the WiFi network is one of the key players in data traffic where 20% of data in the outdoor environment is through the WiFi network, while 60% of data in the indoor environment is in landing WiFi network. Therefore, cellular network operators consider the WiFi network as a complimentary network to offload mobile data.

One of the most important aspects of the 5G network is the capability for managing heterogeneous infrastructure, where it creates a unified programmable platform based on abstracting different RANs as depicted in Figure 4.6. The abstraction platform unifies all the RANs resources and it is shared by different network slices where each slice has the ability to control its users in different access networks such as 5G, LTE or WiFi. As shown in the figure, the mobility management is centrally controlled by a mobility manager (controller). The mobility manager works based on a modular approach, where each slice has its modular unit residing in the controller. Thereby, each module enables the mobility management of a dedicated slice to support different operations, such as resource optimization and data offloading between different access networks and so on.

The general fundamental requirements of offloading data between any networks are:

- Seamless connectivity between two networks such as LTE and unlicensed network (WiFi).

- A common interface of multi-connectivity in the user's mobile device for available networks (offloading networks).

- Considering latency mechanisms to minimize the effectiveness of delay of current service during the offloading procedures, e.g., short path mechanism [102].

Different abstraction parameters are considered for network offloading where these parameters are distinct according to different access networks and most of these abstraction parameters come from physical network resources. Below, we provide brief definitions of potential parameters for abstraction, depending on the network interface [103].

- As we mentioned earlier the abstraction parameters depends on RAN-T. For example, WiFi network parameters include Received Signal Strength Indicator (RSSI), frequency bandwidth, power transmission, etc. whereas in the LTE network abstraction parameters include Quality of service Class Identifier (QCI), Physical Resource Block (PRB), Reference Signals Received Power (RSRP), Reference Signal Received Quality (RSRQ), etc.

- Available bandwidth is an important parameter where it represents the amount of radio resources available at a RAN node. Many factors affect bandwidth availability such as current Quality of Service (QoS) satisfaction requirements, channel capacity and backhaul network load.

- Spectral efficiency represents the capability of how many bit-rate can be transmitted over a current transmission bandwidth (in bps/Hz).

- Node capacity, which represents a composition of available bandwidth and spectral efficiency.

In the Next section, we discuss the seamless mobility management between different access networks, for example, we consider the seamless connectivity between LTE and WiFi networks.

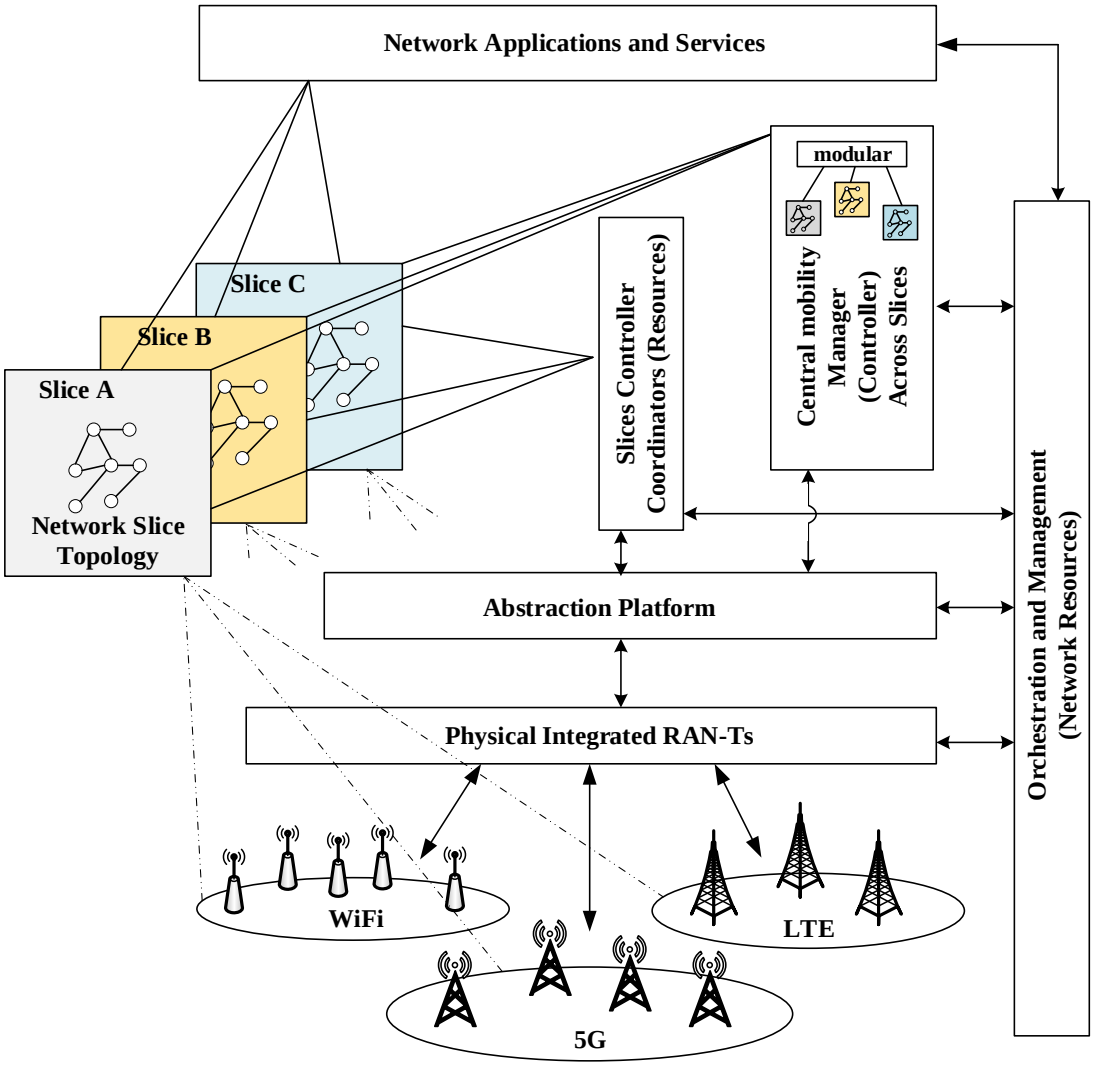


Figure 4.6: Mapping logical abstraction RAN-Ts between different network slices.

4.5.1. Seamless Connectivity of Different RATs in Network Slicing

In heterogeneous network environments, where a user moves between different access networks, the operator would always like to have control of his clients in different access networks in order to introduce better quality of services (QoS) and enhance user experience (QoE). In this work, we introduce a network slicing architecture in order to provide offloading of user flows between different access networks but under the same slice control. In our architecture, we have an abstraction layer that includes different logical shared resources of heterogeneous RAN networks. All network slices share this layer to assign resources to their users in different RANs.

In order to seamlessly assign a user to a certain slice, there is a controller in each slice that manages users in the slice and assigns a number to each user (ID-Slice). ID-Slice represents a slice identification for a user within the slice, meaning that whenever a user switches into different RANs and it has an ID-Slice, this helps to identify the slice to which the user belongs.

Let us consider LTE and WiFi networks to illustrate user seamless connectivity within a network slice. In the traditional LTE network, a network operator holds a user flows (bearers) setup. In the same manner, our proposal encompasses a slice operator that enables a slice operator to setup a user IP-flows. Moreover, the slice controller, during the setup tags an ID-Slice for each flow. In the same context, we assume that a slice operator takes care of the flow admission control to ensure that each flow gets enough resource requirement for QoS guarantee.

In our work, we consider that a UE device has the capability to use both interfaces (LTE and WiFi). Figure 4.7 illustrates the logical connection between network

elements. The P-GW works as an IP anchor, which does all the IP-flows admissions. Another node called Wireless Access Gateway (WAG) implements the necessary functions in the WiFi network. The routing is done between the P-GW and WAG by the LTE-WiFi Controller Flow (LWCF). It takes care of all the signalling between the P-GW and WAG to tunnelling the UE flow mobility from the LTE to the WiFi and vice versa. When a UE changes his network coverage location from LTE to WiFi, the slices controller coordinator assigns a new AP that has enough resource. At this point, all information of the AP is held by the abstraction platform. Note that, all UE information and status are held by the slice controller (e.g., IP addresses, port addresses, OpenFlow rules and ID-Slice which is same as SSID). In the case of any change in the UE locations, the slice controller tells the WAG to update the binding tables in the LWCF. One Home Address (HoA) has a number of Care of Addresses, which may be assigned in the binding cache table. In addition, there is another table called the flow-binding table, which specifies the type of traffic route to a corresponding CoAs. Both tables are sorted with respect to the priorities. The highest prioritized entry is at the top. They are linked together over the Binding Identity (BID) fields. If any item is missing in one of the tables, the highest priority binding entry is used by default. Finally, the novelty of the presented architecture is that seamless individual flows can be implemented for any of the interfaces (LTE and WiFi) under a specific slice.

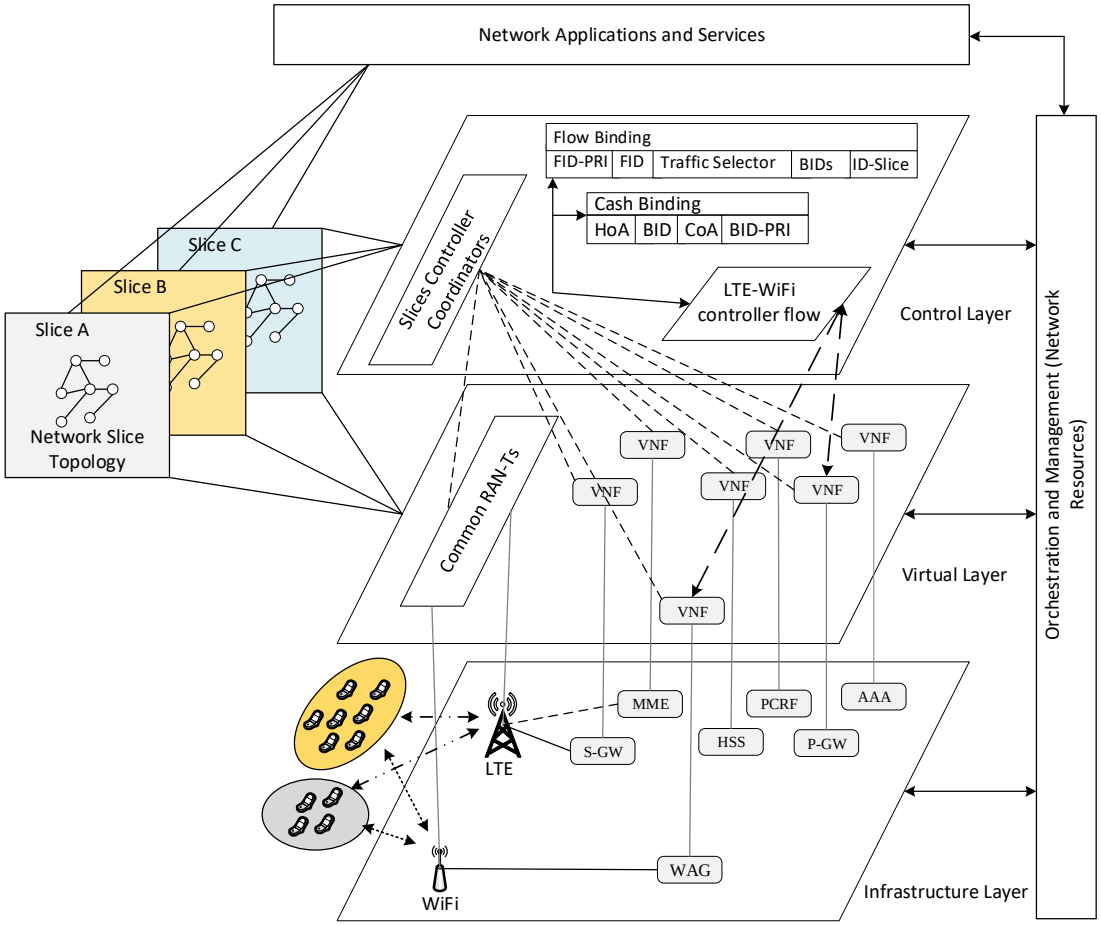


Figure 4.7. Logical connection LTE-WiFi network slicing.

4.5.2. Slicing association between LTE and WiFi networks

The operator would always like to control his clients in order to introduce the best quality service and user experiences. In this work, we introduce a slicing network architecture in a scenario where a UE moves between different access network interfaces (LTE and WiFi). A UE that is within a certain LTE slice network, and for any other reason such as offloading for better RSS, it triggers the handover process to another access network, in this case a WiFi network. Now, the question is how we can keep the UE under the same slice control after switching to a WiFi network coverage area.

If the UE were previously under a certain slice control of LTE networks, it would be controlled and managed by the slice operator. In case of the handover, the slice operator will provide the UE with a list of SSIDs that represents the slice in the WiFi network. When the UE moves to WiFi it will be assigned to one of these SSIDs. At this point, the UE will continue within the same slice that was within LTE network. As a result, we give a slice operator the capability of control over its UEs within a different access network (WiFi network).

4.6. Chapter Summary

In this chapter, we have introduced a new slicing network architecture between LTE and WiFi networks for managing data traffic. Utilizing the concept of SDN and NFV gives a capability to program a network infrastructure and virtualize a network functionality for enabling the network operator to have control over UEs in different access networks. However, in the LTE network, assigning UE to a certain slice is based on the technical requirements and QoS parameters for a service. On the other hand, the WiFi controller allocates a LVAP for each UE who wants to connect to the WiFi network and assigns an individual BSSID and one or more SSID to give an abstraction information about UE status in order to enable the WiFi controller and a slice operator to handle and manage UE mobility between WiFi-APs. Moreover, based on that, we have presented a logical mobility management architectural solution for network slicing based future 5G system. The control mechanisms have been discussed to unified resources of different RATs through the logical abstraction platform. Based on the modular approach, we have shown how each network slice is linked with the module, which is responsible for the mobility management of the slice. In the next

chapter, we propose mobility management handover for managing user mobility between different access networks.

Chapter 5: Mobility Management Handover in Heterogeneous Networks

5.1. Introduction

In this chapter, we address the handover for a user mobile device utilizing the proposed architecture in chapter 4. The proposed solution of this chapter is complementary to the seamless mobility architecture in the chapter 4. This solution considers many metrics when selecting an AP for a certain user, where it generates a list of APs of different interfaces. By considering various parameters, the APs are arranged in descending order in respect of the satisfaction parameters, such as user preferences (e.g., cost and location), services requirements (e.g., audio, video streaming and file sharing applications) and AP capacity in term of user density and throughput.

The main advantage of the proposed solution of the 5G network in this chapter, is that it enables the achievement of a seamless connectivity in the heterogeneous environment and selects an appropriate AP with cooperation between mobile devices and controllers.

5.2. Handover Operations

In the proposed architecture before we explain the handover process in detail, we identify two terminologies regarding the handover processing: homogeneous and heterogeneous handovers. The homogeneous handover in our context means when a user moves between two APs (base stations) belonging to the same type of access network. In contrast, the heterogeneous handover means that a user moves between two APs each one from a different access network such as LTE and WiFi as illustrated

in Figure 5.1. The handover starts when the mobile device requests to join a new AP, it will initially send a request message to the SDN-controller, and then the controller sends the assigned decision back to the current AP which the user is connected to as shown in Figures 5.2 and 5.3.

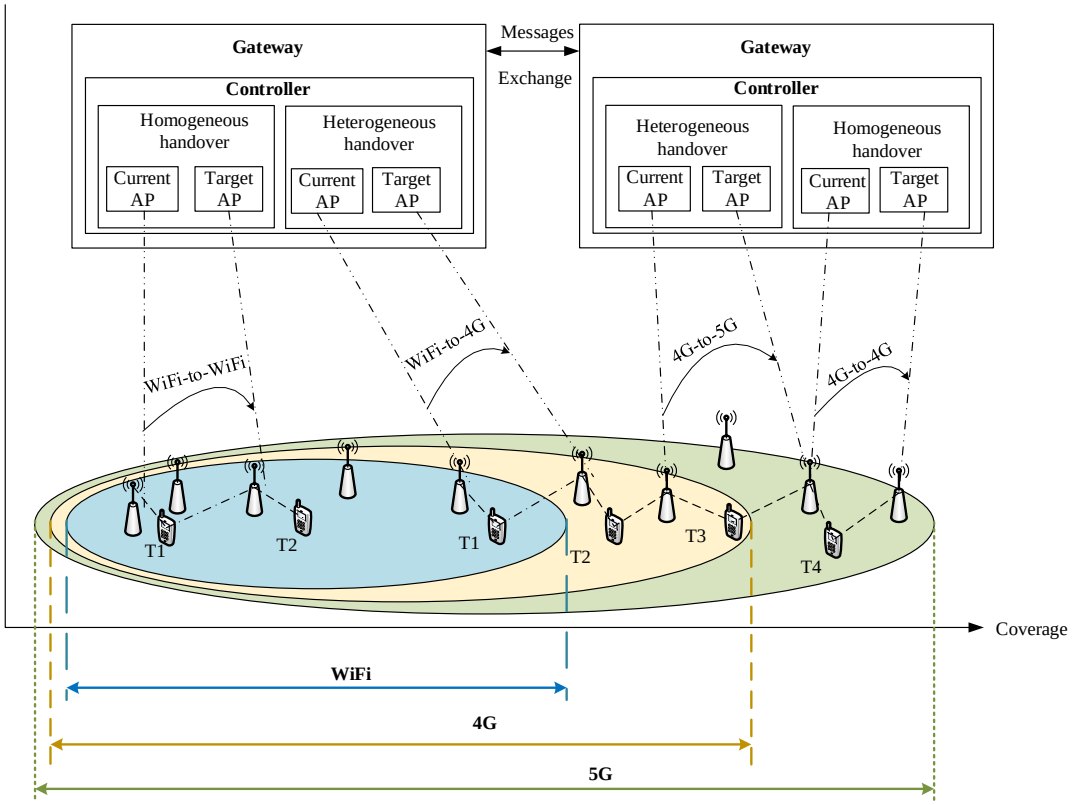


Figure 5.1: Homogeneous and Heterogeneous Handover

5.2.1. The Homogeneous Handover

In case of homogeneous handover, the user mobile device sends a forward request through a current AP to the SDN-controller, then the SDN-controller acknowledges the request to start the handover processing through steps, which is shown in Figure 5.2. The procedure and steps for the homogenous handover are described as follows:

1) The mobile device sends a request message to switch to another AP. This message contains a list of APs arranged according to the highest RSS in multi-interfaces of

access networks along with a set of parameters specified by location, time and current service.

2) The current AP forwards this message to the SDN-controller to select the highest satisfaction AP among different APs.

