Software Defined Network based Architecture and RAN Virtualization for 5G Multi-RAT Networks

A thesis submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

by

Akshatha Nayak M.
(Roll No. 144070029)

Under the guidance of:
Prof. Abhay Karandikar

Department of Electrical Engineering
Indian Institute of Technology Bombay
Powai, Mumbai 400076
March, 2021

To my Grandparents, who fostered the love of learning

Thesis Approval

The Thesis entitled

Software Defined Network based architecture and RAN virtualization for 5G Multi-RAT Networks

by

Akshatha Nayak M.

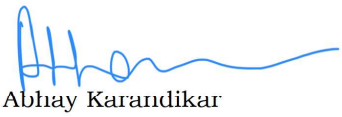
(Roll No. 144070029)

is approved for the degree of

Doctor of Philosophy

Prof. Saif Khan Mohammed
(Examiner)

Prof. S. Vijayakumaran
(Examiner)



Prof. Abhay Karandikar
(Advisor)

Prof. Rushikesh K. Joshi
(Chairperson)

Date: 13-07-2021

Place: IIT Bombay, Mumbai

Declaration

I declare that this written submission represents my ideas in my own words and where others ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Akshatha Nayak M.

Roll No. 144070029

Date: 14-07-2021

Abstract

We live in the information age where we increasingly connect to the world around us through smart devices. We are constantly utilizing services such as video streaming, banking services, social media, etc., which are provided through the mobile network. This phenomenon of growing mobile data utilization, along with a rise in the number of cellular subscriptions, has resulted in an unprecedented growth in cellular data traffic. The required increase in network capacity coupled with the support for diverse services can be provided by the next-generation Fifth Generation (5G) networks.

However, network upgrades are expensive, and the upgrade cost is not offset by the almost constant per-user revenue. To alleviate the capacity problem, network operators employ solutions such as a partial upgrade to 5G and co-located deployment of low-cost technology such as Wireless Local Area Networks (WLANs). Furthermore, they are exploring the use of network slicing to support service diversity in a more efficient way. Consequently, the next-generation networks comprise several Radio Access Technologies (RATs) operating in tandem to provide a diverse set of services. However, at present, the usage of multiple Radio Access Technologies (multi-RATs) poses significant challenges in terms of inter-working and network management due to the fragmented nature of RAT control.

In this thesis, we attempt to investigate several problems pertinent to this topic, such as the unification of multi-RAT network control, network slicing, and enhancement of User Equipment (UE) slice mobility using Software Defined Networking (SDN) and Network Function Virtualization (NFV) based approaches. The proposed solutions reflect the practical nature of the problems and endeavor to make minimal changes to the UE. In principle, this makes it easier to integrate the solution with practical deployment. At the outset, we develop a centralized architecture for 5G New Radio (NR) technology in accordance with the SDN paradigm. Further, we extend this design to incorporate net-

work slicing and include support for multiple RATs. The proposed design is scalable and provides a virtualized view of global network resources. Subsequently, we highlight the lack of a standardized framework for slicing the multi-RAT Radio Access Network (RAN). To address this, we define 'Virtualized RAN (VirtRAN)', a framework for recursively slicing multi-RAT networks. To complement this solution, we propose enhancements to UE multi-connectivity protocol for enabling slice mobility without the need for homogeneous slice deployments. As an auxiliary benefit, we are also able to provide a mechanism for RAN level inter-working between WLAN and 5G NR, which is currently undefined in the standard specification.

In summary, we identify the following multifold outcomes from this thesis. We have identified gaps in the existing standards and suggested plausible solutions to address these gaps. We have also submitted some of the proposals as contributions to relevant standards. Wherever possible, we have demonstrated performance improvements resulting from the proposed designs using network simulator-3 (ns-3) simulations. Finally, we have also indicated open challenges for further exploration.

Contents

Acknowledgments

"Ithaka gave you the marvelous journey,
Without her you wouldn't have set out."

–Constantine P. Cavafy

Graduate school at IIT Bombay has been akin to a personal journey to the fabled Ithaka. The last few years have been exciting and rewarding, thanks to the wonderful people who have supported me every step of the way. My deepest gratitude is to my advisor, Prof. Abhay Karandikar for his invaluable research inputs and guidance. He has been a constant source of inspiration and a role model through his deep technical expertise, commitment to excellence, positive outlook and integrity. I have always admired his tireless dedication and avant-garde vision in using next-generation technology research and standards to solve real world problems while contributing to the nation's progress. He has always made time for discussions with students and has mentored us irrespective of the distance or his busy schedule. I am indebted to him for his tutelage in identifying, formulating, and systematically addressing research problems.

I have also been fortunate to be mentored by Prof. Prasanna Chaporkar over the last two years. He is a brilliant researcher who has provided insightful feedback and always encouraged me to further my work. From him, I have learnt how to present my research in a lucid manner and gained an improved understanding of key concepts.

My heartfelt thanks to Mr. Pranav Jha for always being a pillar of support. His vast knowledge, articulate explanation of various technical concepts, patient critiques, and creativity in finding newer solutions have aided me immensely in evolving into a researcher. I will always treasure the learnings that I have gained through numerous discussions with him over the years. His strong work ethic, attention to detail, kindness, humility, and boundless patience have served as an example to be emulated.

I wish to express my gratitude to Prof. Sarvananan Vijayakumaran and Prof. Gaurav Kasbekar for providing consistent feedback and valuable suggestions as part of the Research Progress Committee. Their inputs have been beneficial in providing newer di-

rections to the research and improving its quality. Over the course of the Ph.D., I have had the privilege to learn from faculty members of various departments through classes and colloquiums. All these interactions have been instrumental in developing a greater appreciation of the subject matter.

Many warm thanks to Sonal, Sangeetha Ma'am, Smitha, Beena, Margaret, Tusharika, Aditya Sir, Chandu Ji, and Rajesh Ji for their help in navigating the requisite paperwork and logistics. I appreciate the efficient help provided by Electrical Engineering office staff during the course of study, particularly Mr. Santhosh, Ms. Tanvi, Ms. Madhu, Ms. Vaishali, and Mr. Mangesh.

It has been an enjoyable experience interacting with various project members and learning from them: Nishant Sir, Nidhya, Priya, Ram, Kumar, Shwetha Ma'am, Swastika, and Rohan Kadam. I am also grateful to Ashish, Abhishek, and Rohan Kharade for their help with the simulator and the interactive discussions on the P1930.1 standard.

I cherish new friendships that have blossomed in IIT and older ones that sustained me through. I am thankful to Aniruddh, Priyanka, Mahak, Sagar, Annapurna, Tejashri, Meera, Ebin, Pavan, Geetanjali, and Suren for the lovely memories and for making the campus feel like home. I am grateful to Priyadarshini and Prasham for always being there. Most importantly, I am thankful to my amazing lab mates: Arghyadip, Indu, Meghna, Shashi, Pradnya, Anushree, and Sadaf. I have learnt so much from them over the years.

Pursuing my research dream would not have been possible without the help of my family. My parents, Usha and Upendra Nayak have lovingly supported my aspirations and have made many sacrifices. I am thankful to Maya and Rajani, who have been my confidantes, partners-in-crime, cheerleaders, and reality checkers. They have always provided unconditional love and support. I am also thankful to my in-laws, Ramadevi and Ravindran V. for their patience and support throughout these years. Lastly, I am grateful to my husband Suhas for steadfastly standing by me through the highs and the lows. He has been my staunchest friend and an incredible support in this endeavor. His understanding nature and pragmatic counsel have been a steadying influence both personally and professionally.

Akshatha Nayak M.

July 2021

List of Abbreviations

2G	Second Generation
3GPP	Third Generation Partnership Project
4G	Fourth Generation
5G	Fifth Generation
5GC	5G Core
ACPF	Application Control and Policy Function
AdpF	Adaptation Function
AF	Application Function
AgrF	Aggregation Function
AMF	Access and Mobility Management Function
AP	Access Point
API	Application Programming Interface
ARPU	Average Revenue Per User
ARQ	Automatic Request Repeat
AUSF	Authentication Server Function
BS	Base Station
BSF	Base Station Function
BSS	Basic Service Set
BSS	Business Support System
CAPEX	Capital Expenditure
CBR	Constant Bit Rate

C-RAN	Cloud RAN
CU	Centralized Unit
CUPS	Control and User Plane Separation
D2D	Device to Device
dBS	data plane Base Station
DFT-s-OFDM	Discrete Fourier Transform-spread-OFDM
DRAL	Device and Resource Abstraction Layer
DSCP	Differentiated Services Code Point
DU	Distributed Unit
E1AP	E1 Application protocol
eAMF	enhanced AMF
eMBB	Enhanced Mobile Broadband
eNB	eNodeB
EPC	Evolved Packet Core
ETSI	European Telecommunications Standards Institute
E-UTRA	Evolved Universal Terrestrial Radio Access
E-UTRAN	Evolved Terrestrial Radio Access Network
F1AP	F1 Application protocol
FR1	Frequency Range 1
FR2	Frequency Range 2
GBR	Guaranteed Bit Rate
gNB	gNodeB
gNB-CU	gNB Central Unit
gNB-CU-CP	gNB-CU Control Plane
gNB-CU-UP	gNB-CU User Plane
gNB-DU	gNB Distributed Unit

GTP	GPRS Tunneling Protocol
GTP-U	GPRS Tunneling Protocol - User Plane
GW	Gateway
HARQ	Hybrid Automatic Request Repeat
HetNet	Heterogeneous Network
IAB	Integrated Access and Backhaul
IMT	International Mobile Telecommunication
IoT	Internet of Things
IP	Internet Protocol
ITU	International Telecommunications Union
IWF	Inter-working Function
KPI	Key Performance Indicator
L1	Layer 1
L2	Layer 2
LA	Local Agent
LDPC	Low Density Parity Codes
LTE	Long Term Evolution
LVAP	Lightweight Virtual Access Point
LWA	LTE WLAN Aggregation
MAC	Medium Access Control
MANO	Management and Orchestrator
MDAF	Management Data Analytics Function
MIMO	Multiple Input Multiple Output
MME	Mobility Management Entity
MMTC	Massive Machine Type Communications
mmwave	millimeter wave