3) Then the controller acknowledges the request to start the handover steps based on the selected AP (homogeneous handover).

4) Handover starts from the current AP, when the current AP sends the message request with measurement parameters to the target AP.

5) The target AP sends acknowledgment message to the request of the current AP, then the current AP forwards the message to the mobile device to confirm the handover to the target AP.

6) The transmission between the mobile device and the current AP will pause for a certain time, and the current AP then buffers the current data service of mobile device.

7) The mobile device sends a status message that contains the location, the device status and the current service, to the target AP.

8) At the same time, the current AP also sends a message to start transferring the buffered data packets to the target AP.

9) The target AP will acknowledge the message of the mobile device and exchange the buffered data.

10) Finally, the target AP sends a request to the controller to establish a new path to the destination to exchange the data service with the mobile device.

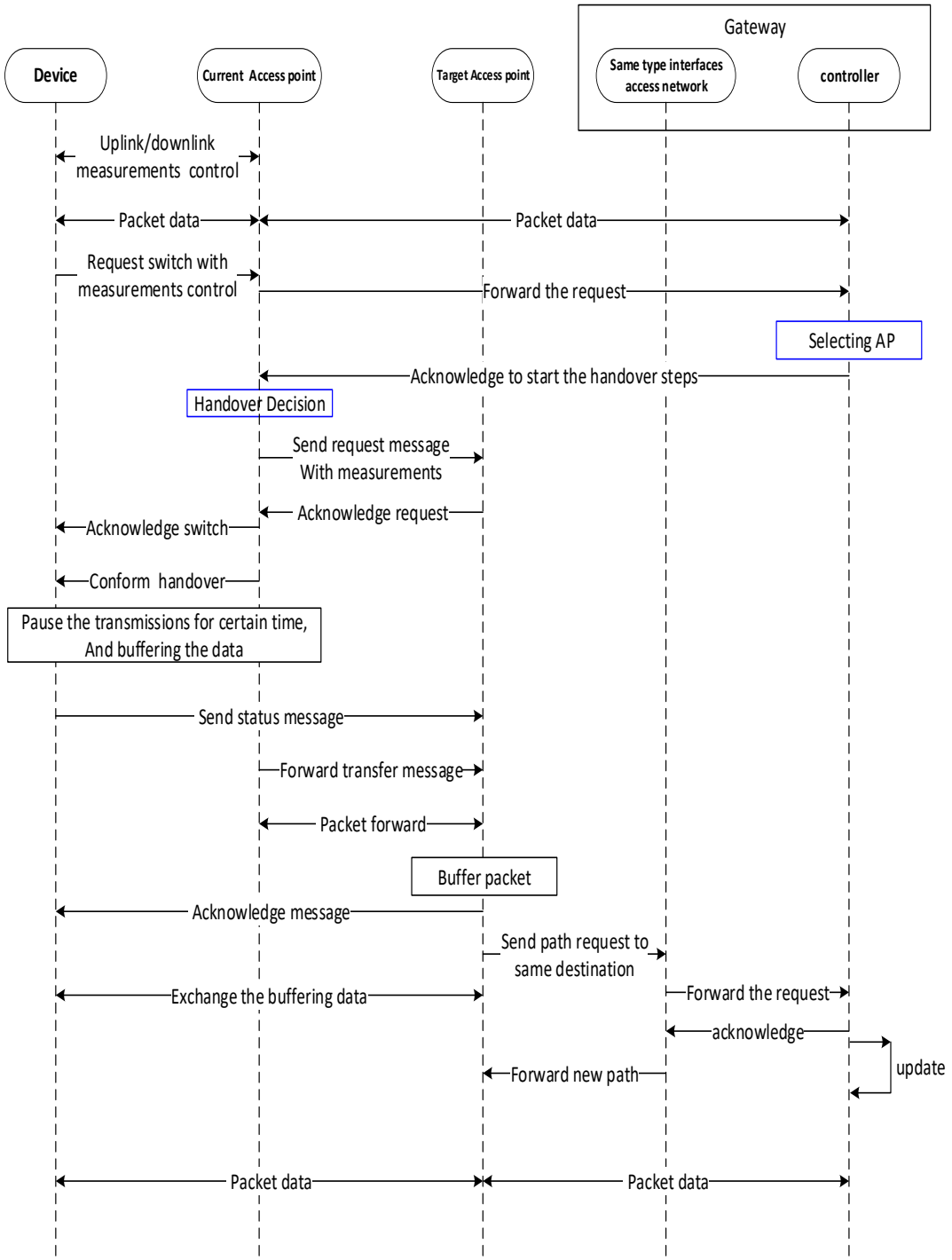


Figure 5.2: Sequence Messaging of Homogeneous Handover

5.2.2. The Heterogeneous Handover

Within the same context, the heterogeneous handover procedure is where the current AP forwards the request message of a mobile device to the SDN-controller. Then, the

SDN-controller sends a control message (serve) to the multi-interface part in the gateway to serve the mobile device request. After sending an acknowledgement message to the current AP of the mobile device, the handover steps will start as shown in Figure 5.3. The procedure and steps of heterogeneous handover are described as follows:

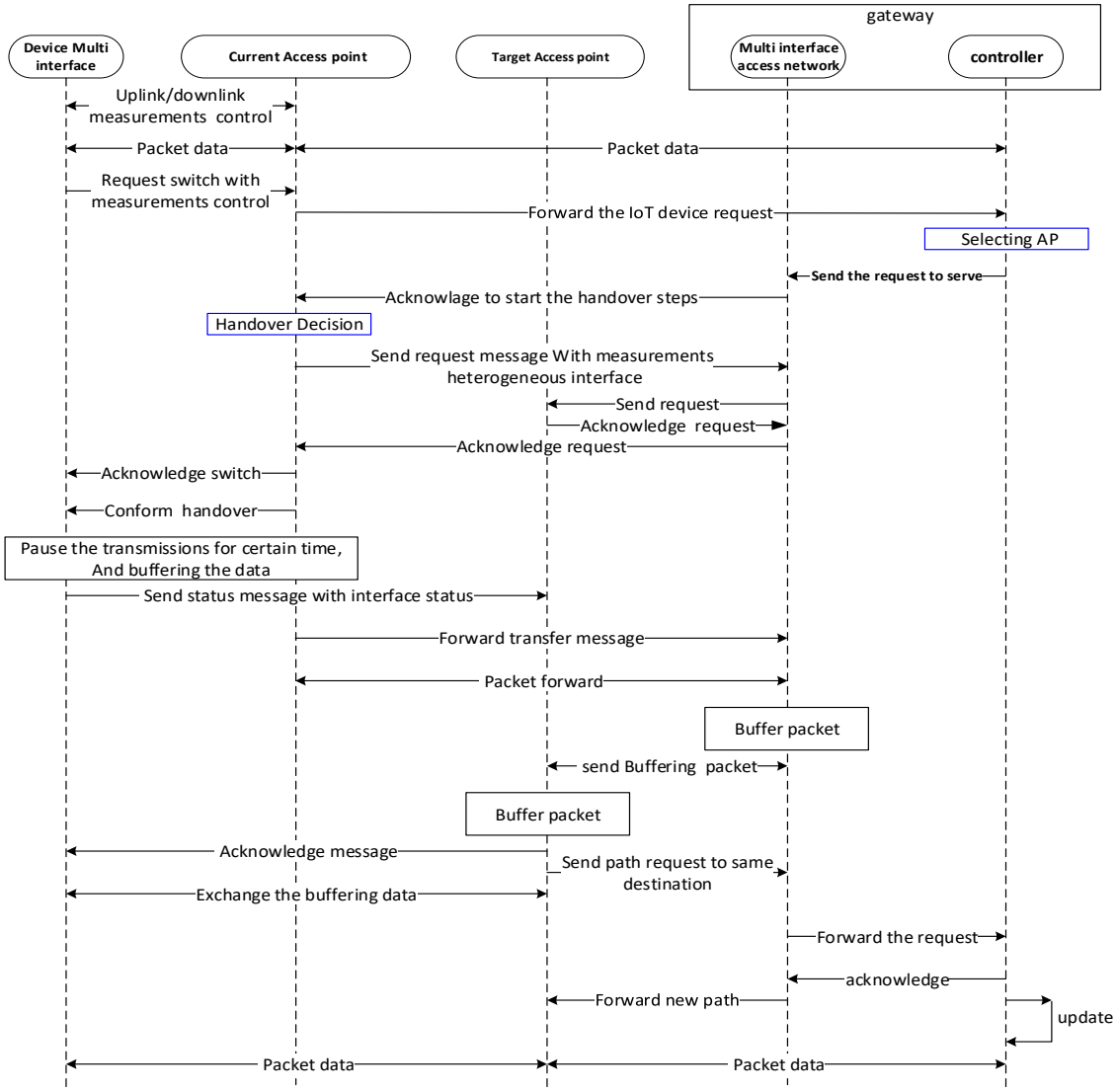


Figure 5.3: Sequence Messaging of Heterogeneous Handover

1) The mobile device sends a request message to the SDN-controller in order to switch to another AP. This message contains a list of APs arranged depending on the highest

RSS in multi-interfaces of access networks along with a set of parameters specified by location, time and current service.

2) The current AP forwards the message to the SDN-controller in order to select the highest satisfied AP among several APs (heterogeneous handover).

3) Then the controller sends the request to the multi-interface access network part in the gateway to start the handover process because the target AP is from a different interface of access point.

4) Then the multi-interface access network part acknowledges the request message to start the handover steps based on the selected AP.

5) Handover starts from the current AP, when the current AP send the message request with measurements to the multi-interface access network part, this part forwards the message to the target AP.

6) The target AP sends the acknowledgement request to the current AP via the multi-interface access network part, then the current AP forwards the message to the mobile device to confirm the handover to target AP.

7) The transmission between device and current AP will pause for a certain time, and the current AP buffers the current data service of the mobile device.

8) The mobile device sends a status message, which contains a location, the device status and the current service, to the target AP

9) At the same time, the current AP also sends a message to the target AP, via the multi-interface access network part, to start transferring the buffered data packets as shown in figure 5.

10) The target AP will acknowledge the message of the mobile device and current AP to exchange the buffered data.

11) Finally, the target AP sends a request to the controller in order to establish a new path to the destination to exchange the data service for the mobile device.

5.3. Selecting AP System Model

In this section, we describe the context information that allows a designer to specify a consideration of suitable parameters for a networking environment, then, explain briefly all parameters that we consider in the proposed solution. Next, we discuss the algorithm of selecting a list of APs in the mobile device and the policy of selecting a suitable AP to satisfy user and network demands.

5.3.1. The Context Information of Network Environment

The network periodically and dynamically catches a context information for different network elements and observes the changes of this context to be able to adapt them based on user device and network environment characteristics. Context information as considered in [104][105] as allowing a network designer to customize the network and application creation at the same time for ensuring that application operation is compatible with all aspects of application design requirements, not just with the preferences of the individual user but also with the preferences of network architecture design and network provider.

There are many networking characteristics that could be extracted from context information:

• Distributed control and management of context sources.

• Help to reduce of network complexity by sharing context-based information

• Dissemination of specific data among different nodes or through cross layer messages inside the same node.

• The integration of autonomics: These enable the efficient representation of available information, needed for context handling and distribution.

Context entity describes the characteristics situation of the entity in the location, environment, identity, activity and time, utilizing the context information to give an answer for questions on who (identity), when (time), where (location) and what (activity) represent context entity. In this work we will consider some context information requirements which are distributed between users (mobile device), networks and services as describe in the Table 5.1.

Table 5.1: Representations Context Information Classes.

Context classes	Definition
Service & application context	Information that describes a service and application requirements. A service context: such as service QoS (response time, availability, execution cost), service roaming state, service network endpoint, etc., service classification such as (streaming, interactive, background, conversational) Application context: bandwidth, packet loss ratio (PLR) packet error rate, jitter, delay
User context	Information that characterizes the user's preference, subscription information and situation, such as location, presence, current activity, social relation and preferences. user preferences: network values: bandwidth, network type, power consumption, security, RSS, power network-independent values: quality, lifetime, cost Subscription information: user profile, user status (location, mobility, etc.)
Network & link context	Information that describes underlying networks. Network context: Topology, network traffic performance (delay, packet loss rate, load), network cost, supported classes of service, network coverage. Link context (network interface): Signal-to-noise ratio (SNR), received signal strength (RSS), bite error rate (BER), signal-to-interference ratio (SIR)
Device context	Information that describes device hardware and software configuration and dynamic information (such as current battery power level, memory consumption, power consumption rate, RSS of available access networks)

5.3.2. Model Parameters for selecting APs

In this work we consider four parameters representing user preferences, network and link status, service requirements and user device status, as shown in Figure 5.4.

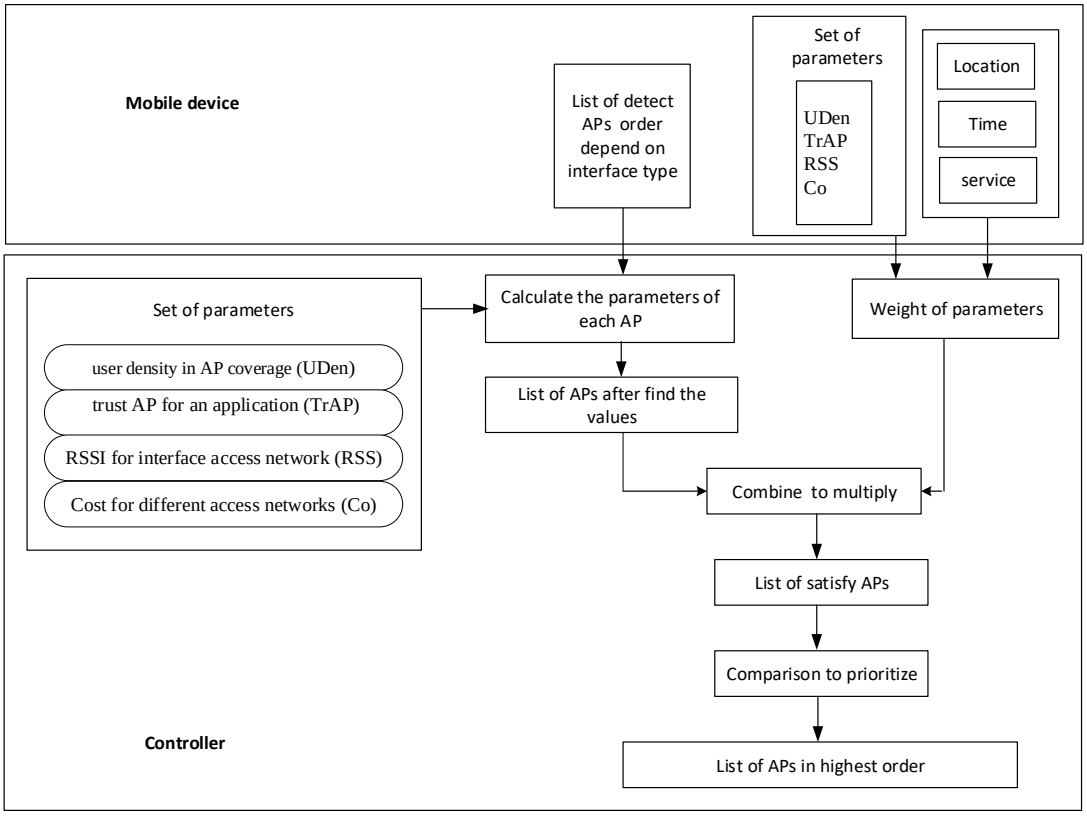


Figure 5.4: model for selecting AP

5.3.2.1. User density

In the proposed architecture, the rationale for computing user density is to optimize the controller selection process for a particular AP in terms of joining a new mobile device. Moreover, the user density value is to enhance an AP selection to a particular user-device with the highest bandwidth to switch to an alternative AP coverage with a sufficient capacity.

In the case where multiple APs are selected with the same capacity but different number of user devices, the AP with fewer of devices will be chosen. Whereas, in the case where a mobile device wants to join a new AP, the controller may not allow the mobile device to switch to a particular AP, unless it forces one of current mobile devices to switch to another AP.

To compute user density within AP, we need to calculate the current number of user devices within the AP coverage area, and the percentage of bandwidth consumptions for each user device by the current service. The AP holds a current percentage of bandwidth use for each user device in the coverage area.

Let us consider the current user mobile device ($UDev$) bandwidth (B) is ($UDev_B$). Then, the Total Access Point (TAP) bandwidth is (TAP_B). Therefore, the percentage (P) of individual user device bandwidth in AP coverage area is ($UDev_P$):

$$UDev_P = \frac{UDev_B * 100}{TAP_B} \tag{5.1}$$

The percentage of the current bandwidth use in overall AP coverage area (CAP_P) is:

$$\mathrm{CAP}_P = \sum_{i=0}^{n-1} UDev_P \tag{5.2}$$

where n represents the number of user devices, i is an integer number. Then we can find the value of user density ($UDen$) as follows:

$$UDen = \begin{cases} n\,, & AP1.CAP_p = AP2.CAP_P \\ CAP_P\,, & AP1.n = AP2.n \end{cases} \tag{5.3}$$

In the equation (5.3) the value of $UDen$ is either equal to the same number of user devices in AP coverage area if the comparison of the CAP_P between two APs are same or equal to CAP_P if the number of user devices are same.

5.3.2.2. Trust AP

In the wireless network connections, the user always looks for a secured connection. Users often use applications requiring highly secured connections because these applications are very critical of the level of risk. Therefore, the user device requires a high level of trust when linking to an access network. Example of such applications are, banking transaction application running on smart device, and in cases of very important calls (calls dealing with secure information). Furthermore, it is very important for a user device to frequently select a reliable AP that the network provider supports with respect to the security policy agreement in the user-SLA.

In this context, we propose that a user should have a List of Visited Locations (LVL) that represents a list of APs identities provided by a network operator to the user, which the user will use for selecting a Trust Access Point ($TrAP$) to connect to a different access network.

Generally, when a user moves carrying a smart device it is crossing different locations and trying to connect to the internet via different APs. Through this mobility situation, the user identifies these APs automatically via the LVL. According to this scenario, there are two possibilities, either the user device identifies a dedicated AP within the list of LVL or the user device does not acknowledge the AP within the LVL.

We assume that the calculation of ($TrAP$) is a Boolean value:

$$TrAP = \begin{cases} 1, & if\ AP\ trusted \\ 0, & if\ AP\ not\ trusted \end{cases} \quad (5.4)$$

When a mobile device interface detects a particular AP to connect with, it checks whether the AP is identified within LVL or not. If a mobile device finds the AP in LVL, then the value of $TrAP$ equals 1 otherwise the value of $TrAP$ is 0.