MN	Master Node
MPLS	Multi Protocol Label Switching
multi-RAT	multiple-Radio Access Technology
N3IWF	Non 3GPP InterWorking Function
NAS	Non Access Stratum
NBI	Northbound Interface
NEF	Network Exposure Function
NETCONF	Network Configuration Protocol
NF	Network Function
NFV	Network Function Virtualization
NFVI	NFV Infrastructure
NGAP	Next Generation Application Protocol
ng-eNB	Next generation eNodeB
NG-RAN	Next Generation Radio Access Network
NR	New Radio
NR-DC	NR-NR Dual Connectivity
NRF	Network Repository Function
ns-3	network simulator-3
NSSAI	Network Slice Selection Assistance Information
NSSF	Network Slice Selection Function
NWDAF	Network Data Analytics Function
OFDM	Orthogonal Frequency Division Multiplexing
ONF	Open Networking Foundation
OPEX	Operating Expense
OptF	Optimization Function
O-RAN	Open RAN

OS	Operating System
OSS	Operations Support System
P1930.1	Project 1930.1
PCF	Policy Control Function
PDCP	Packet Data Convergence Protocol
PDCP-C	Packet Data Convergence Protocol control
PDCP-U	PDCP User Plane
PDU	Protocol Data Unit
PFCP	Packet Forwarding Control Protocol
PGW	Packet Gateway
PGW-c	PGW-control
PGW-u	PGW-user plane
PHY	Physical layer
PRB	Physical Resource Block
QoE	Quality of Experience
QoS	Quality Of Service
RA	Registration Area
RAF	RAT Abstraction Function
RAN	Radio Access Network
RAT	Radio Access Technology
RB	Resource Block
REST	Representational State Transfer
RIC	Radio Interface Controller
RLC	Radio Link Control
RRC	Radio Resource Control
RRM	Radio Resource Management

RSRP	Reference Signal Received Power
RT	Real Time
SBI	Southbound Interface
SCP	Service Communication Proxy
SCTP	Stream Control Transmission Protocol
SDAP	Service Data Adaptation Protocol
SDN	Software Defined Networking
SDO	Standards Development Organization
SF	Security Function
SGW	Serving Gateway
SGW-c	SGW-control
SGW-u	SGW-user plane
SMF	Session Management Function
SN	Secondary Node
SNMP	Simple Network Management Protocol
S-NSSAI	Single Network Slice Selection Assistance Information
SON	Self Organizing Network
SRB3	Signaling Radio Bearer 3
SSID	Service Set IDentifier
TCP	Transmission Control Protocol
TTI	Transmission Time Interval
UAV	Unmanned Aerial Vehicle
UDM	Unified Data Management
UDN	Ultra Dense Network
UDP	User Datagram Protocol
UDR	Unified Data Repository

UE	User Equipment
UFCF	UE and Flow Control Function
UPF	User Plane Function
URLLC	Ultra Reliable Low Latency Communications
V2X	Vehicle to Everything
vdBS	virtual dBS
vGW	Virtual Gateway
VIM	Virtual Infrastructure Manager
VirtRAN	Virtualized RAN
VL	Virtualization Layer
VNF	Virtual Network Function
VNFM	Virtual Network Function Manager
VoIP	Voice over Internet Protocol
VRB	Virtual Resource Block
VSF	Virtual Subsystem Function
WC	WLAN Controller
WCIF	WLAN Control Interface Function
WiMAX	Worldwide interoperability for Microwave Access
WLAN	Wireless Local Area Network

List of Tables

List of Figures

Chapter 1

Introduction

Mobile broadband technology has witnessed widespread adoption in the last decade, with an increase of 1.6 billion subscribers worldwide in the previous year alone [1]. Significant growth in the number of subscriptions, coupled with the rising popularity of multi-media content, has led to a surge in user data consumption. While the demand for data continues to grow, the cellular network itself is in a state of metamorphosis, with an increasing number of smart devices being connected to the network. Diverse applications and devices with varied requirements of latency, power, and throughput are becoming common within the network [2]. All these factors have necessitated the development and deployment of the next-generation technology standards, i.e., Fifth Generation (5G) standards.

1.1 5G - Emergence of a Disruptive Technology

The emergence of 5G is expected to not only enhance network capabilities but also catalyze the rise of digital economies by enabling newer applications, faster innovation, and fostering socio-economic growth [3]. 5G would provide the technology platform to connect people to everything around them, such as cars, homes, etc., in more meaningful ways. Early market studies have predicted that by the year 2035, the adoption of 5G would create 22.3 Million jobs and account for 3.6 Trillion dollars in global economic output. It is also predicted that 5G would prove to be a disruptive technology, transforming entire economies and industries.

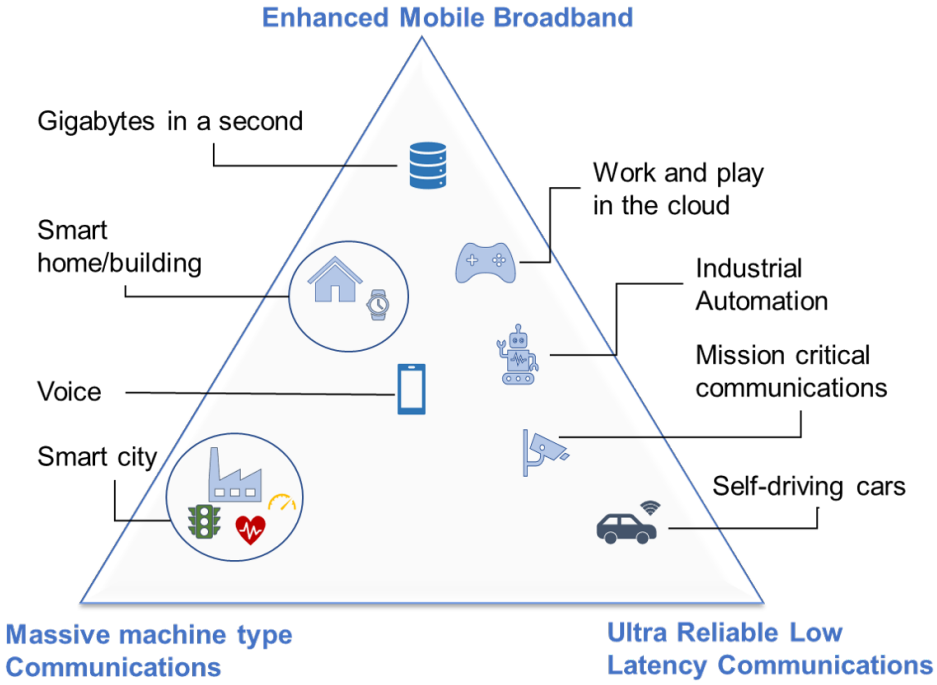


Figure 1.1: Key use-cases of IMT-2020 and beyond (courtesy [4])

The framework and objectives of 5G were first formalized in 2015 by the International Telecommunications Union (ITU) in a report titled "International Mobile Telecommunication (IMT) vision–framework and overall objectives of the future development of IMT for 2020 and beyond" [4]. This report detailed three use cases and the envisaged capabilities of the next-generation networks (also referred to as IMT-2020). The use-cases, together with a few representative applications, are illustrated in Figure 1.1. The intended usage-scenarios and characteristics of the use-cases are described below.

- Enhanced Mobile Broadband (eMBB): eMBB is designed to support scenarios requiring high data rates (upto 20 Gbps), increased mobility, and high user density (upto 10^6 devices/km^2). The expected data rates may vary based on the type of deployment. For example, eMBB may be deployed in both wide-area for highly mobile users (through macro-cells) or as hotspots supporting large user-density (through small-cells). Typically, user data rates provided for small cell deployments are expected to be higher in comparison to data rates for macro-cell deployments [4].

- Massive Machine Type Communications (MMTC): This use-case is expected to support a class of non-delay sensitive applications requiring low data rates and

large user densities while keeping energy consumption to a minimum. Examples include data collected from sensors for applications such as smart metering, air quality monitoring.

- Ultra Reliable Low Latency Communications (URLLC):URLLC comprises a class of applications requiring stringent availability, latency, throughput, and higher accuracy for positioning. Example applications in this category include industrial automation, mission critical applications, and Unmanned Aerial Vehicle (UAV) control.

To understand the scale of performance improvements provided by 5G vis-a-vis the previous Fourth Generation (4G) technology, we compare a few common Key Performance Indicators (KPIs) [4]. An illustration of this comparison is present in Figure 1.2. As shown in the figure, the achievable peak data rate of IMT-2020 ($\approx$ 20 Gbps) exceeds 20 times that of IMT-Advanced (i.e., 4G technology). Similarly, there is a tenfold improvement in traffic capacity, energy efficiency, latency, and maximum supportable connection density. However, it is essential to note that these KPIs vary across usage-scenarios, and all of them may not be supported simultaneously. For example, the MMTC use-case typically supports a high-connection density but lower data rates than that of the eMBB use-case. To understand how these improvements are achieved in practice, we look at the working of the various 5G standards at present.

1.2 5G Standardardization Initiatives

Various Standards Development Organizations (SDOs) are developing different versions of the 5G standard based on ITU's recommendations. The details of the specifications developed by two prominent entities, viz., the Third Generation Partnership Project (3GPP) and Institute of Electrical and Electronics Engineers (IEEE), are described below.