In our work, the needs for a trusted AP depends on the application currently being used by a user. Some applications do not need to trust the AP to connect. For example, when connecting to consume audio or video stream. Where an application needs $TrAP$ to connect with a network and if it cannot find $TrAP$ at the current time, this application will be blocked until the $TrAP$ is found or the user authorizes a particular AP and marks it as a trusted AP for the application.

5.3.2.3. RSS & COST

The values of these two parameters are evaluated from the interface profile of the Access network. The access interface of the mobile device measuring the RSS of APs and each access network identifies the cost of bandwidth used per units.

5.3.3. Algorithm of Selecting APs by User Device

In the proposed system, the user device sends a message to the controller content, a set of above mentioned parameters based on a service currently use and a list of APs ordered according to the type of interface and the values of RSS. The algorithm 5.1 is a priority list for APs, which depends on interface type. It considers access points (AP_i) with the RSS to create a priority list of $PriAP[i][j]$ which consists of a number of Rows and Columns. The rows represent the type of interface (i) that the user device

has to connect with to access networks such as (5G, LTE and WiFi), and the columns illustrates the number of APs ordered according to the highest RSS within a certain interface.

Algorithm 5.1: Access Points sent by the user device to controller

INPUT: AP_i, RSS, I, j

OUTPUT: $PriAP[i][j]$

For all $i = 1$ ***to*** 3 **do**

For all $j = 1$ ***to*** 5 **do**

If $AP_i = 1$ **then**

$n = j$

For all $j = 1$ ***to*** 5 **do**

If $PriAP[i][j-1].RSS >$ $AP_i.RSS$ **then**

$PriAP[i][j] = AP_i$

Else

$Temp = PriAP[i][j-1]$

$PriAP[i][j-1] = AP_i$

$PriAP[i][j] = Temp$

End if

End for

$j = n$

End if

If $AP_i = 2$ **then**

$m = j$

For all $j = 1$ ***to*** 5 **do**

If $PriAP[i][j-1].RSS >$ $AP_i.RSS$ **then**

$PriAP[i][j] = AP_i$

Else

$Temp = PriAP[i][j-1]$

$PriAP[i][j-1] = AP_i$

$PriAP[i][j] = Temp$

End if

End for

$j = m$

End if

If $AP_i = 3$ **then**

$k = j$

For all $j = 1$ ***to*** 5 **do**

If $PriAP[i][j-1].RSS >$ $AP_i.RSS$ **then**

$PriAP[i][j] = AP_i$

Else

$Temp = PriAP[i][j-1]$

$PriAP[i][j-1] = AP_i$

$PriAP[i][j] = Temp$

End if

End for

$j = k$

Let us consider an example to illustrate the algorithm:

In this illustration, we have a mobile device with three interfaces for the access networks, then the device detects many APs around with each interface. Let us assume that the mobile device with each interface detects five APs, then if we apply the algorithm the result will be:

$$PriAP[i][j] = \begin{bmatrix} AP_{1,1}\, AP_{1,2} AP_{1,3} AP_{1,4} AP_{1,5} \\ AP_{2,1} AP_{2,2} AP_{2,3} AP_{2,4} AP_{2,5} \\ AP_{3,1} AP_{3,2} AP_{3,3} AP_{3,4} AP_{3,5} \end{bmatrix} \tag{5.5}$$

where $PriAP[i][j]$ denotes this matrix where the i^{th} represents the type of interface (5G, LTE and WiFi), and j^{th} is the number of APs, in this case $j = 5$ for each i.

5.4. The Policy for Handover

In this section, we describe the working of the policy for the handover of the proposed solution. When the mobile device moves between different types of access network it discovers many APs through various interfaces, consequently it creates a list of the heterogeneous APs. Thus, when a mobile device needs to switch to the most suitable AP, this will be chosen from the list it has created earlier through the handover steps. Our handover policy works based on a utility function where conditional rules depend on the time, location and service. The results of the logical decision are based on these conditions either True or False. So that, these values lead to trigger a set of events guiding the process of assigning an AP.

The selection of an appropriate AP for a mobile device is based on a set of parameters ($SetP$) that is extracted from the contextual information distributed between user preferences and profile, device context, application and service requirements and the environment conditions. The formula of $SetP$ is a set of weighted parameters as follows:

$$SetP = (\mathrm{W_{UDen}}, \mathrm{W_{TrAP}}, \mathrm{W_{RSS}}, \mathrm{W_{Co}}), \quad where\ \mathrm{W_{UDen}}, \mathrm{W_{TrAP}}, \mathrm{W_{RSS}}, \mathrm{W_{Co}}\ \ between\ 0\ and\ 1 \tag{5.6}$$

Note that, the Metrics are considered to evaluate $SetP$ for an application are as follows:

• $UDen$: measured the user density in AP coverage

• $TrAP$: trust access point for an application

• RSS: measured the RSSI for interface access network

• Co: cost for different access networks

Let us consider an example, of our policy of decision-making, where we assign a value of each parameter such as:

$$\text{If Location} = \text{restaurant \& time} = \text{night \& service} = \text{VoIP then} \quad (5.7)$$

$$SetP = (0.4, 0.1, 0.1, 0.4)$$

The mobile device sends a message to the SDN-controller carrying information about current services parameters and a list of APs discovered. Then, the controller calculates these parameters for each individual AP in the list from the user device. Then it compares the set of parameters from the user device with the AP parameters to evaluate the satisfaction value for AP. This process repeats with all APs in the list and prioritizes the list depending on the highest values of each AP (as describe in section 3.5).

5.4.1. Assigning Access Point

In order to assign appropriate AP to a user device, the SDN-controller selects this AP from the list of capable APs. The SDN-controller calculates the same parameters of each AP. Then, it orders the APs depending on which one has the highest sufficient value. Thus, those with the highest values are considered to be more eligible to select for user mobile connection.

Here we describe the method for selecting access points as illustrated in Figure 5.4:

• The controller receives a message from a mobile device containing a list of APs and weight of weighted parameters.

• Each AP has values of the same parameters calculated by the SDN-controller.

• The final result be a vector of metrics for each AP represented as:

$$\overrightarrow{\mathrm{SetAP}(\mathrm{AP_{i,j}})} = \begin{pmatrix} \mathrm{SetAP_{UDen}}(\mathrm{AP_{i,j}}) , \mathrm{SetAP_{TrAP}}(\mathrm{AP_{i,j}}) , \\ \mathrm{SetAP_{RSS}}(\mathrm{AP_{i,j}}) , \mathrm{SetAP_{Co}}(\mathrm{AP_{i,j}}) \end{pmatrix} \tag{5.8}$$

• Then, the SDN-controller will add the $\overrightarrow{\mathrm{SetAP}}$ from a user device with the $\overrightarrow{\mathrm{SetAP}(\mathrm{AP_{i,j}})}$ vector to get the value of satisfied connection of each AP ($QuaAP$):

$$\mathrm{QuaAP}(\mathrm{AP_{I,J}}) = \overrightarrow{\mathrm{SetAP}}.\overrightarrow{\mathrm{SetAP}(\mathrm{AP_{i,j}})} \tag{5.9}$$

$$(\mathrm{W_{UDen}}\ \mathrm{W_{TrAP}}\ \mathrm{W_{RSS}}\ \mathrm{W_{Co}}).\begin{pmatrix} \mathrm{SetAP_{UDen}}\ (\mathrm{AP_{i,j}}) \\ \mathrm{SetAP_{TrAP}}(\mathrm{AP_{i,j}}) \\ \mathrm{SetAP_{RSS}}(\mathrm{AP_{i,j}}) \\ \mathrm{SetAP_{Co}}(\mathrm{AP_{i,j}}) \end{pmatrix} \tag{5.10}$$

$$\begin{aligned} &(\mathrm{W_{UDen}}.\mathrm{SetAP_{UDen}}\ (\mathrm{AP_{i,j}}) + \mathrm{W_{TrAP.}}\ \mathrm{SetAP_{TrAP}}(\mathrm{AP_{i,j}}) \\ &\quad + \mathrm{W_{RSS}}.\mathrm{SetAP_{RSS}}(\mathrm{AP_{i,j}}) \\ &\quad + \ \mathrm{W_{Co}}.\mathrm{SetAP_{Co}}(\mathrm{AP_{i,j}})) \end{aligned} \tag{5.11}$$

For $AP_{i,j}$, the j^{th} represents the number of APs for the i^{th} access network interface.

• Then the controller prioritizes the $QuaAP(AP_{I,J})$ in descending order where the first AP has the highest value of the selected parameters.

Note that, the SDN-controller uses the remaining APs in the $QuaAP(AP_{I,J})$ in case the currently assigned AP is faulty. Then, the SDN-controller will switch the user's device to the second AP in the $QuaAP(AP_{I,J})$ list.

5.5. Use Case Scenarios

Future mobility network (i.e., 5G system) is considered as a big challenge in terms of the variety of user devices and applications that generate a huge data volume. In this work, a mobility management focuses on link continuation. The link connection properties are changing during a movement between different base stations and access points, which are attached to the user's device. The data session has to continue during mobility, where there are two methods of session continuity (seamlessness) either through the fixed IP address or coping with a current attached point address change.

The different request of mobility cannot be handled by a single solution. Therefore, our architecture has many modules to adapt to different network configurations according to the service or slice requirements (this type of approach is called mobility on demand). At this point, the main challenge is how to identify the actual demands in accuracy with respect to selecting an appropriate solution of mobility to fit a scenario demand. Different criteria have to be taken into consideration when selecting the solution such as the end device specification and the surrounding environment (e.g., the smartphone in the dense area or sensor attached to car). Furthermore, the network condition should be taken into consideration (e.g., the load of neighbour access points, different access technologies or QoS parameters).

Taking into account the aforementioned requirements of mobility, different available scenarios could be identified where a user device needs to offload from a current cellular network (e.g., LTE) to the WiFi network. These scenarios are different from each other depending on the current user services. A user may have one or more connection flows representing different services. Consequently, in the case of offloading users into the WiFi network, it is either offloading all user flows or selecting

some of them. Selective flows provide better user experience with services that are sensitive to delay such as online gaming. Therefore, such services have higher priority to stick with a cellular network rather than offloading to WiFi, while services such as FTP download can be offloaded to WiFi because it is not sensitive to delay when switching to WiFi.

Today, network operators pay attention to the WiFi network as a complimentary network to deploy it to offload their data network and extend their customer services such as voice over WiFi (VoWiFi) and video over WiFi. In the context of Voice over LTE (VoLTE), the VoWiFi is a complementary service of VoLTE, both of them utilizing IMS voice specification where the voice delivers across network based IP protocol. The seamless offloading scenario is possible between LTE and WiFi and vice versa. Similarly, video over WiFi follows the video over LTE (ViLTE) in the IMS technology.

The scenarios provided above can be deployed in many real-life situations. For example, when a user is in a region where there is no cellular coverage and that user needs make a call, such as the London tube. Also, in the case where a customer exceeds a monthly subscription bundle, the operator with VoWiFi may be avoided the customers from an extra charging service.

5.6. Chapter Summary

In this chapter, we have shown the proposed handover solution in heterogenous wireless networks to support user mobility management; where identified two procedures steps (homogenous and heterogenous) of handover. Four parameters have been used (i.e. $UDen$, $TrAP$, RSS and Co) to run the handover policy for selecting

AP. The proposed solution was enabled to select an appropriate AP with cooperation of mobile devices and controllers to maximize network performance and satisfying service requirements and user's demands in a dynamically changing of network environments.

Chapter 6: Simulations and Results Evaluation

6.1. Introduction

Two different simulations have been used to evaluate the work that was achieved during the course of this thesis, namely OPNET Modeler and OMNET++. This chapter presents two main parts. The first part describes simulations entities and the second part explains the performance of results evaluation. The OPNET Modeler was used for evaluating the Resource Management in network slicing, whereas, the OMNET++ was used for evaluating the seamless connectivity in mobility management.

The simulations part is presented in the first section of the chapter, which it deals with two simulations as mentioned earlier (OPNET Modeler and OMNET++). OPNET Modeler is a commercial software simulation that implement several networks infrastructure and applications and it designs in a hierarchical modelling environment based on C and C++ programming tools [106]. The hierarchical environment consists of several editors, starting with the project editor, node model editor, process editor and the code editor. On the other hand, the OMNET++ is open source software simulation and is also used for implementing different networks and applications [107]. It is an objected-oriented modular discrete event framework and it has modules in C++. These modules can be combined to build any network simulation and it is truly reusable.

The results evaluation part deals with the overall evaluation of the proposed solutions. It presents the performance evaluation through different scenarios to validate two proposed solutions, namely: Resource Management for network slicing in LTE network and the seamless connectivity for mobility management in network slicing.

6.2. The Simulations tool

We used two different simulation to evaluate our work due to the fact that, at the very early stage of this work we started with the OMNET++. After that, during the evaluation procedure we need to apply some scenarios on the LTE network where different scheduling mechanisms are needed, which is not available in OMNET++. This pushed us to look for alternative simulations. Therefore, we selected the OPNET Modeler because it supports what we need. However, in this section we describe the framework environments of both simulations (OPNET Modeler and OMNET++).

6.2.1. OPNET Modeler framework environment

To validate the proposed models of resource management for network slicing in the LTE network, in this thesis, we use the OPNET Modeler to investigate different scenarios for performance evaluation. The network topology in our simulation is presented in Figure 6.1. This topology illustrates an LTE network with one eNodeB and 10 mobile nodes, which it could be extended to any number of nodes (only limited by the system memory). In the topology, all the wired connections nodes are linked through 100BaseT cable. The scenarios we consider over this topology are based on a comparison study of the performance between the standard LTE network (a legacy network) and the proposed network slicing mechanism (NSRM). Notice that, the implemented model of the LTE network is not the same one that comes with the OPNET installation version; it holds more details to fit our proposed solution, such as mapping the virtual MAC layer. Over the next subsections, we explain some nodes of the network topology.

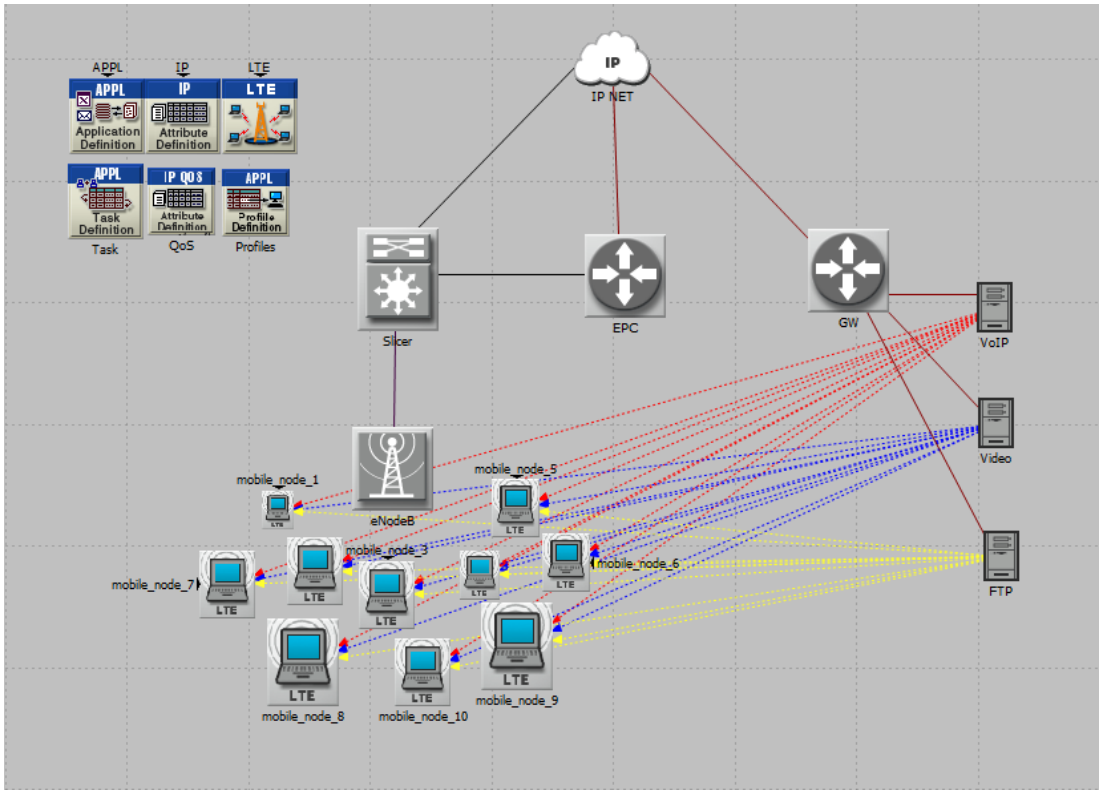


Figure 6.1: Network Topology.

6.2.1.1. Mobile Node Model

The mobile node model is depicted in Figure 6.2. This equipment has many predefined protocols built in based on the 3GPP standard, as shown in the figure (e.g., UDP/IP, TCP/IP, PHY, RLC and MAC).