1.2.1 3GPP 5G Specifications

3GPP 5G specifications are being developed as a candidate for the IMT-2020 family of standards. The main technology enhancements in 3GPP 5G are illustrated in Figure 1.3. Some of the important enhancements include-

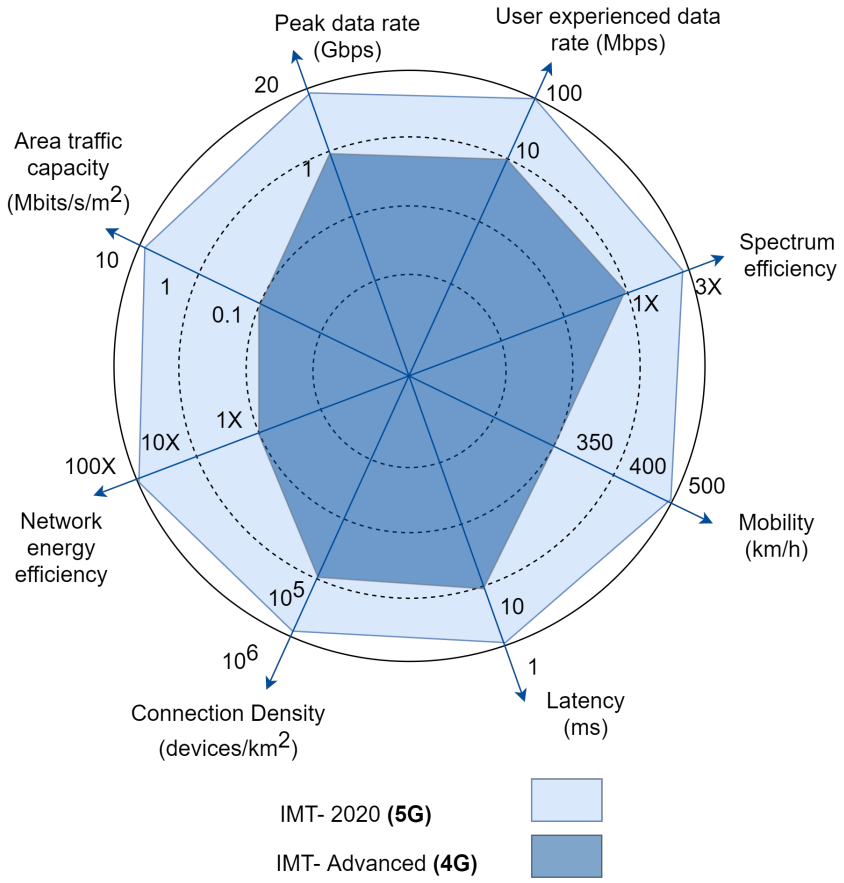


Figure 1.2: Comparison of 4G and 5G KPIs (courtesy [4])

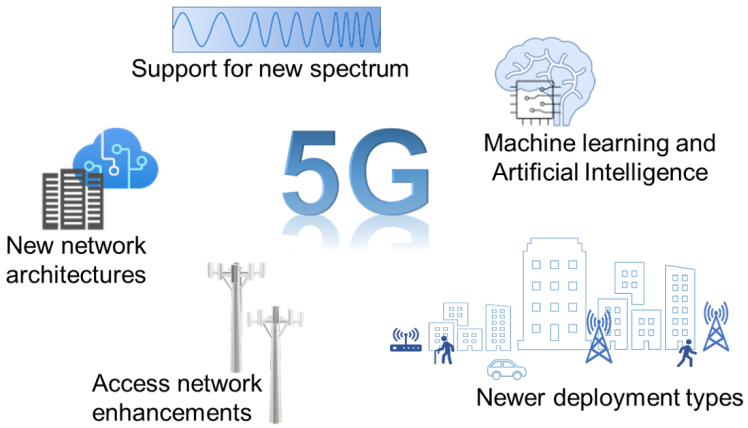


Figure 1.3: 3GPP 5G technology enhancements (courtesy [3])

- **Support for new spectrum bands and higher bandwidths:** 3GPP 5G specifications support a broader set of frequency bands enabling increased transmission bandwidths [5]. The first set of spectrum bands, i.e., Frequency Range 1 (FR1) (commonly known as the 'sub 6 GHz' band), extends from 410 MHz - 7.125 GHz and has a maximum transmission bandwidth of 100 MHz. The second band, i.e., Frequency Range 2 (FR2) also known as the millimeter wave (mmwave) band, supports frequencies between 24.25 GHz – 52.6 GHz. It has a maximum transmission bandwidth of 400 MHz.

- **Enhancements in access network:** 3GPP 5G provides significant performance improvements over its technological predecessor, i.e., 4G Long Term Evolution (LTE). This is done by introducing a new radio interface known as New Radio (NR). 5G NR has higher spectral efficiency in comparison to LTE as it supports beam-forming and massive Multiple Input Multiple Output (MIMO). Unlike LTE, 5G uses Low Density Parity Codes (LDPC) for encoding data channels, as they provide higher throughput and can be decoded in parallel. Polar codes are used for encoding control channels as they have low algorithmic complexity and noise floor. Unlike LTE, 5G NR adopts Orthogonal Frequency Division Multiplexing (OFDM) as the multiple access scheme on both downlink and uplink. 5G also supports Discrete Fourier Transform-spread-OFDM (DFT-s-OFDM) on the uplink [6]. More importantly, 5G introduces a new concept known as 'Numerology,' allowing for flexible configuration of Medium Access Control (MAC) frame structures and radio resource allocations based on the use case, bandwidth, and spectrum availability, while ensuring backward compatibility with LTE.

- **Support for flexible architectures and deployment types:** 3GPP 5G incorporates Software Defined Networking (SDN) and Network Function Virtualization (NFV) as key technology enablers. It introduces a new core network with a service-based architecture comprising network functions with well-defined Application Programming Interfaces (APIs) for interacting with each other. The 5G Radio Access Network (RAN) also allows for different functional split options. A detailed description of the 3GPP 5G architecture is provided in Chapter 3. The use of SDN and NFV enables us to build different architectural solutions customized to the use case.

For example, Cloud RAN (C-RAN) is an architecture where data is processed in a logically centralized server [7] within the cellular core network. Other architectural elements like 'Fog' place network intelligence at the network edge and process data close to the user. 3GPP 5G also supports the creation of multiple end-to-end logical networks known as 'network slices' over a shared infrastructure. A brief overview of SDN, NFV, and network slicing is presented in Section 1.3. 5G can also enable newer deployment types such as Ultra Dense Networks (UDNs), wherein the number of base stations deployed would be more than the number of users.

- **Use of machine learning and artificial intelligence:** 5G defines new network functions such as Network Data Analytics Function (NWDAF) and Management Data Analytics Function (MDAF) for collecting and performing data analytics on network and management data, respectively [8]. Based on its implementation, NWDAF enables the use of machine learning algorithms such as regression and deep learning algorithms to perform functions like optimizing User Equipment (UE) connection management, UE mobility management, etc. [9].

At present, three phases of the 3GPP 5G specifications have been defined viz., Releases 15, 16 and 17. The third phase, i.e., Release 17, is expected to be completed in June 2022. Cellular networks confirming to Release 15 specifications have already been deployed in a few regions around the world. A summary of a few important features in each release is provided in Table 1.1.

As mentioned in the table, Release 15 was the base-line release for 3GPP 5G. It has primarily focused on the architectural evolution from LTE by defining a 'non standalone' architecture where a 5G NR base station can be connected to LTE core network to provide a subset of 5G functionality. It also defines 5G 'standalone' architecture wherein 5G NR base stations could connect to 3GPP 5G Core (5GC), supporting all 5G features. Release 15 specifies inter-working enhancements with Wireless Local Area Network (WLAN) such as voice over WLAN, identification, and charging for traffic transported over WLAN. An exhaustive list of Release 15 feature details is available in [10].

Release 16 enhancements [11] include support for notifications for Vehicle to Everything (V2X) services, wherein any change in Quality Of Service (QoS) is notified

Table 1.1: Summary of 3GPP Release 15, 16 and 17 features

Release No.	**Feature Summary**
Release 15	5G System Phase-1
	Introduction of New Radio (NR)
	Improvements to LTE
	Wireless Local Area Network (WLAN) and unlicensed spectrum
	Service based architecture
	Slicing: end-to-end logical networks
	Third party access to 5G services through Application Programmable Interface (API) exposure
	Vehicle to Everything (V2X) Phase 2
Release 16	5G System Phase-2
	Industrial Internet of Things (IoT)
	V2X Phase 3
	URLLC
	Integrated Access and Backhaul (IAB)
	NR based Access to Unlicensed Spectrum
Release 17	NR MIMO and sidelink enhancement
	Dynamic spectrum sharing enhancements
	RAN slicing
	5G Multicast broadcast
	Network slicing Phase 2
	RAN slicing
	Unmanned Aerial Systems
	5G wireless and wireline convergence

to the platooning vehicles for automatically readjusting inter-vehicular distances. Also, inter-working for network slices, for UE mobility between 4G core network and 5GC are specified. Similarly, the tentative list of features for Release 17 is summarized in Table 1.1. However, due to the ongoing standardization process, details on the finalized features and enhancements would only be available at a later point in time.

1.2.2 IEEE 5G Standards

Although IEEE has not proposed a candidate technology for IMT-2020, it specifies multiple standardization activities as a part of its 5G initiatives. Among the set of IEEE 5G standards, two primary standards are IEEE 802.11ax and IEEE 802.11ay. IEEE 802.11ax is specified for operations in the 'sub 6 GHz' band with maximum bandwidths of 160 MHz and 40 MHz for the 5 GHz and 2.4 GHz bands, respectively.

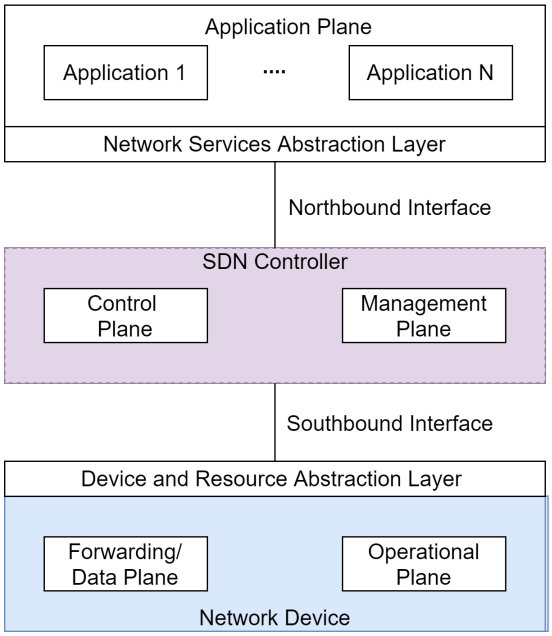


Figure 1.4: SDN architecture (adapted from [14])

It supports data rates upto 9.6 Gbps and adopts OFDMA as the multiple access protocol on both uplink and downlink [12]. Other enhancements from the previous generation (IEEE 802.11ac) include support for opportunistic power saving and virtualization (i.e., creation of multiple virtual Access Points (APs) over a single physical AP). The feature enables the creation of multiple (upto 32) independent basic service sets. The final version of this standard has been approved in February 2021.

IEEE 802.11ay operates in the unlicensed mmwave band at frequencies above 60 GHz. The maximum operating bandwidth is 8.64 GHz, and the maximum supported data rate is 100 Gbps. The enhancements in this standard include support for channel bonding and carrier aggregation, support for beamforming for MIMO and multi-user MIMO, etc [13]. IEEE 802.11ay standard is expected to be approved in March 2021.

1.3 Key Technology Enablers of 5G - An Introduction

Telecom industry is increasingly turning to softwarized technologies such as SDN and NFV to create flexible networks that support a diverse set of services.

1.3.1 Software Defined Network (SDN)

SDN is a networking paradigm that originated through experimentation in Internet Protocol (IP) based networks and has since then been extensively used in data centers [15]. As defined in [14], SDN is - *"A programmable networks approach that supports the separation of control and forwarding planes via standardized interfaces"*. The component planes within a software defined network are illustrated in Figure 1.4. As illustrated in the figure, network control and management may be logically centralized at an entity called the SDN controller. The control plane is a collection of functions responsible for controlling network devices. The control plane is responsible for making packet forwarding decisions and configuring the rules for packet forwarding on network devices. Examples of control plane functionalities include topology discovery, packet route selection and instantiation, etc. The management plane is responsible for monitoring, configuring, and maintaining one or more network devices. Examples of management plane functions are fault management, configuration management, etc. Note that the control plane interacts mostly with the forwarding plane, whereas the management plane interacts with the operational plane.

The forwarding plane/data plane/user plane is a collection of resources across all network devices responsible for forwarding traffic [14]. For example, in the case of a Switch, the forwarding plane may comprise ports used for the reception and transmission of packets and a forwarding table with its associated logic. The forwarding plane is responsible for handling packets in the data path based on the instructions received from the control plane. It performs several actions on packets in the data-path such as forwarding, dropping or modification based on instructions from the control plane. The operational plane consists of information regarding the state of the forwarding device, e.g., number of ports, active/inactive status, etc. Sometimes, the operational plane is considered to be a part of the data plane and not a separate entity in itself. The SDN controller and data planes are separated through a standard interface known as the 'Southbound Interface (SBI)'. Various protocols such as OpenFlow [16], OF-CONFIG, etc., can be used on the SBI. The separation of control and data plane functionality through a standardized interface enables each plane to be scaled independently.

SDN based network architecture also defines an independent application plane on top of the controller [14]. These two planes are connected through a standard interface known as the 'Northbound Interface (NBI)', which is typically a Representational State Transfer (REST) based API. The abstraction layers shown in Figure 1.4 are responsible for abstracting the resources from the plane below and presenting it to the plane above. For example, the Device and Resource Abstraction Layer (DRAL) abstracts details of the network device and presents it to SDN Controller. Due to the presence of standardized interfaces between application, control, and forwarding planes, the development and deployment of new applications in SDNs can be achieved with ease. This scenario is unlike the present-day network where the clear separation of planes is absent [17]. SDN also simplifies network management with policy-based rules, thus eliminating the need for vendor-specific device configurations.

SDN is often used in conjunction with NFV for providing increased network flexibility. NFV is defined by European Telecommunications Standards Institute (ETSI) specifications [18] as - *"the principle of separating network functions from the hardware they run on by using virtual hardware abstraction"*. NFV is used to decouple/virtualize network functions that run on specialized hardware into virtual network functions/software that could be installed on general-purpose hardware. While SDN provides interface separation across control and forwarding planes, NFV helps in their dynamic instantiation and scaling. NFV provides the ability to virtualize hardware resources such as storage, compute, and network. It also enables the flexible allocation of resources to different network functions.

1.3.2 Network Function Virtualization (NFV)

The architectural framework for NFV illustrating the component blocks and their interfaces is illustrated in Figure 1.5. The NFV Infrastructure (NFVI) can span across multiple locations and is defined as *"the totality of all hardware and software components which build up the environment in which Virtual Network Functions (VNFs) are deployed, managed and executed"* [19]. The hardware components comprise compute resources, e.g., general-purpose servers, storage resources, e.g., network attached storage servers, and network resources, e.g., routers, wired/wireless links.

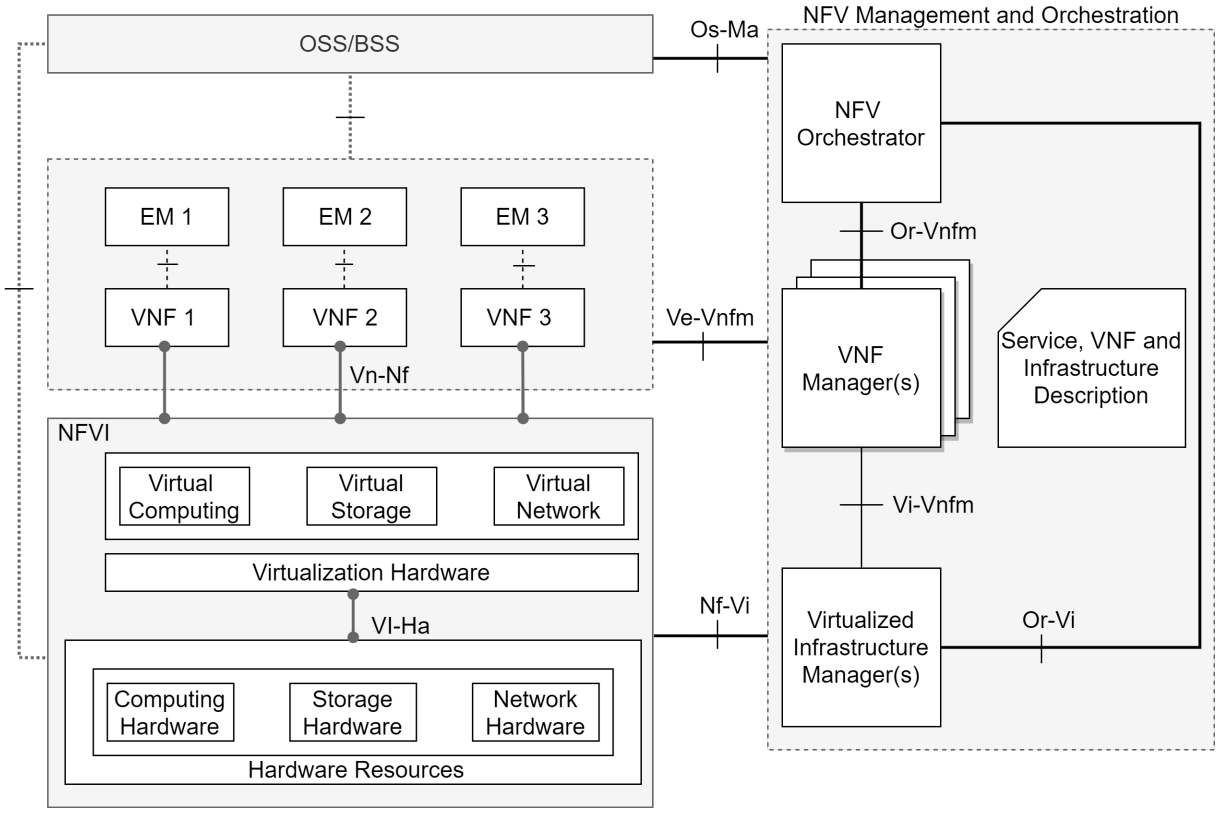


Figure 1.5: NFV architecture framework (courtesy [19])

The virtualization layer (also known as the hypervisor) abstracts the hardware resources into functional blocks having well-defined external interfaces and functional behavior, i.e., Network Functions (NFs) through software virtualization techniques. These virtualized network functions are referred to as 'VNFs'. The Operations Support System (OSS) and Business Support System (BSS) are a group of applications used for providing service delivery to customers. OSS is responsible for managing network operations such as inventory, service provisioning, network configuration, and fault management. BSS is used for business-related management tasks such as customer management, revenue management, and order management.