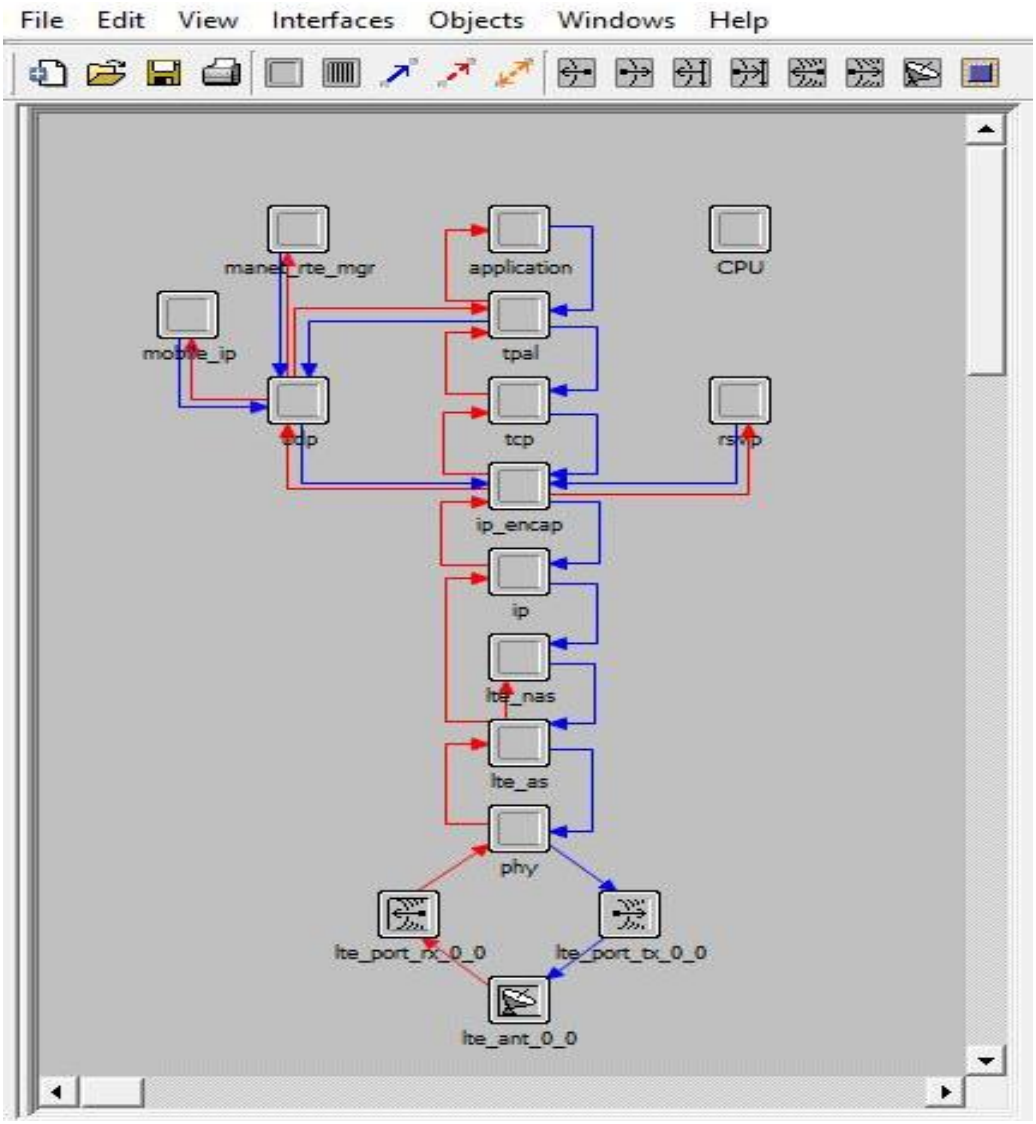


Figure 6.2: the mobile node model.

This node model has particular attributes that can be independently configured. One of these attributes relates to the LTE network configuration. For example, the EPC ID, the number of serving eNodeB, MAC layer specifications and speed node. These attributes facilitate a link association between the mobile node and a particular eNodeB.

6.2.1.2. eNodeB node model

The eNodeB connects the mobile node with the EPC. This means, the eNodeB includes all the necessary Radio Access Protocols and all the other wire links and above protocols to tunnel the user data plane between the mobile node and the core network (EPC). Figure 6.3 illustrate the eNodeB with all the protocols and radio interfaces of the physical layer. Each eNodeB also has individual attributes which configures the eNodeB according to a particular scenario. In this work, the MAC address has been modified depending on the proposed solutions, as described earlier in chapter 3.

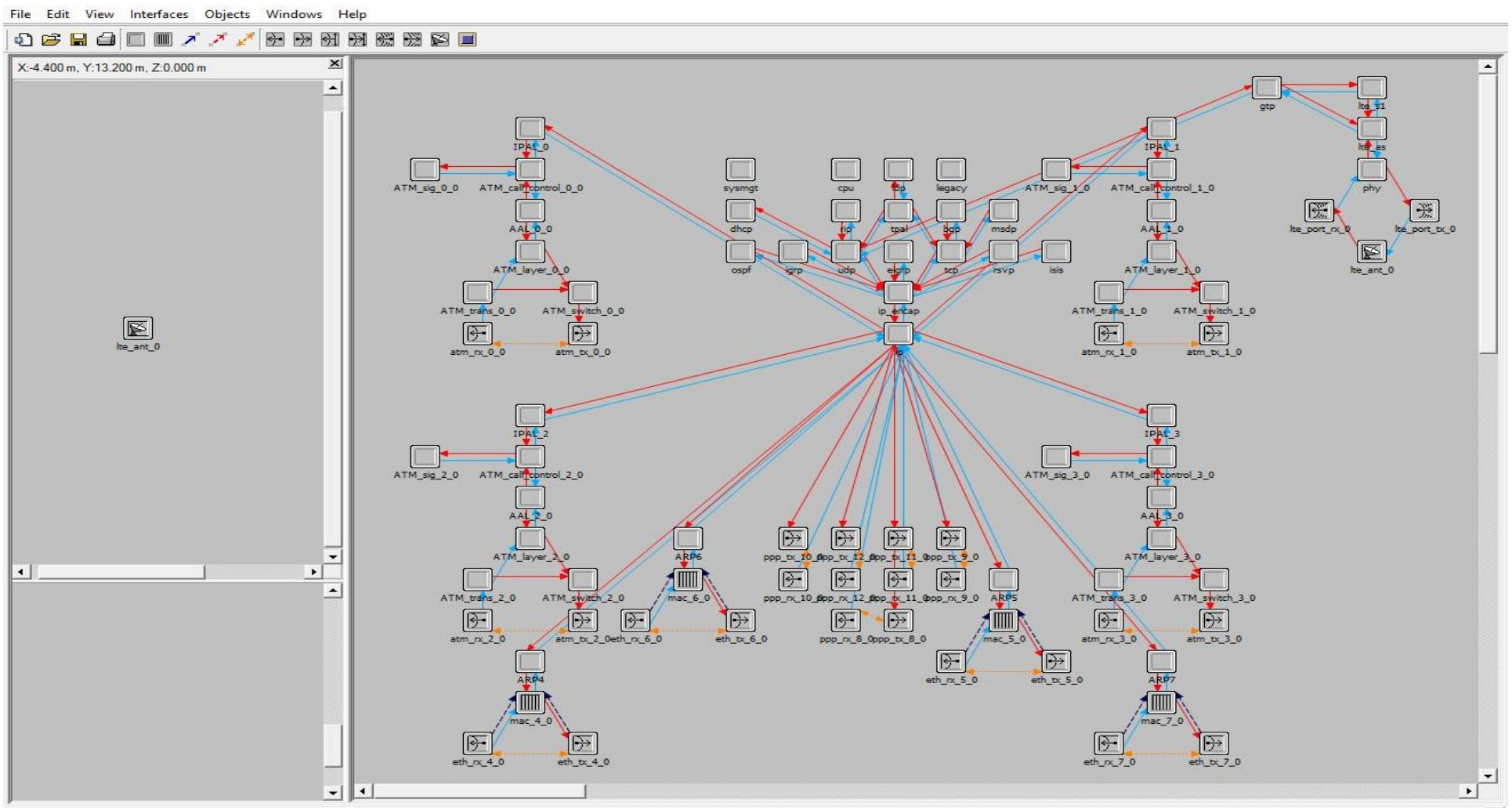


Figure 6.3: the eNodeB node model.

6.2.1.3. Slicer node model

The Slicer node model is presented in figure 6.4. This node collaborates with the eNodeB to create a virtual layer of resources allocation. This layer enables sharing of network resources between virtual networks (slices). Each slice has ability to manage its users individually.

The Slicer node has a set of attributes that can configured each slice separately and share the same bandwidth. This is done by creating a set of IP-flows and the slices share them by scheduling the buffering packets of each slice (all the processes described in chapter 3).

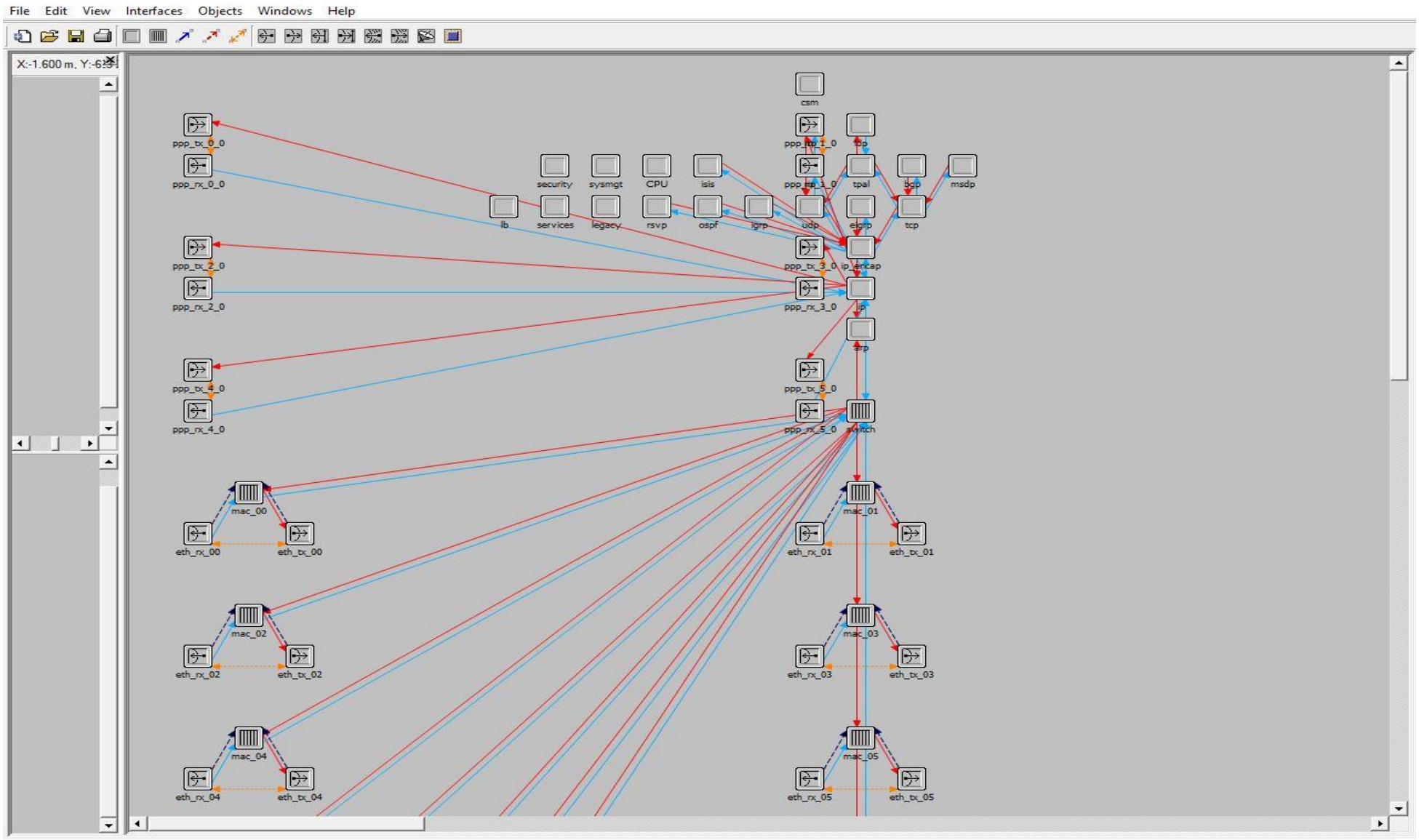


Figure 6.4: Slicer node model

6.2.1.4. Application Configuration node model

This is one of the global nodes in OPNET that it responsible for configuring the different application parameters. This node is called a global node, due to the fact that if it uses in a particular scenario to configure an application, this application will be accessible for all other nodes in the scenario. Figure 6.5 shows the node with a number of applications, such as Video Conferencing, FTP, HTTP, Voice and even custom applications.

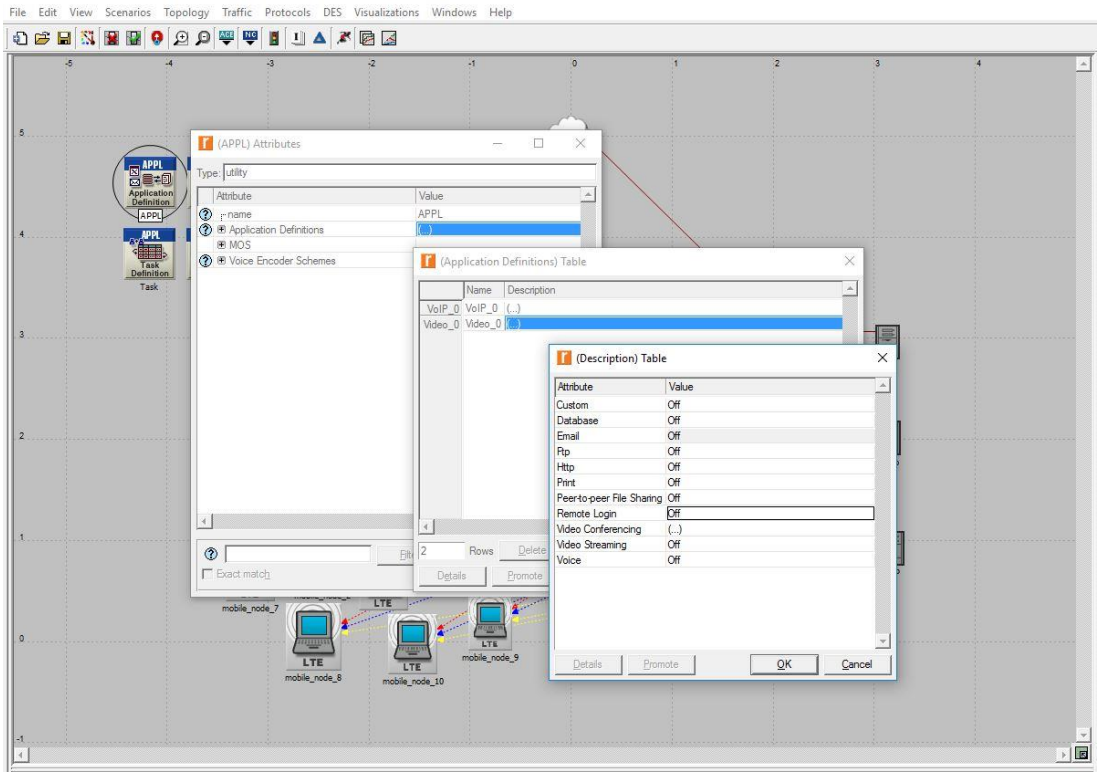


Figure 6.5: Application Configuration node model with a list of applications.

6.2.2. OMNET++ framework environment

For evaluating our proposed mechanism, the seamless connectivity for mobility management in network slicing (MMNS), we use OMNeT++ simulation to simulate the network topology which is illustrated in Figure 6.6. This topology consists of two

servers (corresponding nodes) each for a slice, one controller for integrating flows between LTE and WiFi networks, P-GW for LTE network to assign IP-Flow for each mobile node (MN), WAG for managing WiFi access network that is all the WiFi APs connecting with it and 10 MNs. Additionally, we have two network slices and each has 5 MNs. The network topology runs based on IPv6, therefore, we follow the 3GPP specification to implement the mobile IPv6. In the next subsections we give a brief explanation about some simulation entities that help us to evaluate the performance of the proposed solution.

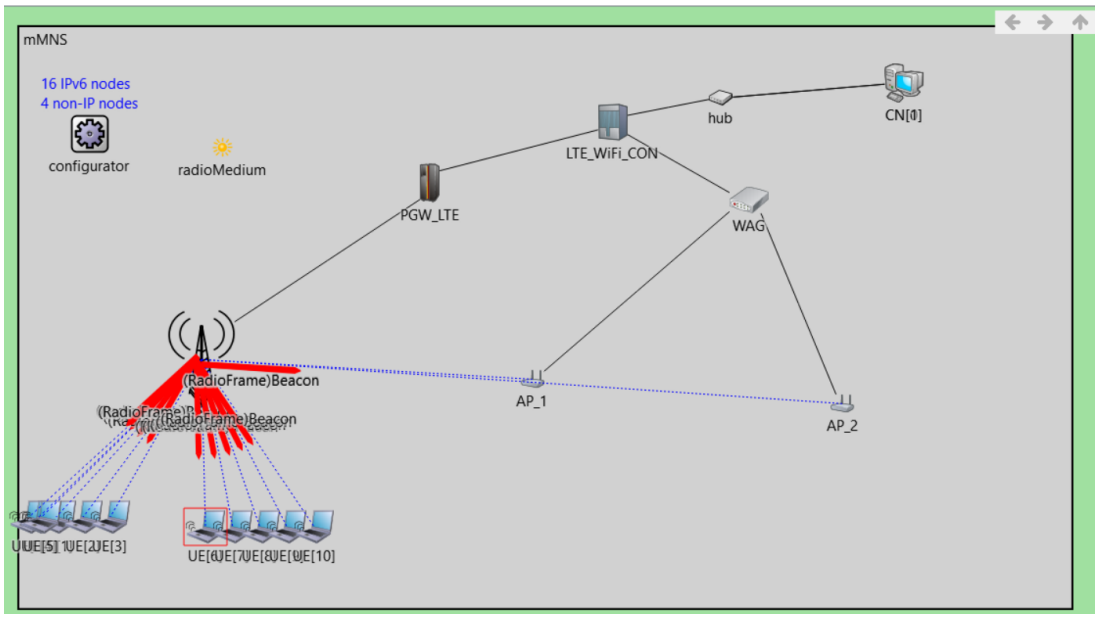
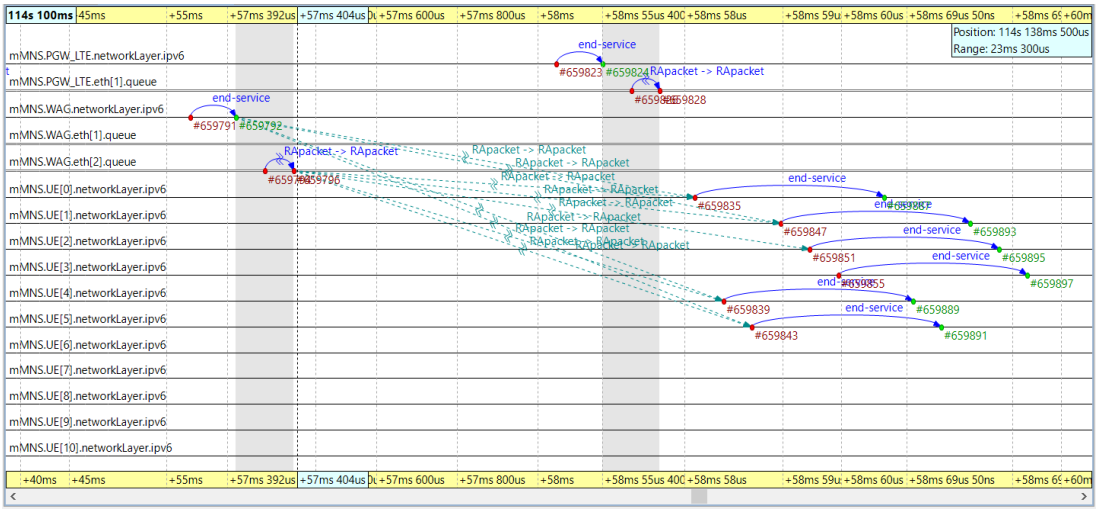


Figure 6.6: The network topology

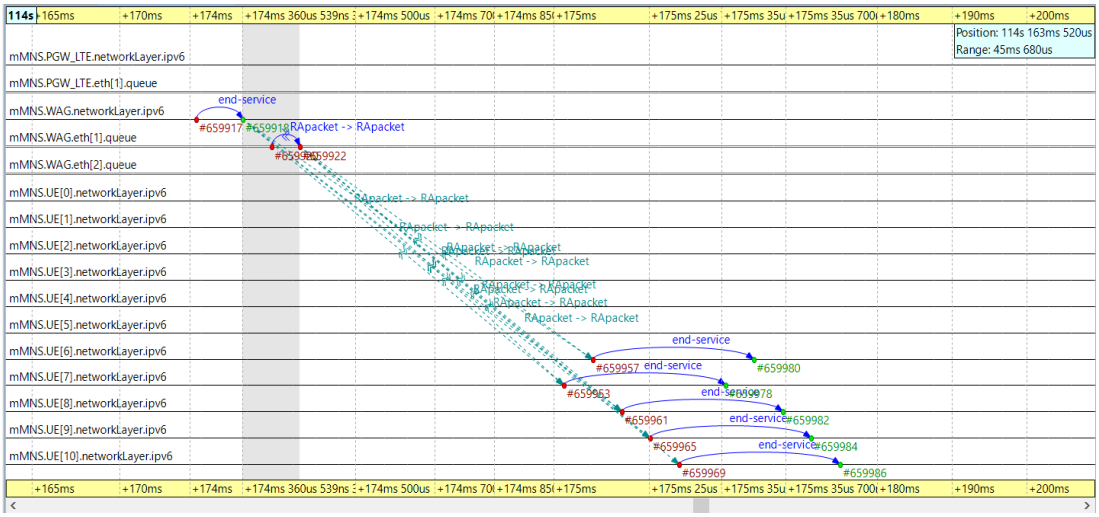
6.2.2.1. Visualizing Behaviour of Model

OMNET++ helps the user to understand interactions between modules by recording all sequence events to a file, so that, it enables the user to visualize a particular model and follow it sequences interaction during the simulation time. The OMNET++ design has a sequence chart diagram tool that provides a clear view about how the events follow each other. Moreover, by utilizing this facility it can be focused on selected or

all modules at the same time. Figure 7 illustrates the snapshot of tunnelling IPv6 for all the users in both slices. For the slice 1 users, when they move from LTE network to the WiFi network the IP sessions also move and tunnelling, Figure 6.7 (a) shows the IPv6 sequence for the users in slice 1. In the same manner, figure 6.7 (b) presents the IPv6 Sequence mobility for the users belonging to slice 2.