The NFV Management and Orchestrator (MANO) component in the NFV framework connects to the NFVI and the OSS/BSS. It comprises VNF Manager, Virtual Infrastructure Manager (VIM), and an NFV Orchestrator. VIM virtualizes VNFs and also controls and manages their interactions with NFVI. It performs several functions, e.g., allocating VMs to hypervisors, allocating and reclaiming resources dynamically, etc. The life-cycle of a VNF, i.e., instantiation, update, query, scaling, termination, is man-

aged by a VNF Manager. NFV Orchestrator manages the life-cycle of network service and coordinates the management of network service life-cycle, VNF life-cycle (supported by the Virtual Network Function Manager (VNFM)), and NFVI resources (supported by the VIM) to ensure an optimized allocation of the necessary resources and connectivity [18].

1.3.3 Network Slicing

SDN enables the creation of logical (virtual) networks, also known as network slices, over a shared infrastructure [20]. The 3GPP 5G specifications define a network slice as '*A logical network that provides specific network capabilities and network characteristics*' [21]. A slice consists of a set of NFs and the corresponding resources. Each network slice can be used to support a different service without affecting other slices. For example, applications with large variations in QoS requirements can be supported using separate slices over the same physical infrastructure. This allows a service provider to deploy newer services over a common network infrastructure without affecting existing services or allows multiple operators to share a common physical infrastructure. A user can access multiple slices concurrently to avail different services at the same time.

A network slice instance is identified by an identifier known as Single Network Slice Selection Assistance Information (S-NSSAI) [21]. S-NSSAI comprises of two sub-identifiers viz., Slice/Service type and Slice Differentiator, respectively. Slice/service type provides information regarding the features and services provided by the slice. The slice differentiator is an optional component used to differentiate between multiple network slice instances of the same service type. Note that the life-cycle of a network slice instance is independent of the life-cycle of a network service [22].

An example wireless network resulting from an application of the SDN paradigm together with network slicing is illustrated in Figure 1.6. As illustrated, the usage of SDN results in the logical centralization of network intelligence at the SDN controller entity, providing it with a global view of the network resources. The controller forms the control plane of the network and is responsible for control-related decisions and configurations. The remaining nodes in the network, such as base stations, switches, routers, etc., are integrated with a software program generally known as an 'agent', which allows the controller to communicate the control decisions to these nodes. As a result, the nodes

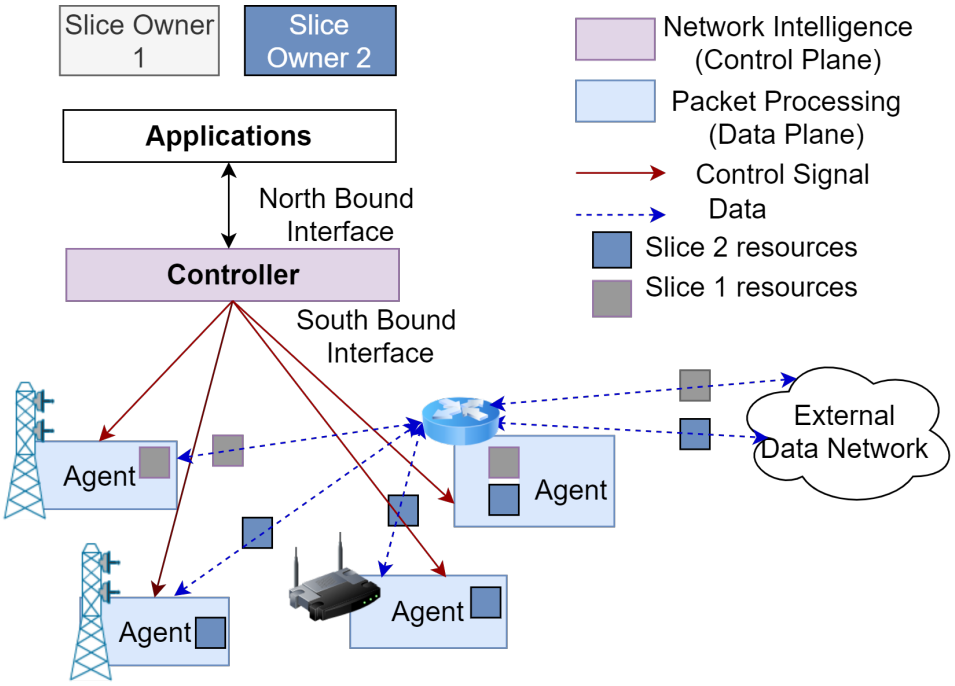


Figure 1.6: Illustration of SDN and network slicing in a wireless network

forward data packets based on the instructions provided by the controller to the agent and hence, comprise the data plane. This is in contrast to conventional networks, where network intelligence is distributed, and each node may perform calculations for forwarding/resource allocation, etc. Applications running on top of the controller may also be used to make policy decisions, which are then enforced by the controller. As observable from the figure, it is easier to roll out newer applications in an SDN based network, as the required changes affect only the controller and may not affect other applications.

The figure also illustrates the concept of network slicing wherein two operators designated as 'Slice owner 1' and 'Slice owner 2' are using the shared multiple-Radio Access Technology (multi-RAT) network. As seen from the figure, both these operators are guaranteed end-to-end resources on every node such as compute, storage, etc., as well as link resources such as bandwidth over the network slice that they own. This also ensures isolation between the operators, where an event in one of the slices (such as a surge in traffic) does not affect the slices owned by other operators. Slice isolation, in turn, would result in providing better performance to the individual subscribers.

1.4 Motivation for the Thesis

Migration from 4G to 5G technology is vital to avail the benefits provided by the 5G economy. However, the cost of network upgrades from present-day 4G LTE networks to 5G mobile network is very high, especially in RAN [23]. This is in contrast to the decrease in Average Revenue Per User (ARPU), especially in emerging markets like India [24]. Therefore, to provide an evolutionary path from 4G to 5G networks without incurring a considerable expenditure for network upgrades, service providers are considering two approaches.

The first approach involves architectural enhancements for coordinated operation with existing 4G systems [25]. The second approach involves supplementing cellular network capacity through the deployment of low-cost solutions in unlicensed spectrum, such as WLANs. As a result, next-generation networks would intrinsically comprise multi-RATs functioning in a harmonized manner to provide various services. However, at present, usage of multiple and diverse Radio Access Technologies (RATs) poses significant challenges in terms of inter-working and network management [26]. Moreover, due to the tightly coupled nature of the data and control plane in present-day networks, the implementation of newer services for supporting application and device diversity is a complicated task. Therefore, it is imperative that newer approaches unifying control and management across RATs are necessary for controlling next-generation networks.

Our work attempts to address several issues related to the centralization of network control, multi-RAT RAN control, and network slicing through the use of SDN and NFV. The usage of SDN and NFV in RAN presents several challenges. SDN was originally designed to be used in wired networks which typically do not have broadcast channels. Therefore, SDN controllers and protocols that were developed to control and manage wired networks do not have the capability to control and manage broadcast channels required in wireless networks.

Wireless cellular networks are also subject to spectrum and transmit/receive power regulations. The users in these networks are typically mobile and experience variations in radio-link quality due to Doppler effects, fading, and shadowing. In comparison to wired networks, the fluctuation/changes in radio link require complex physical and Medium

Access Control (MAC) layers in cellular wireless networks, which need to perform functions like Adaptive Modulation and Coding (AMC) and power control in addition to data transfer. Moreover, to utilize the scarce radio resources there may also be a need for compression of IP packet headers in data plane. Cellular architecture also requires the use of tunneling mechanisms for handling the mobility of a user when it moves across cells. These requirements bring additional complexity to the data plane protocols in the wireless networks. All this necessitates the support of additional functionality in an SDN Controller for controlling wireless networks.

Similarly, NFV requires virtualization of physical resources of different types such as network, compute and storage resources. Virtualization with the help of hypervisor in an NFV environment can be designed relatively easily when the physical resources to be sliced are indistinguishable from each other, for example, storage slices (e.g., parts of random access storage), compute slices (in terms of number of processing cycles of a CPU) or network slices (in terms of bandwidth on an optical fiber link) are indistinguishable. However, the radio resources are not indistinguishable due to factors such as frequency selective fading. Hence, using a hypervisor (over the physical layer) for slicing the radio resources is a difficult task. Details of the research contributions and the organization of the thesis are presented in Section 1.5.

1.5 Contributions and Organization of the Thesis

The primary focus of this thesis is on the design and evaluation of SDN based architectures for control and management of multi-RAT RANs. The proposed architectures are designed in a manner that allows evolutionary upgrades to present-day networks without requiring changes in UE hardware. The thesis also provides insights into slicing multi-RAT RANs and proposes a framework for slicing them recursively. Additionally, it also suggests protocol enhancements over 3GPP multi-connectivity for improving network slice mobility in the access network. Some of the proposed ideas, such as the recursive slicing framework and dis-aggregated RAN model, have been submitted as contributions to ongoing 5G standardization activity in IEEE [27]. A summary of the key contributions of the thesis is illustrated in Figure 1.7. The remainder of the thesis is organized as follows-

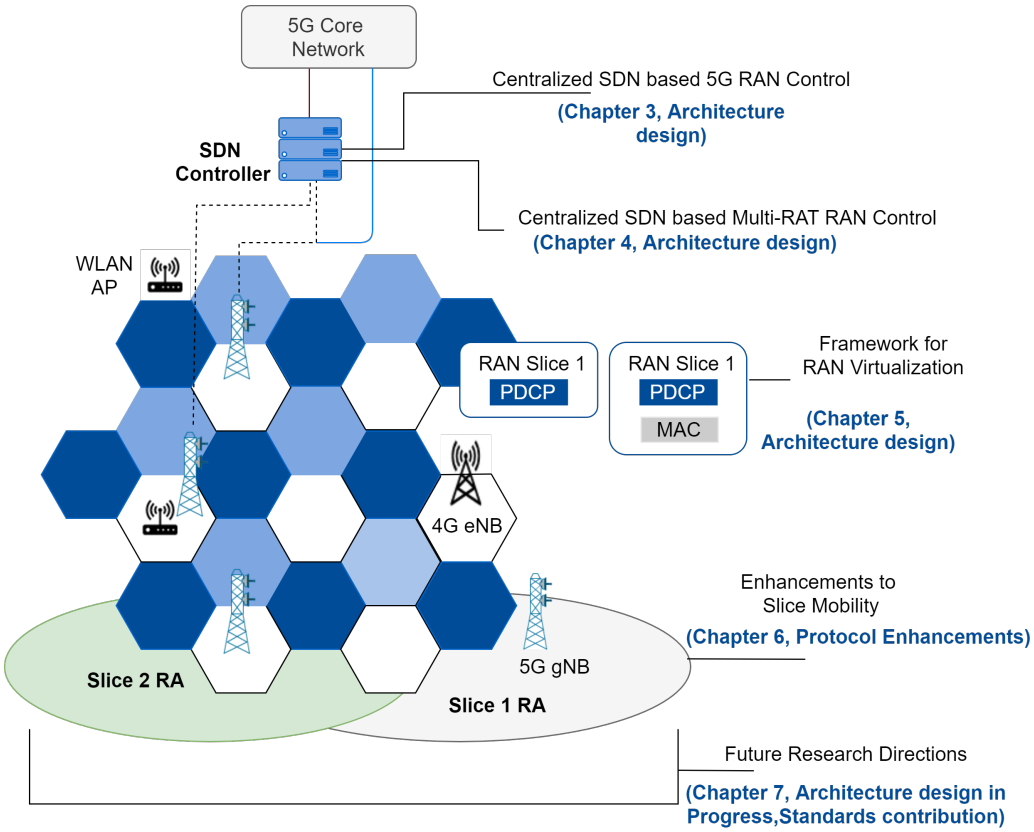


Figure 1.7: Summary of key contributions of the thesis

- In **Chapter 2**, we posit a review of key literature that has had made a significant impact on the study of software defined multi-RAT RAN architecture design and network slicing. We also describe the relevant standardization efforts for controlling and managing multi-RAT RANs. Furthermore, we attempt to highlight gaps in the existing literature and the open challenges in related areas.

- In **Chapter 3**, we provide an overview of the 3GPP 5G architecture. We address the issues related to SDN based control of the present-day 5G RAN and propose architectural enhancements for enabling centralized control for the end-to-end network. The proposed architecture relocates the control functionality present in 5G RAN to the core network. This results in logical centralization of network control while reducing signaling costs between 5G NR RAN and core network. We demonstrate the advantages of the proposed architecture in terms of signaling and latency reductions in comparison to the 3GPP 5G network through call flows. The proposal is further validated through simulations using network simulator-3 (ns-3) [28].

- In **Chapter 4**, we extend the ideas presented in Chapter 3 for enabling centralized control in a multi-RAT network. In this work, we propose an architecture and suggest mechanisms for providing end-to-end SDN based control in a network that supports multiple slices. The proposed architecture is scalable and designed to have minimal impact on the UE for facilitating an easier integration with operator network. The architecture provides RAT-agnostic interfaces to applications and a virtualized view of the network resources, enabling simplified control and management. The performance of the proposed architecture is evaluated using a custom-built SDN based multi-RAT simulator developed using ns-3[1]. The performance improvements of the proposed architecture in comparison with existing architecture are demonstrated for evaluation of KPIs.

- In **Chapter 5**, we describe open issues in slicing multi-RAT RANs today and emphasize the need for recursive framework(s) for RAN slicing. We also propose a novel framework, 'Virtualized RAN (VirtRAN)' which is designed as a recursive SDN/NFV based framework for multi-RAT RAN slicing. This framework can be used to address some of the prevalent gaps and support features like network slicing and user mobility management in 5G networks in an efficient manner. The work described in this chapter also provides insights and recommendations regarding the preferred protocol layer (at the base station) at which network slicing policies can be enforced based on different scenarios. These recommendations are validated through ns-3 simulations.

- In **Chapter 6**, we identify the gaps in 3GPP Release-15 specifications with regards to slice mobility, which require homogeneous slice deployments within a Registration Area for supporting slice mobility. We propose a multi-connectivity based approach for supporting slice mobility in areas where slice deployments are non-homogeneous. The proposed mechanisms result in lowering the Capital Expenditure (CAPEX) while providing performance improvements by reducing handover failures. This mechanism also improves slice availability and reduces handover failures in comparisons to deployments where UEs are not multi-connected. The improvements in

[1]The ns-3 based simulation platform was developed by Ashish Sharma, Rohan Kharade, and Abhishek Dandekar.

handover failure reductions and slice availability are demonstrated with the help of ns-3 simulations.

- In the concluding chapter, **Chapter 7**, we summarize the impact of our research work with a focus on standardization. We also provide a glimpse into ongoing early-stage work, such as the idea of dis-aggregating the multi-RAT RAN. The proposed RAN dis-aggregation solution identifies a set of modular functions within RAN nodes, such as base stations that are common for every RAT. These functions can be chained in a suitable manner to instantiate a base station for a particular RAT. RAN dis-aggregation not only introduces flexibility in the network but also simplifies multi-RAT network control. As mentioned earlier, this idea has been submitted to an ongoing 5G standardization activity viz., "Project 1930.1 (P1930.1)-Recommended Practice for SDN based Middleware for Control and Management of Wireless Networks" in IEEE. Lastly, possible directions for future research have also been explored.

Chapter 2

SDN for Multi-RAT Networks - State of the Art and Open Issues

As mentioned in the previous chapter, SDN/NFV based approaches for network control and management are predicted to be the mainstay for control and management of 5G and beyond networks. There has been significant scientific interest in this area in recent years, leading to a plethora of research proposals and a few standardization activities. This chapter attempts to outline the notable contributions and standards in this area with a view to describe the approaches considered. To achieve this, we first review key works related to SDN based control for a given access technology viz., WLAN, LTE and 5G and then progress to prior art supporting multi-RAT control. Wherever the solutions support network slicing, the mechanisms used for the same have been described. This is followed by a brief discussion of the standardization activities in this area. Finally, we highlight the key issues/challenges that are of interest to us and form the basis for the research described in subsequent chapters.

2.1 SDN based Wireless Network Control

2.1.1 SDN based WLAN Control

Odin [29] is an implementation of SDN controlled WLAN and is amongst the first solutions to illustrate the working of a user-centric WLAN architecture. The Odin architecture consists of a controller that controls APs within a WLAN. AP control is achieved with

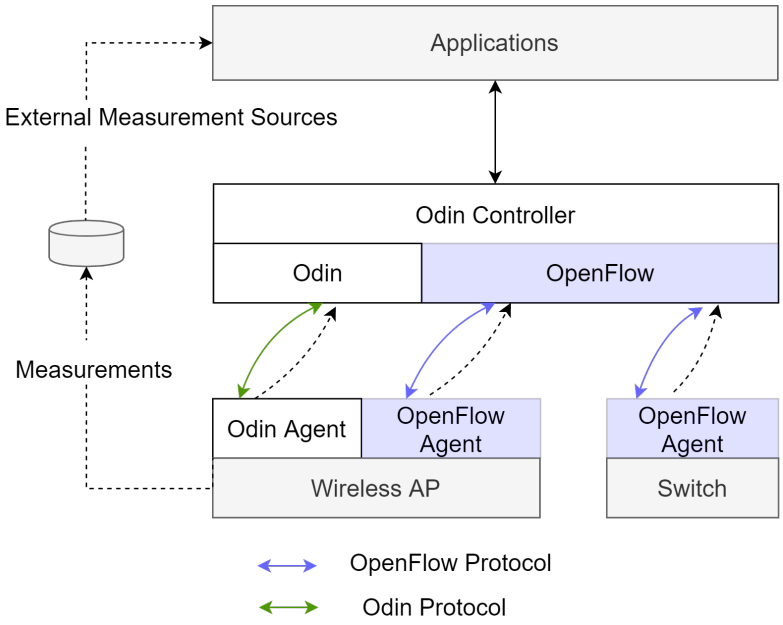


Figure 2.1: System architecture of Odin (courtesy [29])

the help of two agents running on the APs, i.e., Odin agent for radio control and OpenFlow agent for flow control. The authors also introduce the concept of a Lightweight Virtual Access Point (LVAP), which is an abstraction of the WLAN AP for a given user. LVAPs are installed on APs and are assigned to a given user by the Odin controller. As a result, it appears that each user has been assigned a dedicated AP. An LVAP is responsible for all communication with a single UE. LVAPs are migrated across APs when a user moves, and through this mechanism, handovers across APs are transparent to a given UE. The Odin framework also demonstrates a method for slicing WLAN. Within this framework, a network slice consists of a set of WLAN APs, clients with their associated LVAPs, and Service Set IDentifier (SSID) along with the associated network application.

Despite its elegance, the solution has certain shortcomings. Odin may not be scalable for a large number of UEs, as beacons are sent over multiple unicast channels to individual UEs by the corresponding LVAPs. Also, the proposed scheme may not work when the target AP is on a different channel than the source AP as mobility is no longer transparent to the UE. Moreover, this scheme may not be easily adapted for use over LTE and 5G NR due to the changes in protocol required.

Lasagna [30] is a WLAN control and slicing solution defined over 5G-EmPOWER [31].

It is an SDN based multi-RAT testbed designed to control LTE and WLAN. The authors implement slicing by using a programmable hypervisor on top of the Linux WLAN stack. They also extend OpenFlow match-action rule for WLAN environments and introduce a new abstraction known as 'traffic rule' for mapping a given portion of the flow space with a scheduler. The traffic rule identified by a tuple comprising SSID for identifying destination WLAN and Differentiated Services Code Point (DSCP) for indicating priority for an IP packet.