(a) Slice 1



(b) Slice 2

Figure 6.7: Screenshot of a sequence chart of IPv6 tunnelling for users in slice 1 and slice 2.

6.2.2.2. eNodeB Model

Figure 6.8 shows the eNodeB node with different models that collaborate together to implement the functions of eNodeB. The eNodeB has many interfaces to connect with surrounding devices or even to the internet. For example, the PPP interface to connect it to the internet, the X2 interface to connect the eNodeB to other eNodeBs and NIC interface to connect the mobile device to the eNodeB. Notice that, both devices (mobile node and eNodeB) have the Network Interface Card (NIC) interface, which holds the LTE protocol stack. When the mobile node and eNodeB try to connect each other, they implement protocol stack inside the NIC and the stack layers implement as a sub-module per-layer (e.g., Radio Link Control (RLC), Packet Data Convergence Protocol (PDCP), MAC and PHY).

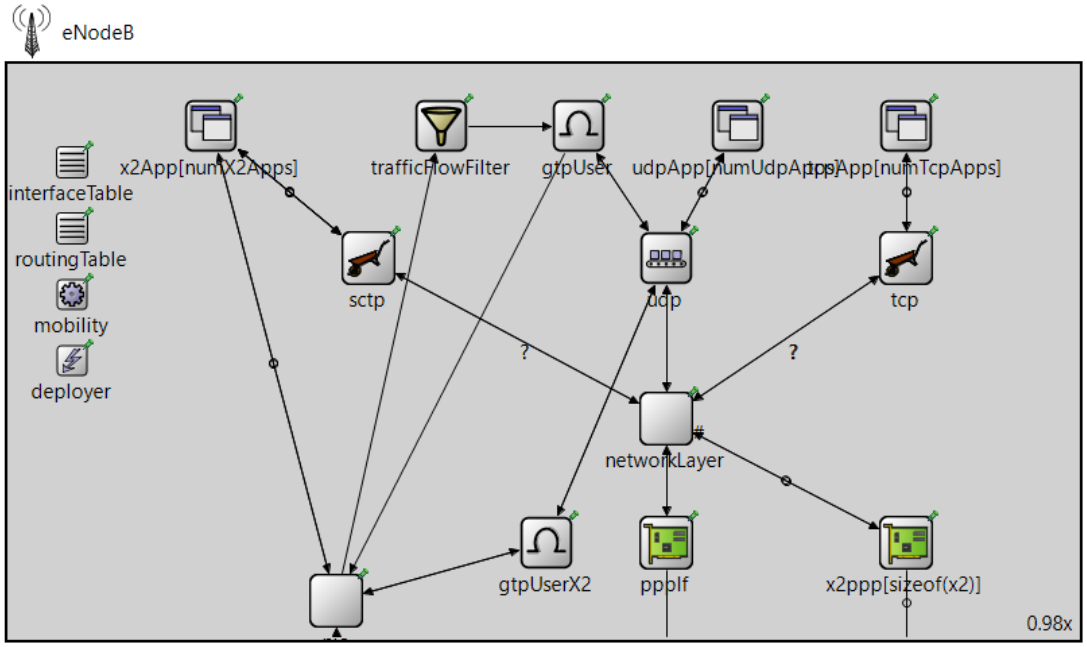


Figure 6.8: the different models of eNodeB node in OMNET++.

6.2.2.3. Controller (LTE_WiFi_CON)

The controller model can be seen in Figure 6.9. The controller includes the OFA_Controller application module that perceives the controller functionality. Moreover, it contains the TCP/IP stack modules that enable the control plane part to handle all the control messages that needed to guild the packets in the user plane.

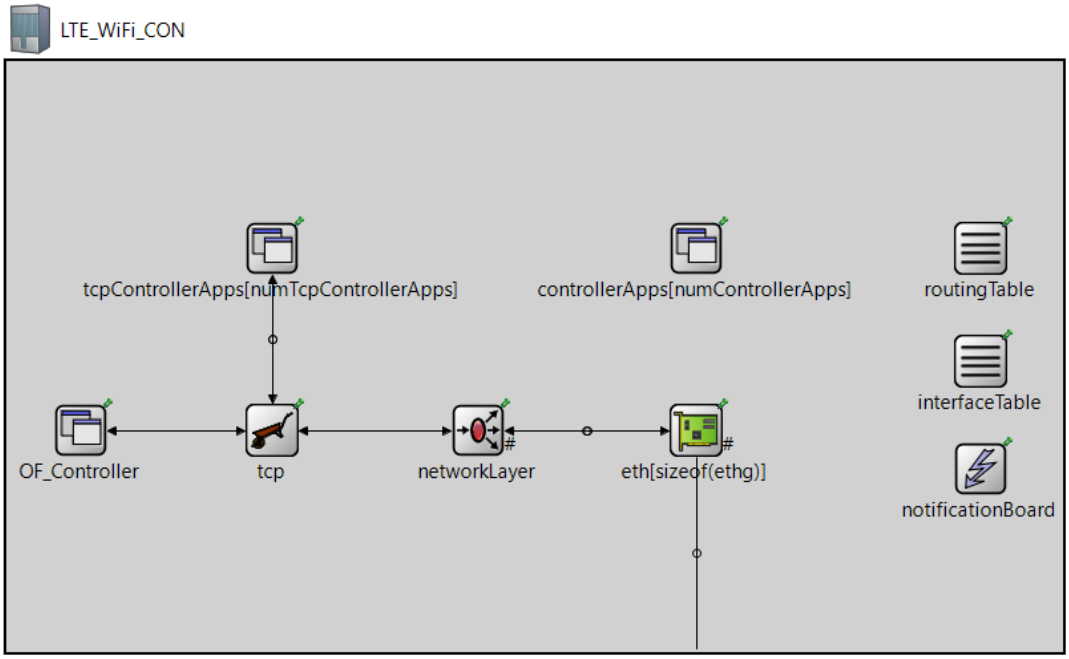


Figure 6.9: the controller node and the models

6.3. Results evaluation

In this section, we describe the results evaluation of resource management (NSRM) and the mobility management (MMNS).

6.3.1. Results evaluation for NSRM

This section is describing the performance evaluation of NSRM utilizing OPNET simulation to implement the network topology in figure 6.1. The simulation

configuration parameters are shown in Table 6.1. The next sub-sections present different scenarios to validate the proposed solution and explain the significance of our results.

Table 6.1: Simulation parameters

Parameter Name	Value
Simulation run time	720 (in seconds)
Network Slicing	2 network slices
Estimate α	0.5
Slicer resolution	1 second
Mobility model	Random Way Point (RWP), Users are initially distributed uniformly in a cell
Channel Model	Path loss: 128.1 + 37.6 log 10 (R), R in km [108]. Slow fading: Correlated Log normal, zero mean, 8db std. and 50 m correlation distance. Fast fading: Jake's like model.
Users speed	5 km/h
Total Number of PRBs	99 (corresponds to about ~ 20 MHz)
CQI reporting	Ideal
Modulation schemes	QPSK, 16 QAM, 64 QAM
eNodeB coverage area	Circular with one cell, R = 300 meters
Link-2-System interface	Effective Exponential SINR mapping [109].
FTP traffic model	File size: constant 3Mbyte Inter-arrival time: exponential (20s)
Video traffic model	24 Frames/sec, frame size: 1562 bytes (300 kbps)
VoIP Traffic model	Encoder Scheme: G. 711 (64 kbps). Talk period / Silence period: exponential (3s).

The scenarios in subsection 6.3.1.1 are presented in order to demonstrate that NSRM ensures effective bandwidth reservation for coexisting slices. Through this scenario, we will highlight the effectiveness of the proposed exponential smooth model that takes into account predefined agreements in measuring allocated bandwidth. On the other hand, in subsections 6.3.1.2 and 6.3.1.3, we consider our simulation scenarios in order to present the importance of our proposed Max-Min model under NSRM. In particular, in these subsections, we delineate how our proposal can successfully manage users' flow isolation and customization that belong to the same slice.

6.3.1.1. Bandwidth Reservation

This subsection presents different scenarios of bandwidth reservation based on predefined contracts of slices with an InP as follows:

For the fixed guaranteed bandwidth contract, we consider a video traffic model. In this scenario, we assume that the downlink (DL) of an eNodeB provides 30 PRBs (a fixed guaranteed user data rate).

Figure 6.10 shows the average user throughput under a legacy LTE network and the proposed NSRM. As depicted in the figure, both networks show approximately the same per user throughput performance. This happens because in the case of fixed guaranteed bandwidth both solutions follow the same mechanism, as we mentioned before in section 3.3.2. Unsurprisingly, due to the same reason, both of the solutions present similar average end-to-end delay performance (see Figure 6.11).

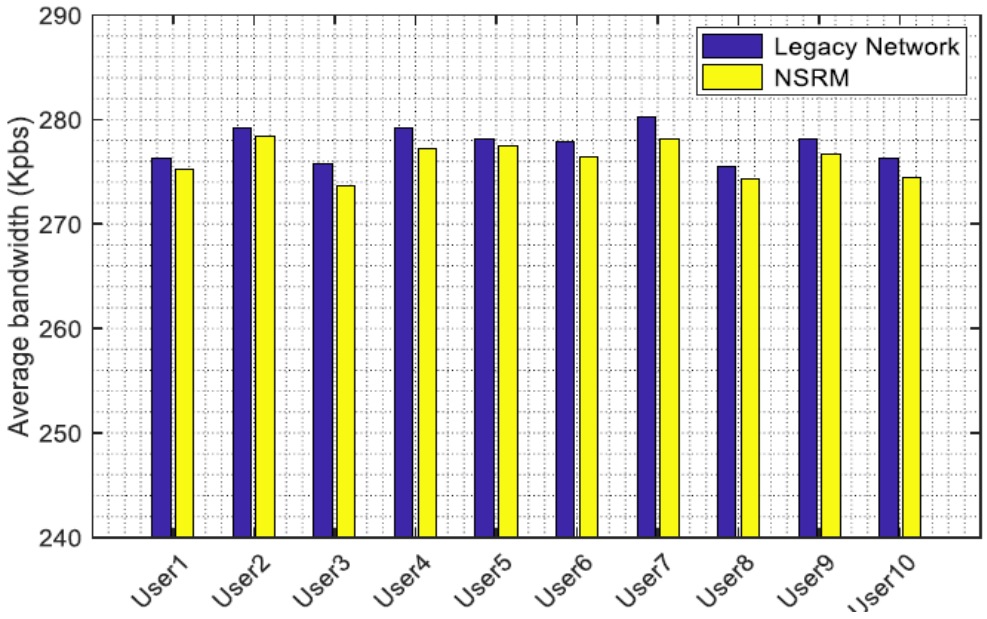


Figure 6.10: DL fixed guaranteed average per user throughput.

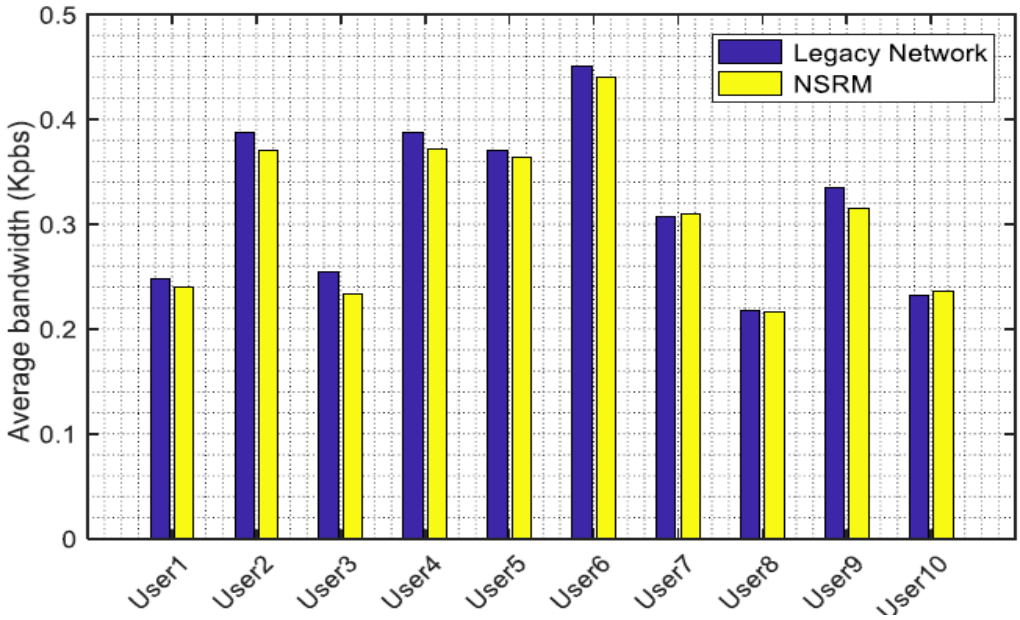


Figure 6.11: The DL average per user application end-to-end delay.

The next scenario is based on a dynamic guaranteed bandwidth contract with the VoIP traffic model application. In this scenario, the DL user data rate dynamically changes based on users' requirement and the maximum guaranteed boundary of resource reservation is 30 PRBs. Figure 6.12 demonstrates the throughput performance

comparison between these two solutions. The result shows that the average throughput per user in both networks is similar. The reason for this is that under both solutions the bandwidth reservation is guaranteed even with dynamic changes of user throughput. This scenario proves that the NSRM solution is able to dynamically reserve PRBs of a slice according to users' requirements.

At this point, we are interested to observe how the proposed solution can contribute to the increment in the utilization of radio resources. Figure 6.13 demonstrates resource blocking performance comparison between the two solutions. The results depicted in this figure confirm that in NSRM resource blocking is approximately 35% less compared to the legacy LTE network. The rationale for this result is that unlike the legacy LTE (see LTE bandwidth allocation mechanism in section 2.1.2), our proposed NSRM allocates bandwidth based on slice requirement, resulting in increasing utilization of PRBs (i.e., there would not be any unused PRBs). Consequently, in NSRM, the resource blocking will be less compared to the legacy LTE network.

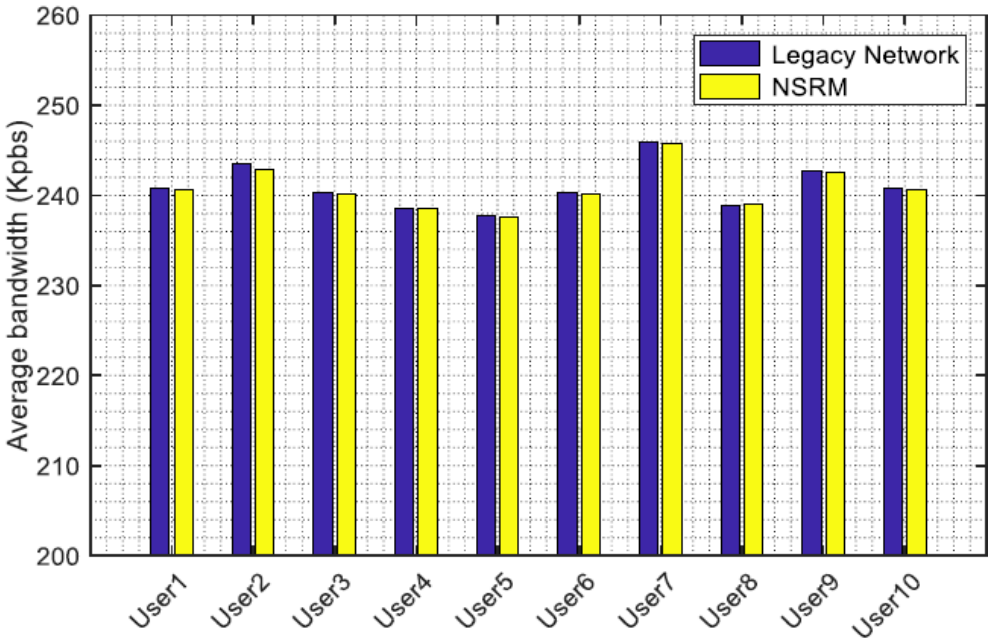


Figure 6.12: The DL dynamic guaranteed throughput average per user.

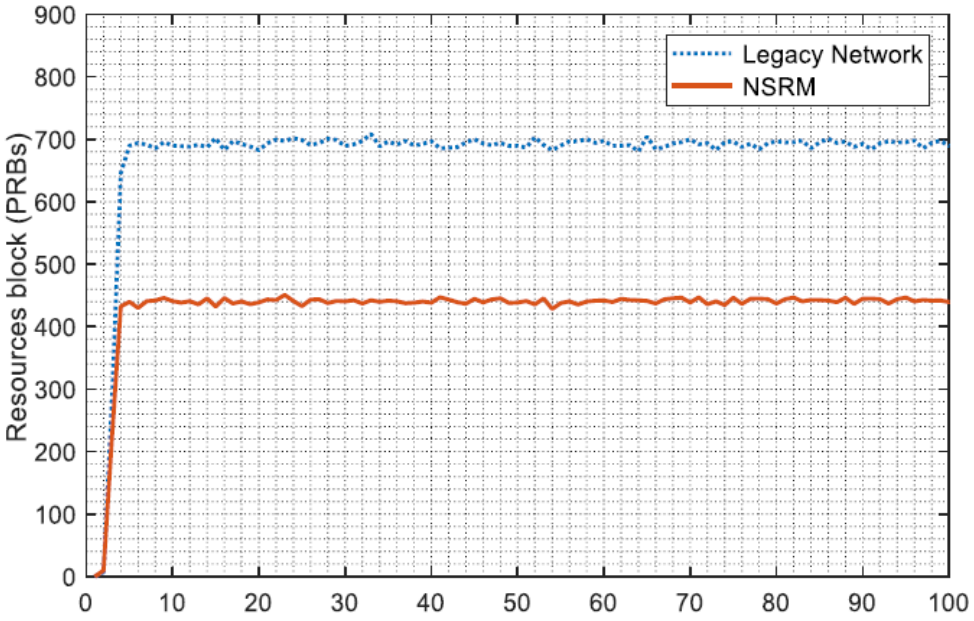


Figure 6.13: Bandwidth reservation in both scenarios.

In addition, we are interested to observe the importance of the proposed solution when a network has best effort traffic. In our simulation, in this case, we consider three types of traffic: best effort, guarantee bandwidth and dynamic guarantee bandwidth. Traffic of VoIP and video application services is considered as best effort in our simulation. Both of these applications have minimum and maximum guaranteed data rates of 30 PRBs and 50 PRBs, respectively.

Figure 6.14 shows the average bandwidth of VoIP service per user in a legacy network and NSRM solution. In this figure, we can note that both networks have the same performance per user bandwidth. Note that both networks assign the remaining PRBs to the best effort applications after satisfying resource demand of the guaranteed bandwidth applications. In the case of VoIP traffic, both solutions can meet the bandwidth requirement. Consequently, their performance for VoIP service is the same. However, the results for average bandwidth allocation for a video service depicted in Figure 6.15, show that the NSRM outperforms the legacy LTE network. It needs to

highlight that VoIP traffic is given more priority than the video traffic in an LTE network [81]. Therefore, after meeting the VoIP traffic bandwidth requirement, the legacy network allocates the residual bandwidth to the video services. The NSRM does the same; however, the amount of residual bandwidth in NSRM is larger than a legacy LTE network due to applying the dynamic bandwidth allocation mechanism. Consequently, in NSRM, a user gets more bandwidth compared to a user in a legacy LTE network (a user approximately gets 15 Kbps more bandwidth in NSRM).

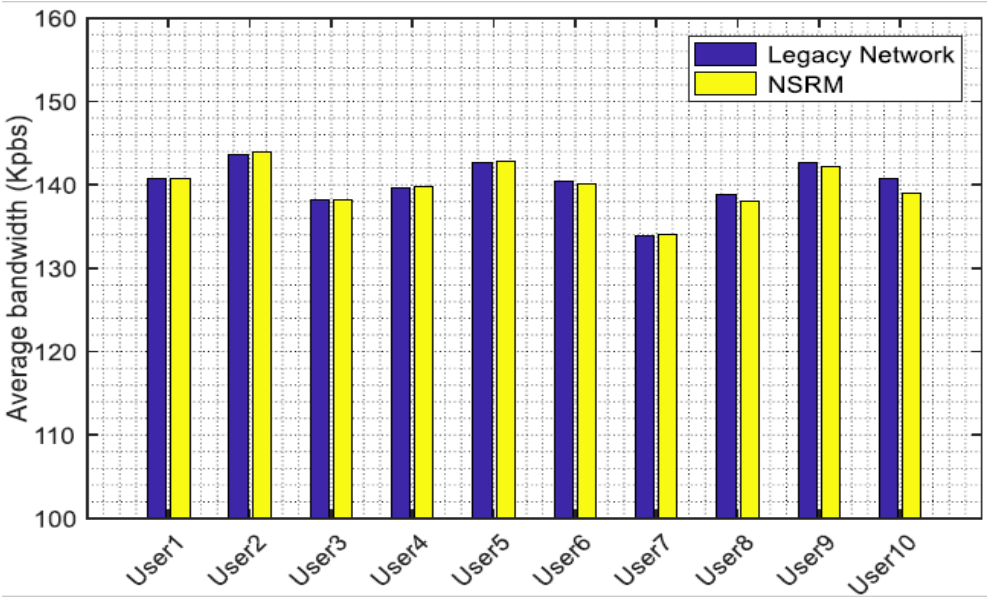


Figure 6.14: DL best effort average bandwidth of VoIP service per user.

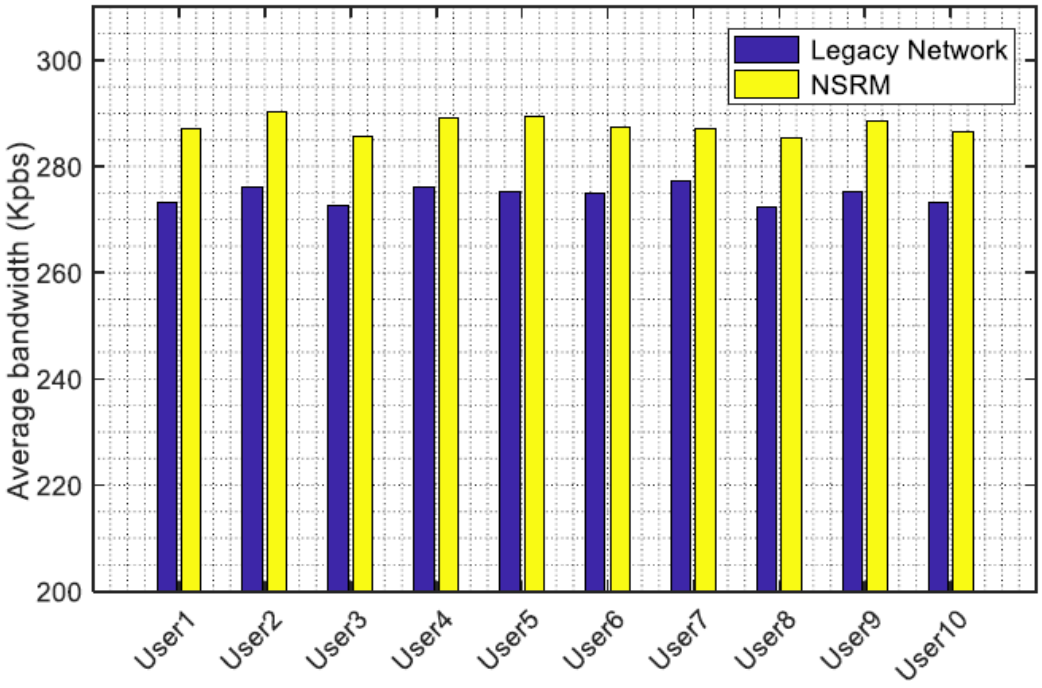
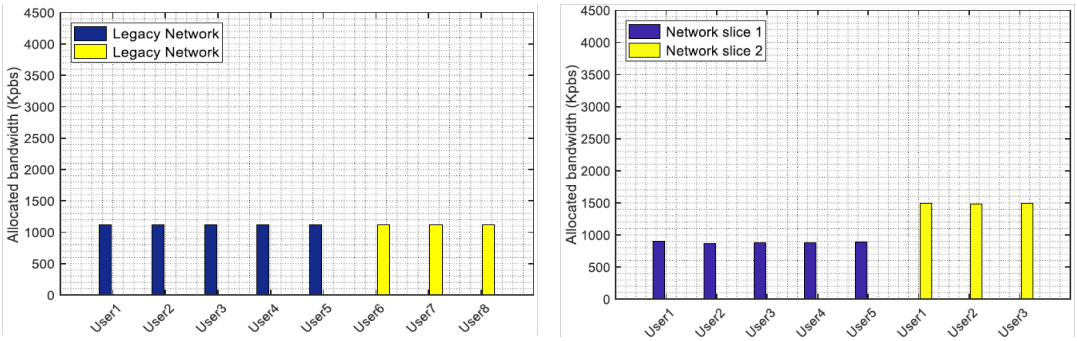


Figure 6.15: DL best effort average bandwidth of Video service per user.

6.3.1.2. Evaluation of Isolation Model

In this section, we demonstrate how our solution can successfully maintain the isolation for both inter slices (among the slices) and intra slice (among the users belong to the same slice). Under the same scenario, we compare NSRM's results in front of a legacy network. In this simulation scenario, we consider FTP traffic flows. Here, we consider two groups of users. First group (slice 1) and the second group (slice 2) have 5 users and 3 users, respectively. All the users in our simulation are located at equal distance from an eNodeB, which applies 64 QAM for Modulation and Coding Schemes (MCS). Furthermore, we assume as an aggregation, bandwidth requirement is 9 Mbps and each of the slices needs 4.5 Mbps. Additionally, in this performance evaluation, we assume all the users in a slice have the same bandwidth requirement. Simulation results are presented in Figures 6.16 (a) and (b) for a legacy network and NSRM, respectively.

Looking at Figure 6.16 (b), we observe that NSRM successfully isolates resources between the two slices. That is, NSRM provides both of the slices with an equal amount of bandwidth (each slice gets 4.5 Mbps). From the same figure, we can also realize that, under each slice all the users are provided with almost the same amount of bandwidth. These results clearly highlight that NSRM can successfully isolate not only the inter slice bandwidth but also it can isolate users' bandwidth within a slice (e.g., in the case of slice 1, each of the five users gets around 0.9 Mbps).



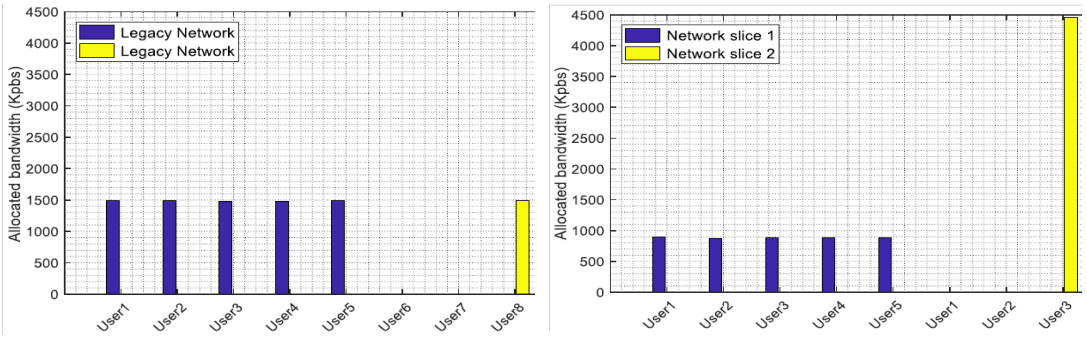
(a) Legacy network (b)NSRM

Figure 6.16: Bandwidth isolation performance evaluation.

The next simulation scenario we consider aims at illustrating how our proposed NSRM can dynamically reallocate bandwidth and successfully isolate resources with the change of network condition. We narrate the scenario as follows. In this case, our assumptions are the same as the previous scenario. Further, in this simulation, we consider, initially, each of the 8 users connected with an eNodeB is allocated 1.125 Mbps (i.e., the eNodeB provides total 9 Mbps to these users). After 200s from the simulation starting time, two users (users 6 and 7 in Legacy LTE, and 1 and 2 of Slice 2 in NSRM) turn off their mobile, releasing around 2.25 Mbps bandwidth in each scenario. In the case of Legacy LTE, the scheduler will redistribute the released

bandwidth equally to the remaining users. However, for the NSRM, the slice controller (scheduler) of the slice will reallocate the released resources of the slice and distribute them to the users according to their current requirements. The simulation results from this scenario are presented in Figure 6.17 (a) and 14 (b).

Considering Figure 6.17 (a), we observe that in a legacy network overall bandwidth of each user is increased by 0.375 kbps after two users left the network (see Figure 6.16 (a)). It happens because, in legacy network, the eNodeB redistributes the released bandwidth across the users equally. In the case of NSRM, as we notice from Figure 6.17 (b), the user 3 of Slice 2 is reallocated the released bandwidth (See Figure 6.16 (b)). However, the bandwidth allocated to each user in Slice 1 remains the same (i.e., the change of bandwidth allocation in Slice 2 does not influence the users of Slice 1). This result clearly proves that NSRM does not only successfully isolate resources between the slices but that it can also dynamically reallocate the resources.



(a) Legacy network (b)NSRM

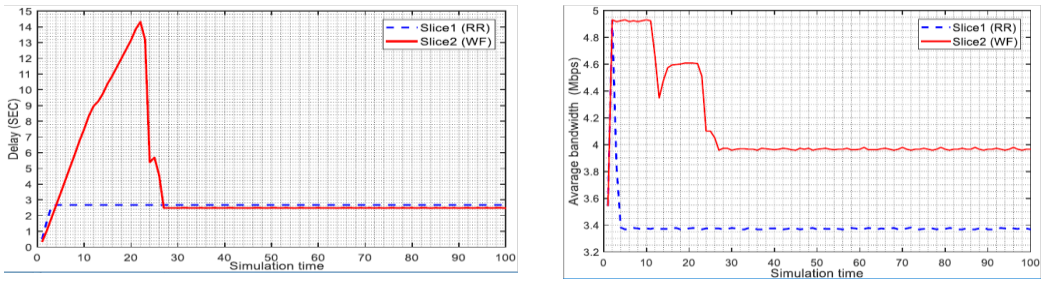
Figure 6.17: Isolation scenarios when the bandwidth increasing.

6.3.1.3. Customization

In this subsection, we want to demonstrate that in our NSRM each slice can have its own scheduling policy (i.e., different slices can have different scheduling policies).

Let us assume that Slice 1 and Slice 2 each has 4 users with heavy video traffic flows. In this simulation scenario, we consider that the Slice 1 uses a Priority Round Robin (PRR) and the Slice 2 applies Weighted Fair (WF) scheduling policy. Moreover, we suppose that all users in both slices have the same configuration setup (see video traffic model in Table 2). Simulation results are presented in Figure 6.18 (a) and (b) for traffic delay and DL traffic received.

Figure 6.18 (a) shows the delay performance for each slice in NSRM. From this figure, we can realize that despite having the same number of users with the same configuration in both slices, their delay performances are not identical. In fact, this result is quite obvious. As these two slices have two different scheduling policies, their delay performance is not the same. And for certain reasons, they have different downlink throughput performances, see Figure 6.18 (b). Therefore, these findings delineate that the proposed NSRM can allow dispensing different scheduling policies for each of the slices in an eNodeB. Note that the explanation of the performance of these two scheduling polices is not in the scope of this thesis.



(a) Traffic delay (b) DL traffic received

Figure 6.18: Flow schedulers' performance of different slices in NSRM.

6.3.2. Results Evaluation for MMNS

In this section, we present a comprehensive performance evaluation of our architecture for mobility management called MMNS (mobility management in network slicing). We compare our mechanism against two well-known mobility management approaches, namely PMIPv6 [110] and HMIPv6 [111]. As we mention earlier, for evaluating our proposed mechanism (MMNS), we use OMNeT++ simulation to simulate the network topology in the figure 6.6. In this simulation topology, we apply heavy video traffic service, which is composed of 24 frames/second and the frame size is 1562 bytes (300 Kbps).

In order to simulate the mobility management of MMNS between LTE and WiFi we use IPv6 to tunnel the IP-Flow of the MN when it moves between LTE and WiFi and vice versa.

6.3.2.1. Handover latency

We measured the handover latency in our simulation through the comparison between three different mobility mechanisms, as shown in Figure 6.19. In this figure, there are three handovers during the simulation time. The first handover from LTE eNodeB to WiFi AP and the second and third handovers between the WiFi APs, because this first handover has the highest latency. Looking at the figure, we observed that the HMIPv6 has the highest handover latency between the all three handover events at 2.208s and 1.537s in the first and second handovers, respectively. This is because HMIPv6 mobility management is Host-based, which it means that the Mobile Node (MN) involves the most mobility events resulting high latency such as binding update to the mobility anchor point, the on-link care-of-address creation and the wireless media

access. On the other hand, PMIPv6 and MMNS have low latencies compare to HMIPv6 mainly because both of them are Network-based mobility management schemes, which it means that all the mobility events that are mentioned above are controlled by a network without the need for MN interaction. Additionally, we note that the latency in PMIPv6 is notably higher than the latency in MMNS. This is due to the fact that when the MN moves between different AP, the signalling messages exchanged between MAG and LMA are needed in order to update the binding table for a new LMA of the MN. Whereas, these messages are not required in MMNS because all the binding updates are done in the LTE-WiFi-controller.