2.1.1.1 Software Defined Control for Cellular Networks

The bulk of available literature for SDN based control of cellular networks is based on the application of SDN principles to 3GPP LTE networks. However, they can be applied to the 5G network with equivalent modifications as the networks have been designed to work in a backward compatible manner, especially within the sub-6-GHz band. The existing works can be classified into three major categories, viz., work related to only RAN control, only core control, and end-to-end control. We summarize relevant work applying the SDN paradigm to the LTE core as well as RAN. However, we place more emphasis on the RAN related literature as it stems to be our area of interest.

2.1.2 Software Defined Core Network Control

SoftCell [32] is an SDN based core network solution aimed at improving the scalability of the core network by replacing the Packet Gateways (PGWs) in present-day networks with switches configured using an SDN controller. Softcell provides the ability to provide the fine-grained UE specific routing that is provided by the Evolved Packet Core (EPC) with an alternative cost-effective architecture illustrated in Figure 2.2. This architecture provides the advantages of location based routing (e.g., IP based routing) as well as tag based routing e.g., Multi Protocol Label Switching (MPLS) routing by aggregating UE traffic along multiple dimensions with identifiers such as policy, base station and UE ids. As illustrated in Figure 2.2, the proposed architecture consists of a Local Agent (LA) placed at a base station. The LA is responsible for caching flow rules provided by the controller. As a result, only the flows that do not have a matching rule at the LA are redirected to the controller. This reduces traffic directed to the controller and results in

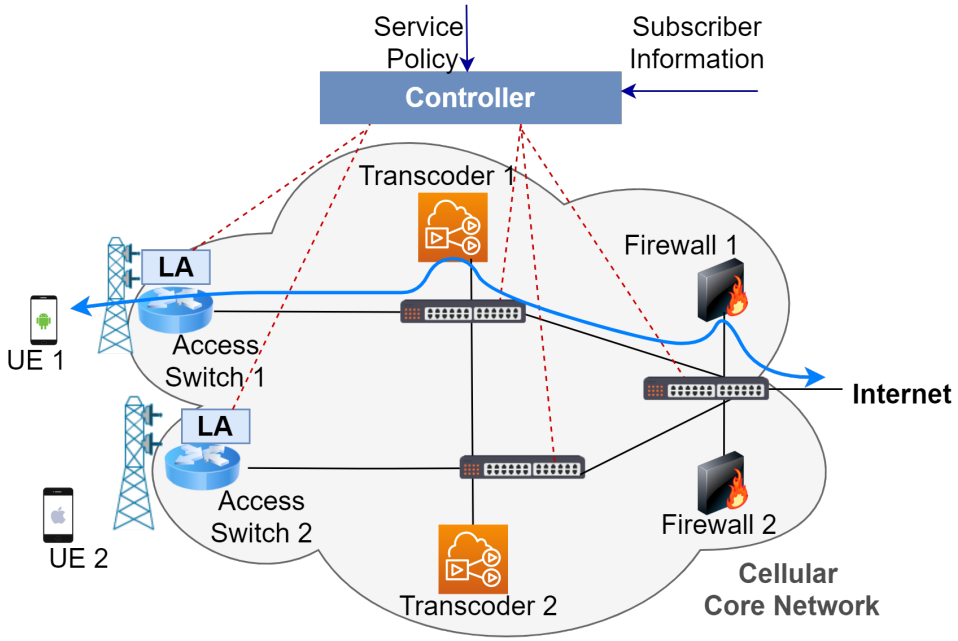


Figure 2.2: SoftCell architecture (courtesy [32])

lower delay. Unlike today's networks, where traffic is classified at the Gateway edge (at PGW), Softcell proposes to classify traffic at the access edge, i.e., at a switch closer to the base station. This is based on the assumption that most of the traffic is initiated by a UE. Once a particular flow is classified, it is tagged with multiple tags viz., service policy, base station, and UE ids. This tag is piggybacked on the downlink traffic as well as ensures that traffic is routed along the same middleboxes (e.g., firewalls, transcoders, etc.) in both directions. It also provides the flexibility of placing middleboxes anywhere within the network. The architecture also introduces a concept of location based IP address at the access edge. This IP is transparent to the UE and is only used for routing traffic within the core network. The authors provide fault tolerance in the network by using a backup SDN controller, which maintains the state coherently with the main controller. Similarly, failure of LAs are handled by restarting them and re-installing the policy rules from the controller. Although the architecture proposes a few novel ideas, there seem to be a few problems with the same. Adding LAs at every node increases the cost of the RAN. Further, flow rules may need periodic updates to maintain consistency.

Mobileflow [33] is an SDN based architecture for the 4G EPC, which was among the first works to demonstrate a proof-of-concept implementation. In this architecture, the

control plane element from the mobile core, i.e., the Mobility Management Entity (MME), is managed by the mobile controller. Other elements such as Serving Gateway (SGW) and PGW are configured using the Mobile Flow Forwarding Elements (MFFE). Authors have suggested the deployment of MFFEs as OpenFlow [16] switches with additional functionality such as Layer 3 (L3) tunneling and flexible charging. The paper also extends the Mobileflow model as an example application for implementing an SDN based end-to-end Evolved Packet System (EPS). In this case, the control functionality from all other data plane elements such as the eNodeB (eNB), SGW, and PGW have been abstracted out into the controller, and MFFEs are configured in such a manner that they can work with existing 4G architecture. The control and data plane split suggested within EPC, namely the SGW-control (SGW-c), PGW-control (PGW-c), SGW-user plane (SGW-u), PGW-user plane (PGW-u), are in line with the SDN features introduced in release 14 [34].

2.1.3 Software Defined RAN Control

SoftRAN [35] is an early work for SDN based architecture for cellular RAN. It is similar to a centralized Self Organizing Network (SON)/Radio Resource Management (Radio Resource Management (RRM)) solution proposed by 3GPP. In this work, the authors propose a hierarchical SDN controller for dense cellular deployments. The control functionality of a base station, which may impact its neighboring cells, such as handover decisions, transmit power control for mitigating interference, UE uplink resource block allocations, etc., are transposed into a global controller. The local controller is responsible for localized decisions of a physical base station. The underlying network resources are abstracted into a three-dimensional resource grid of base station index, time, and frequency slots. A solution for slicing SoftRAN was proposed in Radiovisor [36].

Radiovisor highlights the fact that interference is an additional factor for slice creation and management within wireless networks. Hence, spectrum resources allocated for each slice must be isolated and not interfere with one another. Radiovisor supports the inclusion of a per-slice controller, application(s) and deployment of layered configuration, e.g., scheduling for MAC, Physical layer (PHY) configuration, etc., for a specific slice flexibly and independently. However, the procedure for slicing control plane resources and ensuring isolation is unclear from the information provided.

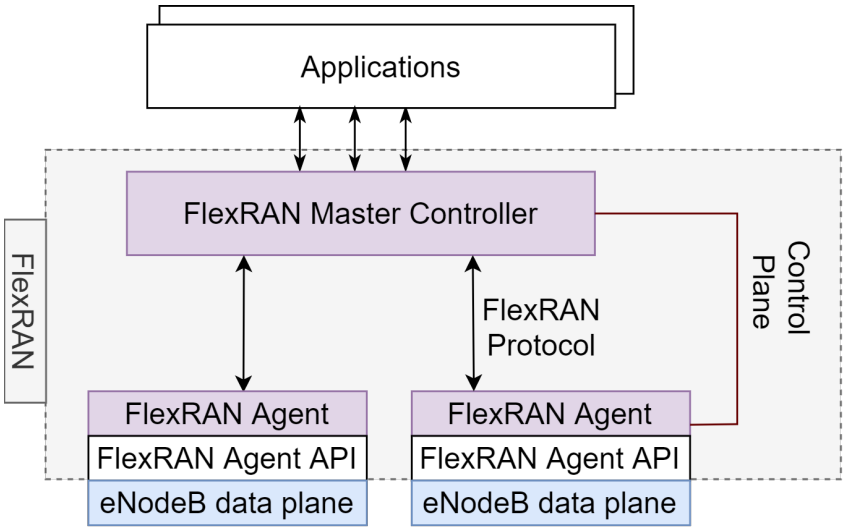


Figure 2.3: System architecture for FlexRAN (courtesy [37])

FlexRAN [37] is another proposal that introduces and implements the idea of software defined RAN for cellular networks. Although FlexRAN has been designed and implemented for LTE networks, the authors state that it is extensible for future RATs and also describe some of the necessary steps for the same. As illustrated in Figure 2.3, it is a hierarchical architecture with a centralized master controller and a FlexRAN agent (local controller) deployed at each LTE base station, i.e., the eNB. The control functionality within the eNB is transposed into the master controller. The master controller can perform scheduling and radio resource scheduling decisions centrally for eNBs under its control. The proposed architecture also provides the flexibility to use FlexRAN in bandwidth-constrained environments by introducing control modules known as Virtual Subsystem Functions (VSFs) within the FlexRAN agent where both scheduling policies and resource configurations can be provided and updated. This allows for localized operation at eNBs when necessary.

An interesting work that uses both C-RAN [7] architecture and FlexRAN controller to develop a prototype for a dynamic slicing solution for LTE is given in [38]. Here, the authors isolate the control part of the MAC layer, which is responsible for scheduling decisions, and instantiate it as an application on the controller. A slice manager function is deployed at the south bound interface of the controller to translate the scheduling decisions to Resource Block (RB) allocations. As a result, the eNB is only responsible

for RB allocation. The authors evaluate the performance of a conventional SDN based LTE network, which is referred to as a baseline system, and compare it to two scenarios. In the first scenario, the resources are statically sliced, and in the second scenario, the resources are dynamically sliced. Within the slicing scenario, the slice owners request the slicing controller for the required number of resources, and the slicing decision is conveyed to the 'slice manager' function through the SBI of the controller. The authors also demonstrate that user satisfaction in the case of sliced scenarios is higher, with dynamic slicing providing better performance in comparison to static slicing scenarios.