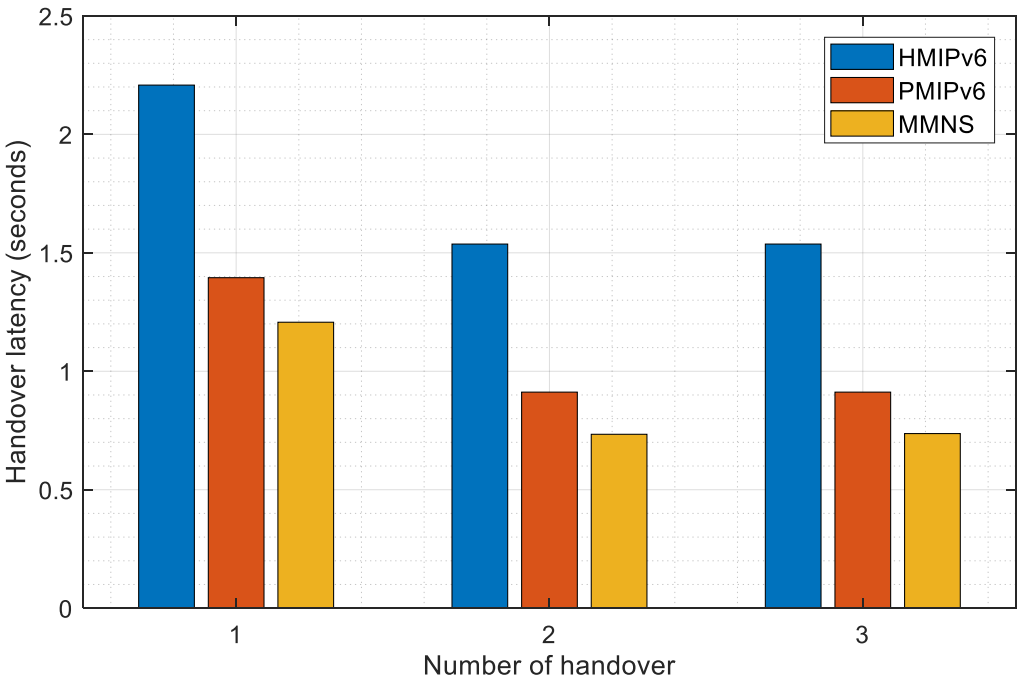


Figure 6.19: Handover latency evaluation for HMIPv6, PMIPv6 and MMNS.

Figure 6.20 illustrates the performance of throughput in MMNS for each slice during the mobility of MNs. From this figure, we notice that there are three handover occurring and the first one has lower throughput than the others. This is because the handover between two different Radio Access Technologies (RATs) (i.e., LTE and

WiFi) needs the messages overhead to update the binding addresses of each MN during the mobility. On the other hand, the other two handovers happened in the WiFi coverage areas and the throughput was slightly better compared to the first handover. This is due to the fact that the WAG of WiFi network will work as a MGA for all APs, the signalling messages exchanged between MAG and LMA to update the binding table are not needed because the IP address of MN does not change in MAG, it changes just in LMA with a new AP of the MN.

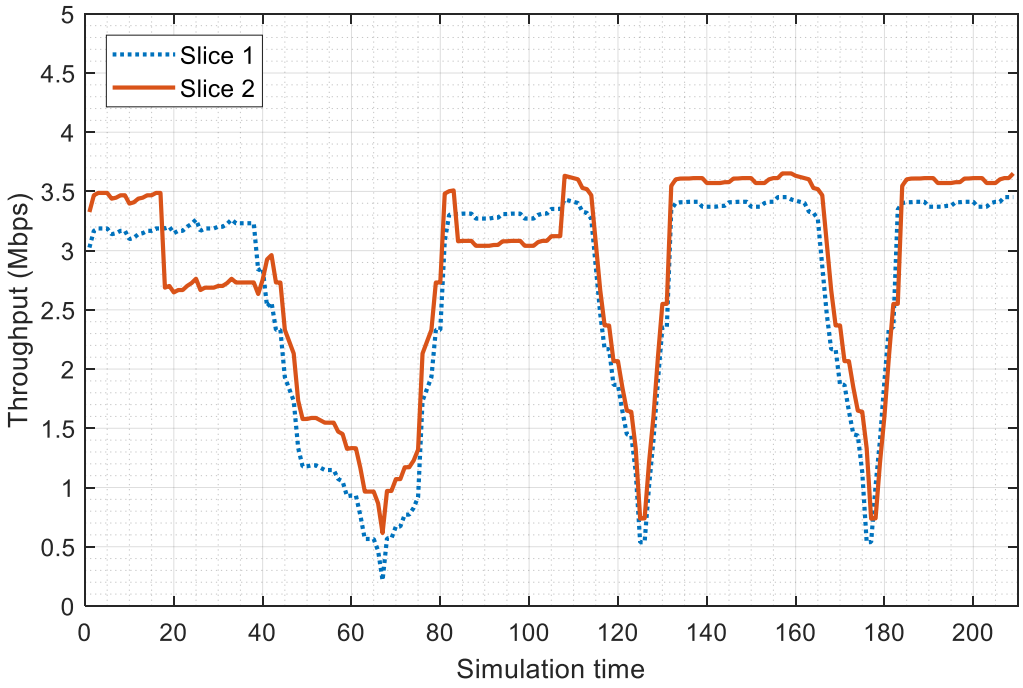


Figure 6.20: Throughput of each slice during the handover.

6.3.2.2. Traffic overhead (packet delivery cost)

In this section, we describe the total traffic delivery packets between the Corresponding Node (CN) to the MN. The total traffic overhead is depending on the number of hops between the CN and MN in respect of mobility and the packet size. Figure 6.21 illustrates the comparison of traffic delivery with respect to different flow rates for the three mobility management schemes (MMNS, PMIPv6 and HMIPv6).

However, from the figure, we can see that the MMNS has lower packets delivery cost than the other mobility management schemes. This is due to the fact that during the mobility of MN the selection of mobility anchor in MMNS is based on selecting the nearest LMA for each MN's flow. As a result, this reduces the number of hops of packets between the MN and CN.

We notice that the PMIPv6 has the highest signalling overhead. This is because of the binding table updates and the acknowledge messages exchanged between the previous MAG of AP and the LMA for the current AP. Moreover, due to the mobility the MAG will be considered out of the mobility domain, which required extra messages exchanged to find the optimal route between the MN and CN. In case of HMIPv6, we notice that it outperforms of the PMIPv6 because it allocates a set of LMAs close to the MN, resulting in route optimization due to the ability to select a shorter route. Moreover, due to the global address of MN remaining unchanged, there is no signalling required between the MN and CN.

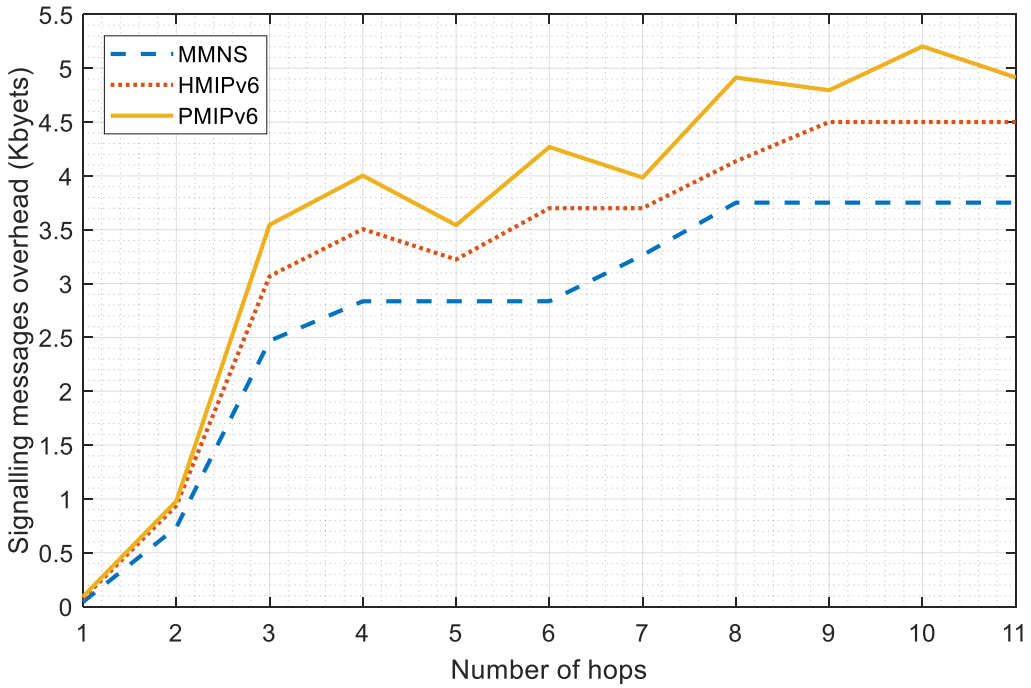


Figure 6.21: Traffic signalling overhead for HMIPv6, PMIPv6 and MMNS.

6.3.2.3. Seamless Session Continues

In this section, we show the link continues of an MN under certain slice control, during a mobility management for the MN movement. As mentioned earlier, when any MN is assigned to a certain slice it receives a list of names (IDs), which represents the slice IDs use as a SSIDs when the MN moves to the WiFi coverage area during the handover.

Figure 6.22 shows the scenario for two different slices, the first slice (slice 1) has sharing resources in all WiFi APs, which means it has SSIDs with all APs. While, in the second slice (slice 2), it also has sharing resources with all WiFi APs except one AP, meaning that there is no SSID for the slice 2 with this AP. This resulting, there is no service for the slice 2 when it's MNs move across this AP. In the figure we observe that there is no throughput for the MNs of slice 2 in the second handover, due to no SSID for slice 2 with the AP. Notice that, the throughput packets for the MNs of slice

in the third handover as shown in the figure. However, the seamless links session for the MNs of slice 1 are continuous across all three handovers.

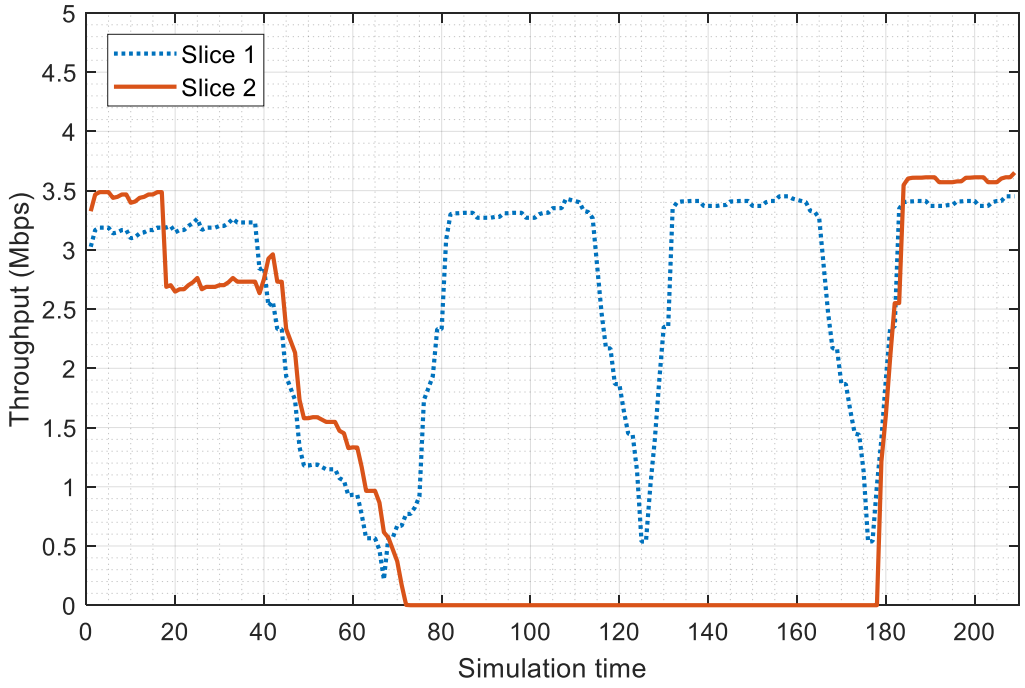


Figure 6.22: the seamless links session during a mobility of MNs under slices controls (slice 1 and slice 2).

6.4. Chapter Summary

In this chapter, we have presented the software simulations that used for evaluating the proposed contributions of this thesis; where we have conducted OPNET Modeler and OMNET++ simulations. Initially, the brief introduction about these simulations have been discussed including some classes (object-oriented programming) and sub-routines that used in this work. The OPNET Modeler was used to run the network topology for evaluating the NSRM architecture and the OMNET++ was used to evaluate the proposed MMNS architecture. Then, the results of performance evaluation for both architectures were analysis and discussed separately.

Chapter 7: Conclusions and Future works

Generally, this chapter summaries the contributions of the thesis and discusses the work to be carried out in the future. In the section 7.1, we will summarize the solutions described in this thesis, then we point out different future study directions and possible extensions of our work in section 7.2.

7.1. Conclusions

In this thesis, we introduced a resource management mechanism for network slicing as well as mobility management in different RATs utilizing the concept of network slicing to facilitate a user flows seamless continuity session in heterogeneous point's attachment. In fact, we applied the proposed solution mechanism of resources management in LTE network and addressed different emerging issues for resources provisioning in network slicing including resource allocation, resource isolation and resource customization. On the other hand, the proposed solution of mobility management is targeting the heterogeneous environment of 5G and future networks. However, because of the difficulty of simulating these new networks, we demonstrated the solution between LTE and WiFi as different access networks for a user movement. The key contributions of the thesis are listed below:

- A network slicing mechanism for resource allocation in LTE networks has been presented, where we utilizing the exponential smoothing model and Max-Min fairness mechanism. The simple exponential smoothing model is responsible for managing the inter resource allocation between slices, where it takes into consideration the estimated bandwidth that each slice needs periodically. In

contrast, the Max-Min fairness mechanism is managing the inter resource allocation for a certain slice, where it responsible of guarantee for isolating and fair sharing of a distributed bandwidth between users. Our simulation results show that the proposed mechanism satisfied the user service requirements and that it can implement different customized flow traffic for different isolated slices simultaneously.

- A logical mobility management architecture presented for network slicing based future 5G system. The control mechanisms have been discussed to unified resources of different RATs through the logical abstraction platform. Based on the modular approach, we have shown how each network slice is linked with the module, which is responsible for the mobility management of the slice. Moreover, we have introduced different use case scenarios of data offloading in cellular networks.

- Handover functionalities have been explained, where a mobile device could join and leave between different access network controllers. The proposed solution enables the selection of an appropriate AP with the cooperation of mobile devices and controllers to maximize network performance and satisfying users' demands in dynamically changing network conditions.

7.2. Future works

Based on the work introduced in this thesis, we suggest several interesting points for extending the current work and targeting future research trends as follows:

- We are aiming at investigating how network slicing can be actualized in order to share resources from different heterogeneous access networks (develop a unified

network slicing platform). In this unified network slicing platform, among the important resources issues, we are planning to study are: (i) QoS aware mobility management, and (ii) energy efficient dynamic network slice selection for user devices.

- Several improvements can be made to enhance isolation across flows belonging to different slices for the same user by modifying the client LTE drivers. WiFi-APs interference should be resolved when APs are advertising on the same channel. To address these issues, it is essential to design a distribution of APs on network topology through the WiFi controller.

- Security is rising as a critical issue in network slicing due to the nature of sharing resources between different slices. Therefore, considering secure end-to-end isolation will allow the network slicing to serve different types of services at various levels of security policy requirements. Based on that, the orchestration security management mechanism needs to be designed. In particular, the infrastructure domain requires policy coordination mechanisms to handle resources isolation among different network slices [112].

References

[1] Cisco, “Cisco Visual Networking Index : Forecast and Methodology, 2016–2021,” 2017.

[2] R. Pepper, “Cisco Visual Networking Index (VNI) Global Mobile Data Traffic Forecast Update,” 2013.

[3] ONF, “Framework for SDN : Scope and Requirements,” no. June, 2015.

[4] V. 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.

[5] 5G PPP Architecture Working Group, “View on 5G Architecture,” 2016.

[6] Jingchu Liu, T. Zhao, S. Zhou, and Z. Niu, “CONCERT : A C LOUD -B ASED A RCHITECTURE FOR N EXT -G ENERATION C ELLULAR S YSTEMS,” IEEE Wirel. Commun., vol. 21, no. 6, pp. 14–22, 2014.

[7] M. Richart, J. Baliosian, J. Serrat, and J. Gorricho, “Resource Slicing in Virtual Wireless Networks : A Survey,” IEEE Trans. Netw. Serv. Manag., vol. 13, no. 3, pp. 462–476, 2016.

[8] M. I. Kamel, L. B. Le, and A. e Girard, “LTE Wireless Network Virtualization : Dynamic Slicing via Flexible Scheduling,” in IEEE 80th Vehicular Technol. Conf., 2014, pp. 0–4.

[9] R. Kokku, R. Mahindra, H. Zhang, and S. Rangarajan, “NVS : A Substrate for Virtualizing Wireless Resources in Cellular Networks,” IEEE/ACMTRANSACTIONS Netw., vol. 20, no. 5, pp. 1333–1346, 2012.

[10] M. Yang, Y. Li, L. Zeng, D. Jin, and L. Su, “Karnaugh-map Like Online Embedding Algorithm of Wireless Virtualization,” in in Proc. 15th Int. Symp, 2012, pp. 594–598.

[11] H. Yu, G. Iosifidis, J. Huang, and L. Tassiulas, “Auction-based coopetition between LTE unlicensed and Wi-Fi,” IEEE J. Sel. Areas Commun., vol. 35, no. 1, pp. 79–90, 2017.

[12] B. Divecha, A. Abraham, C. Grosan, and S. Sanyal, “Impact of node mobility on MANET routing protocols models,” JDIM, vol. 5, no. 1, pp. 19–23, 2007.

[13] T. Spyropoulos, “Performance analysis of mobility-assisted routing,” Proceedings of the 7th ACM international symposium on Mobile ad hoc networking and computing (MobiHoc’ 06). pp. 46–90, 2006.

[14] “Ofcom Infrastructure Report,” 2014.

[15] 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, 2016.

[16] P. J. Pietraski, R. Yang, K. Li, C. Wang, T. Deng, S. Kaur, E. Bala, and R. V Pragada, “Method and apparatus for enhancing cell-edge user performance and signaling radio link failure conditions via downlink cooperative component carriers.” Google Patents, 2017.

[17] M. Y. Naderi, P. Nintanavongsa, and K. R. Chowdhury, “RF-MAC: A medium access control protocol for re-chargeable sensor networks powered by wireless energy harvesting,” IEEE Trans. Wirel. Commun., vol. 13, no. 7, pp. 3926–3937, 2014.

[18] S. Sesia, M. Baker, and I. Toufik, LTE-the UMTS long term evolution: from theory to practice. John Wiley & Sons, 2011.

[19] M. J. Yang, S. Y. Lim, H. J. Park, and N. H. Park, “Solving the data overload: Device-to-device bearer control architecture for cellular data offloading,” IEEE Veh. Technol. Mag., vol. 8, no. 1, pp. 31–39, 2013.

[20] V.-G. Nguyen, T.-X. Do, and Y. Kim, “SDN and virtualization-based LTE mobile network architectures: A comprehensive survey,” Wirel. Pers. Commun., vol. 86, no. 3, pp. 1401–1438, 2016.

[21] K. Doppler, M. Rinne, C. Wijting, C. B. Ribeiro, and K. Hugl, “Device-to-device communication as an underlay to LTE-advanced networks,” IEEE Commun. Mag., vol. 47, no. 12, 2009.

[22] Ofcom, “Communications Market Report,” 2017.

[23] 3GPP TS 24.302, “Access to the 3GPP Evolved Packet Core (EPC) via non-3GPP access networks,” 2016.

[24] ETSI TS 123 402, “Architecture enhancements for non-3GPP accesses,” 2011.

[25] IMT Vision, “Framework and overall objectives of the future development of IMT for 2020 and beyond,” 2015.

[26] “Supported by the 5G Vision The 5G Infrastructure Public Private Partnership: the next generation of communication networks and services,” 2015.

[27] M. R. Palattella, M. Dohler, A. Grieco, G. Rizzo, J. Torsner, T. Engel, and L. Ladid, “Internet of things in the 5G era: Enablers, architecture, and business models,” IEEE J. Sel. Areas Commun., vol. 34, no. 3, pp. 510–527, 2016.

[28] J. Ordonez-Lucena, P. Ameigeiras, D. Lopez, J. J. Ramos-Munoz, J. Lorca, and J. Folgueira, “Network slicing for 5g with sdn/nfv: Concepts, architectures, and challenges,” IEEE Commun. Mag., vol. 55, no. 5, pp. 80–87, 2017.

[29] N. Alliance, “5G white paper,” Next Gener. Mob. networks, white Pap., 2015.

[30] N. Alliance, “Description of network slicing concept,” NGMN 5G P, vol. 1, 2016.

[31] A. Devlic, A. Hamidian, D. Liang, M. Eriksson, A. Consoli, and J. Lundstedt, “NESMO: Network slicing management and orchestration framework,” in Communications Workshops (ICC Workshops), 2017 IEEE International Conference on, 2017, pp. 1202–1208.

[32] S. A. Baset, “Open source cloud technologies,” in Proceedings of the Third ACM Symposium on Cloud Computing, 2012, p. 28.

[33] P. Barham, B. Dragovic, K. Fraser, S. Hand, T. Harris, A. Ho, R. Neugebauer, I. Pratt, and A. Warfield, “Xen and the art of virtualization,” in ACM SIGOPS operating systems review, 2003, vol. 37, no. 5, pp. 164–177.

[34] E. Haletky, VMware ESX and ESXi in the enterprise: planning deployment of virtualization servers. Pearson Education, 2011.

[35] A. J. Younge, R. Henschel, J. T. Brown, G. Von Laszewski, J. Qiu, and G. C. Fox, “Analysis of virtualization technologies for high performance computing

environments," in Cloud Computing (CLOUD), 2011 IEEE International Conference on, 2011, pp. 9–16.

[36] E. Bugnion, S. Devine, M. Rosenblum, J. Sugerman, and E. Y. Wang, "Bringing virtualization to the x86 architecture with the original vmware workstation," ACM Trans. Comput. Syst., vol. 30, no. 4, p. 12, 2012.

[37] A. Rehman, S. Alqahtani, A. Altameem, and T. Saba, "Virtual machine security challenges: case studies," Int. J. Mach. Learn. Cybern., vol. 5, no. 5, pp. 729–742, 2014.

[38] M. G. Xavier, M. V Neves, F. D. Rossi, T. C. Ferreto, T. Lange, and C. A. F. De Rose, "Performance evaluation of container-based virtualization for high performance computing environments," in Parallel, Distributed and Network-Based Processing (PDP), 2013 21st Euromicro International Conference on, 2013, pp. 233–240.

[39] J. Antony, H. J. Ganesan, M. Gangadharan, and R. Shanmugam, "Resource management for containers in a virtualized environment." Mar-2018.

[40] D. Merkel, "Docker: lightweight linux containers for consistent development and deployment," Linux J., vol. 2014, no. 239, p. 2, 2014.

[41] A. Pahlevan, J. Picorel, A. P. Zarandi, D. Rossi, M. Zapater, A. Bartolini, P. G. Del Valle, D. Atienza, L. Benini, and B. Falsafi, "Towards near-threshold server processors," in Proceedings of the 2016 Conference on Design, Automation & Test in Europe, 2016, pp. 7–12.

[42] M. G. Xavier, M. V. Neves, and C. A. F. De Rose, "A performance comparison of container-based virtualization systems for mapreduce clusters," in Parallel, Distributed and Network-Based Processing (PDP), 2014 22nd Euromicro International Conference on, 2014, pp. 299–306.

[43] 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, 2015.

[44] B. A. A. Nunes, M. Mendonca, X.-N. Nguyen, K. Obraczka, and T. Turletti, "A survey of software-defined networking: Past, present, and future of

programmable networks," IEEE Commun. Surv. Tutorials, vol. 16, no. 3, pp. 1617–1634, 2014.

[45] M. Berman, J. S. Chase, L. Landweber, A. Nakao, M. Ott, D. Raychaudhuri, R. Ricci, and I. Seskar, "GENI: A federated testbed for innovative network experiments," Comput. Networks, vol. 61, pp. 5–23, 2014.

[46] Open Networking Foundation, "SDN Architecture Overview," 2013.

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

[48] S. Palkar, C. Lan, S. Han, K. Jang, A. Panda, S. Ratnasamy, L. Rizzo, and S. Shenker, "E2: a framework for NFV applications," in Proceedings of the 25th Symposium on Operating Systems Principles, 2015, pp. 121–136.

[49] 5GPPP, "View on 5G Architecture (Version 2 . 0)," 2017.

[50] A. S. Thyagaturu, Y. Dashti, and M. Reisslein, "SDN-based smart gateways (Sm-GWs) for multi-operator small cell network management," IEEE Trans. Netw. Serv. Manag., vol. 13, no. 4, pp. 740–753, 2016.

[51] P. Rost, C. Mannweiler, D. S. Michalopoulos, C. Sartori, V. Sciancalepore, N. Sastry, O. Holland, S. Tayade, B. Han, D. Bega, and others, "Network slicing to enable scalability and flexibility in 5G mobile networks," IEEE Commun. Mag., vol. 55, no. 5, pp. 72–79, 2017.

[52] K. Katsalis, N. Nikaein, E. Schiller, A. Ksentini, and T. Braun, "Network slices toward 5G communications: Slicing the LTE network," IEEE Commun. Mag., vol. 55, no. 8, pp. 146–154, 2017.

[53] 3GPP TR 28.801, "Study on Management and Orchestration of Network Slicing for Next Generation Network," 2017.

[54] 3GPP TR 28.801, Study on Management and Orchestration of Network Slicing for Next Generation Network, Rel.15, May 2017.

[55] 5G Americas, "Network Slicing for 5G Networks and Services," 2016.

[56] W. Rankothge, F. Le, A. Russo, and J. Lobo, "Optimizing Resource Allocation

for Virtualized Network Functions in a Cloud Center Using Genetic Algorithms," IEEE Trans. Netw. Serv. Manag., vol. 14, no. 2, pp. 343–356, 2017.

[57] B. Guan, J. Wu, Y. Wang, S. U. Khan, and S. Member, "CIVSched : A Communication-Aware Inter-VM Scheduling Technique for Decreased Network Latency between Co-Located VMs," IEEE Trans. CLOUD Comput., vol. 2, no. 3, pp. 320–332, 2014.

[58] A. Testolin, M. D. E. F. D. E. Grazia, and M. Zorzi, "Cognition-Based Networks : A New Perspective on Network Optimization Using Learning and Distributed Intelligence," IEEE Access, vol. 3, pp. 1512–1530, 2015.

[59] C. Liang and F. R. Yu, "Wireless Network Virtualization: A Survey, Some Research Issues and Challenges," IEEE Commun. Surv. Tutorials, vol. 17, no. 1, pp. 358–380, 2015.

[60] M. Kalil, A. Shami, and Y. Ye, "Wireless resources virtualization in LTE systems," Proc. - IEEE INFOCOM, pp. 363–368, 2014.

[61] K. Samdanis, A. Kunz, M. I. Hossain, and T. Taleb, "Virtual bearer management for efficient MTC radio and backhaul sharing in LTE networks," in IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, PIMRC, 2013.

[62] X. Jin, L. E. Li, L. Vanbever, and J. Rexford, "SoftCell: Scalable and Flexible Cellular Core Network Architecture," in the ninth ACM conference on Emerging networking experiments and technologies, 2013, pp. 163–174.

[63] J. Van De Belt, H. Ahmadi, and L. E. Doyle, "A Dynamic Embedding Algorithm for Wireless Network Virtualization," in IEEE VTC, 2014, pp. 1–6.

[64] M. Jiang, M. Condoluci, and T. Mahmoodi, "Network slicing management & prioritization in 5G mobile systems," in European Wireless, 2016, pp. 1–6.

[65] L. J. Zhang, W. K. Chen, Y. Zhang, A. Quintero, and S. Pierre, "Seamless Mobility Management Schemes for IPv6-based Wireless Networks," Mob. Networks Appl., vol. 19, no. 6, pp. 745–757, 2014.

[66] and M. A. S. Chen, Y. Shi, B. Hu, "Mobility Management Reference Models,"

in Berlin, Heidelberg: Springer Berlin Heidelberg, 2016, pp. 29–62.

[67] J. Kempf, “Problem statement for network-based localized mobility management (NETLMM),” 2007.

[68] A. S. S. Navaz and R. Barathiraja, “Security Aspects of Mobile IP,” J. Nano Sci. Nano Technol., vol. 2, no. 3, 2014.

[69] H. Modares, A. Moravejosharieh, J. Lloret, and R. Bin Salleh, “A survey on proxy mobile IPv6 handover,” IEEE Syst. J., vol. 10, no. 1, pp. 208–217, 2016.

[70] A. Gani, G. M. Nayeem, M. Shiraz, M. Sookhak, M. Whaiduzzaman, and S. Khan, “A review on interworking and mobility techniques for seamless connectivity in mobile cloud computing,” J. Netw. Comput. Appl., vol. 43, pp. 84–102, 2014.

[71] A. B. Johnston, SIP: understanding the session initiation protocol. Artech House, 2015.

[72] D. Liu and P. Seite, “Distributed Mobility Management: Current practices and gap analysis,” 2015.

[73] Y. Bi, H. Zhou, W. Xu, X. S. Shen, and H. Zhao, “An efficient PMIPv6-based handoff scheme for urban vehicular networks,” IEEE Trans. Intell. Transp. Syst., vol. 17, no. 12, pp. 3613–3628, 2016.

[74] 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.

[75] V. Yazici, U. C. Kozat, and M. O. Sunay, “A new control plane for 5G network architecture with a case study on unified handoff, mobility, and routing management,” IEEE Commun. Mag., vol. 52, no. 11, pp. 76–85, 2014.

[76] H. Railway, H. Song, X. Fang, and L. Yan, “Handover Scheme for 5G C / U Plane Split Heterogeneous Network in,” IEEE Trans. Veh. Technol., vol. 63, no. 9, pp. 4633–4646, 2014.

[77] M. Chen, Y. Hao, M. Qiu, J. Song, D. Wu, and I. Humar, “Mobility-aware caching and computation offloading in 5G ultra-dense cellular networks,”

Sensors, vol. 16, no. 7, p. 974, 2016.

[78] X. Ge, J. Ye, Y. Yang, and Q. Li, “User mobility evaluation for 5G small cell networks based on individual mobility model,” IEEE J. Sel. Areas Commun., vol. 34, no. 3, pp. 528–541, 2016.

[79] V. Sharma, I. You, and R. Kumar, “Resource-based mobility management for video users in 5G using catalytic computing,” Comput. Commun., 2017.

[80] M. Agiwal, A. Roy, and N. Saxena, “Next Generation 5G Wireless Networks : A Comprehensive Survey,” IEEE Commun. Surv. TUTORIALS, vol. 18, no. 3, pp. 1617–1655, 2016.

[81] F. Capozzi, G. Piro, L. A. Grieco, G. Boggia, and P. Camarda, “Downlink Packet Scheduling in LTE Cellular Networks : Key Design Issues and a Survey,” IEEE Commun. Surv. TUTORIALS, vol. 15, no. 2, pp. 678–700, 2013.

[82] R. P. Jover, “LTE PHY Fundamentals,” 2015.

[83] M. Alasti, B. Neekzad, J. Hui, and R. Vannithamby, “Quality of service in WiMAX and LTE networks,” IEEE Commun. Mag., vol. 48, no. 5, pp. 104–111, 2010.

[84] B. Sadiq, R. Madan, and A. Sampath, “Downlink scheduling for multiclass traffic in LTE,” Eurasip J. Wirel. Commun. Netw., p. 18, 2009.

[85] A. S. D. Alfoudi, M. Dighriri, G. M. Lee, R. Pereira, and F. P. Tso, “Traffic Management in LTE-WiFi Slicing Networks,” in Innovations in Clouds, Internet and Networks (ICIN), 2017, pp. 268–273.

[86] P. Rost, C. Mannweiler, Di. S. Michalopoulos, C. Sartori, V. Sciancalepore, N. Sastry, O. Holland, S. Tayade, B. Han, D. Bega, D. Aziz, and H. Bakker, “Network Slicing to Enable Scalability and Flexibility in 5G Mobile Networks,” IEEE Commun. Mag., vol. 55, no. 5, pp. 72–79, 2017.

[87] J. Brown, J. Y. Khan, and S. Member, “A Predictive Resource Allocation Algorithm in the LTE Uplink for Event Based M2M Applications,” IEEE Trans. Mob. Comput., vol. 14, no. 12, pp. 2433–2446, 2015.

[88] R. Hwang, S. Member, C. Lee, and Y. Chen, “Cost Optimization of Elasticity Cloud Resource Subscription Policy,” IEEE Trans. Serv. Comput., vol. 7, no. 4, pp. 561–574, 2014.

[89] Y. Zaki, L. Zhao, C. Goerg, and A. Timm-giel, “LTE Wireless Virtualization and Spectrum Management,” in Wireless and Mobile Networking Conference (WMNC), 2010 Third Joint IFIP, 2010, pp. 1–6.

[90] T. Taleb, B. Mada, M. Corici, A. Nakao, and H. Flinck, “PERMIT : Network Slicing for Personalized 5G Mobile Telecommunications,” IEEE Commun. Mag., vol. 55, no. 5, pp. 88–93, 2017.

[91] R. Sherwood, G. Gibb, K. Yap, G. Appenzeller, M. Casado, N. Mckeown, G. Parulkar, R. Sherwood, G. Gibb, K. Yap, and G. Appenzeller, “FlowVisor : A Network Virtualization Layer,” in OpenFlow Switch Consortium, Tech. Rep, 1, p.132., 2009.

[92] P. Mogensen, W. Na, I. Z. Kovács, F. Frederiksen, A. Pokhariyal, K. I. Pedersen, T. Kolding, K. Hugl, and M. Kuusela, “LTE Capacity compared to the Shannon Bound,” in IEEE VTC, 2007, pp. 1234–1238.

[93] 4G Americas, “Bringing Network Function Virtualization to LTE,” 2014.

[94] Y. Zaki, L. Zhao, C. Görg, and A. Timm-Giel, “LTE mobile network virtualization - Exploiting multiplexing and multi-user diversity gain,” MONET, vol. 16, pp. 424–432, 2011.

[95] Cisco, “Campus LAN and Wireless LAN Design Guide,” 2015.

[96] L. Suresh, J. Schulz-Zander, R. Merz, A. Feldmann, and T. Vazao, “Towards programmable enterprise WLANS with Odin,” in Proceedings of the first workshop on Hot topics in software defined networks, 2012, pp. 115–120.

[97] V. G. Nguyen and Y. H. Kim, “Slicing the next mobile packet core network,” 2014 11th Int. Symp. Wirel. Commun. Syst. ISWCS 2014 - Proc., pp. 901–904, 2014.

[98] T. Shimojo, Y. Takano, A. Khan, S. Kaptchouang, M. Tamura, and S. Iwashina, “Future mobile core network for efficient service operation,” 1st IEEE Conf. Netw. Softwarization Software-Defined Infrastructures Networks, Clouds, IoT

Serv. NETSOFT 2015, 2015.

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

[100] 3GPP, “TR 23.799, V0.7.0, Technical Specification Group Services and System Aspects, Study on Architecture for Next Generation System,” 2016.

[101] C. Tsirakis, P. Matzoros, P. Sioutis, and G. Agapiou, “Load balancing in 5G Networks,” MATEC Web Conf., vol. 125, pp. 1–6, 2017.

[102] R. Jmal and L. Chaari Fourati, “Implementing shortest path routing mechanism using Openflow POX controller,” 2014 Int. Symp. Networks, Comput. Commun. ISNCC 2014, pp. 1–6, 2014.

[103] A. R, J. Dielissen, K. Goossens, E. Rijpkema, and P. Wielage, “An Efficient On Chip Network Interface Offering Guaranteed Services , Shared Memory Abstraction , and Flexible Network Configuration,” in Design, Automation and Test in Europe Conference and Exhibition, 2004, vol. 24, pp. 1–6.

[104] B. Martini and F. Paganelli, “A service-oriented approach for dynamic chaining of virtual network functions over multi-provider software-defined networks,” Futur. Internet, vol. 8, no. 2, pp. 1–21, 2016.

[105] F. Paganelli, M. Ulema, and B. Martini, “Context-aware service composition and delivery in NGSONs over SDN,” IEEE Commun. Mag., vol. 52, no. 8, pp. 97–105, 2014.

[106] R. Dunaytsev, Network Simulators: OPNET overview and examples. 2010.

[107] A. Varga and R. Hornig, “An overview of the OMNeT++ simulation environment,” in Proceedings of the 1st international conference on Simulation tools and techniques for communications, networks and systems & workshops, 2008, p. 60.

[108] A. B. Saleh, S. Redana, H. Jyri, and B. Raaf, “On the Coverage Extension and Capacity Enhancement of Inband Relay Deployments in LTE-Advanced Networks,” J. Electr. Comput. Eng., vol. 2010:4, 2010.

[109] R. Giuliano and F. Mazzenga, “Exponential Effective SINR Approximations

for OFDM/OFDMA-Based Cellular System Planning," IEEE Trans. Wirel. Commun., vol. 8, no. 9, pp. 4434–4439, 2009.

[110] S. M. Raza, D. S. Kim, and H. Choo, "Leveraging PMIPv6 with SDN," in Proceedings of the 8th International Conference on Ubiquitous Information Management and Communication, 2014, p. 13.

[111] S. Pack, T. Kwon, and Y. Choi, "A performance comparison of mobility anchor point selection schemes in Hierarchical Mobile IPv6 networks," Comput. Networks, vol. 51, no. 6, pp. 1630–1642, 2007.

[112] X. Li, M. Samaka, H. A. Chan, D. Bhamare, L. Gupta, C. Guo, and R. Jain, "Network slicing for 5g: Challenges and opportunities," IEEE Internet Comput., vol. 21, no. 5, pp. 20–27, 2017.