2.1.4 Software Defined Multi-RAT Networks

OpenRoads [23] is an early work that demonstrated SDN based control and management of multi-RAT networks, such as WLAN and Worldwide interoperability for Microwave Access (WiMAX). In this solution, the authors propose using OpenFlow for configuring routes within a network and the use of Simple Network Management Protocol (SNMP) [39] to configure radio-related parameters on the APs. This architecture treats WLAN AP as an OpenFlow switch. OpenRoads architecture comprises three layers, i.e., flow layer consisting of OpenFlow tables and SNMP, slicing layer consisting of a virtualization solution known as Flowvisor [40] for slicing the network and controller layer consisting of a controller for centralized network control. The Flowvisor is placed in between the APs and SDN controllers and divides the flow-space manifested by individual APs into smaller sub-spaces and maps these individual sub-spaces to separate network slices. Individual network slices may be controlled by different SDN controllers.

Since IEEE 802.11 MAC layer has many similarities to Ethernet MAC, it is possible to view WLAN APs (with 802.11 MAC) as Ethernet switches and use OpenFlow protocol to control them. However, there are inherent limitations in using flow level abstraction as the interface between control and data planes in wireless networks for a few reasons. For example, if we use this interface, the allocation of underlying radio resources, e.g., bandwidth to each of the network slices (flow space), is completely hidden from the SDN controller. Instead, it becomes the responsibility of the data plane entities (APs), thereby defeating the purpose of having an SDN based network architecture. Also, as Flowvisor is responsible for creating slices over the flow-space manifested by APs, the APs themselves

are unaware of the network slices. Further, due to time and user-specific variation in radio channels, the allocation of radio resources to different slices may vary over time. Due to APs' unawareness of the network slices, they are unable to maintain slice-specific separation of radio resources. As far as other mobile technologies, such as LTE and 5G NR are concerned, the applicability of OpenFlow as an interface between the RAN control plane and data plane functions is even more limited as their radio protocol structure is more complex in comparison to IEEE 802.11 WLANs. Besides, they also use abstractions such as radio bearers, which may need to be manipulated by the controller.

5G-EmPOWER [31] is one of the first works that has implemented SDN based multi-RAT controller. The solution provides a framework with an SDN controller as a network operating system to control and manage LTE and WLAN with the help of a unified controller. The proposal defines a new management protocol known as OpenEmpower and an Operating System (OS) known as 5G-EmPOWER. The architecture transposes management functionalities from RAN nodes and moves them to the management plane running over 5G-EmPOWER OS. The 5G-EmPOWER operating system behaves like a controller, and a 5G-EmPOWER agent is placed on each RAN node so that it can be configured by the OS. The 5G-EmPOWER OS is responsible for functions such as allocating data plane resources for users, providing isolation between users, providing a RAT-agnostic view of resources to users by abstracting network resource details, etc.

The 5G-EmPOWER framework also demonstrates RAN slicing for LTE network. The proposed slicing mechanism places a hypervisor above the physical layer. The hypervisor performs the abstraction of Physical Resource Blocks (PRBs) into virtual PRBs, which can then be grouped into virtual PRB groups for use. A slice resource manager placed at the MAC layer above the hypervisor is used for managing the slice. Multiple slices with independent schedulers can be created at the MAC. The virtual PRB groups created with the help of the hypervisor are then mapped to be allocated to be used by slice-specific schedulers for performing slicing. However, the authors do not provide details on how slicing could be performed over WLAN. Similar to SoftRAN, the 5G-EmPOWER framework has been developed with a view to support centralized SON server functionality over multi-RAT RAN.

A more comprehensive survey of SDN architectures for the cellular RAN and core network can be found in [41]and [42], respectively. Similarly, further reading on network slicing is available in works such as [43–45].

2.2 Industry Initiatives and Standardization Activities

SDN and network slicing have garnered considerable interest from operators, resulting in initiatives such as Open RAN (O-RAN) [46]. O-RAN was started in 2017 by a consortium of cellular network operators with an objective to develop SDN based smart RAN for Second Generation (2G) and beyond cellular RATs. It is also aimed at providing open interfaces to cellular RAN for enabling vendor inter-operability and usage of artificial intelligence/machine learning algorithms for optimized network decisions. O-RAN APIs are defined using 3GPP specifications as their basis. This standard promotes the usage of open-source software and off-the-shelf hardware for reducing CAPEX.

The system architecture of O-RAN is illustrated in Figure 2.4. The radio interface control functions in O-RAN are decoupled as non-Real Time (RT) ($> 1s$) and near-RT ($< 1s$) based on the time scale of operation. The non-RT Radio Interface Controller (RIC) is responsible for longer time-scale decisions such as policy management, configuration, training of learning models from the collected data, etc. On the other hand, the near-RT RIC interfaces with the non-RT RIC through the A1 interface and provides RRM related functionality such as mobility management, QoS management, etc. It also enables third-party applications to be easily incorporated into the network and maintains a near-RT network state by gathering data from the layers below through the E2 interface. O-RAN supports 4G LTE and 5G NR RATs at present. As within the 3GPP 5G specs [47], the radio protocol stack has been split into Centralized Unit (CU) and Distributed Units (DUs). The interfaces defined by 3GPP, such as E1 (between control and data plane) and F1 (between CU and DU), are being extended for use within the O-RAN standard. O-RAN is built as an extension to 3GPP and hence does not provide any specific guidelines for slicing the RAN. It is intended that the mechanisms defined by 3GPP would be used as-is unless explicitly mentioned within O-RAN specifications. As a result, it is inferred that slicing within O-RAN is also implementation dependent. Moreover, O-RAN does not support the control and management of other RATs such as WLAN at present.

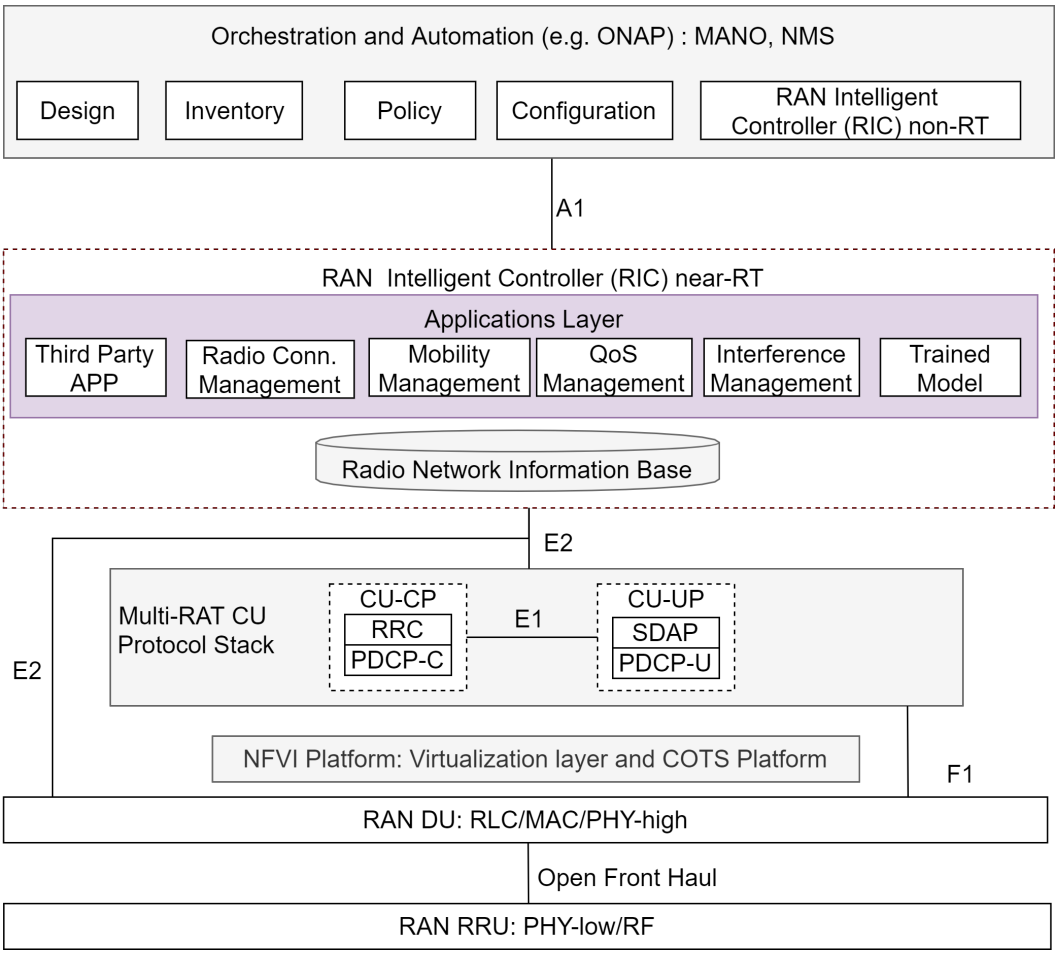


Figure 2.4: O-RAN architecture (courtesy [46])

Significant steps to address the gaps in SDN based multi-RAT control and management are also being undertaken in IEEE through Project 1930.1 [27]. This standardization activity was initiated in 2017. It endeavors to develop an SDN based framework for control and management of multi-RAT networks. The standard also provides a mechanism to achieve vendor-independent control of existing networks through the introduction of a 'middleware' between the SDN controller and the physical infrastructure. Several related aspects, such as middleware and interface design, inter-working mechanisms between various RATs through the middleware, dis-aggregation mechanisms for multi-RAT RANs, etc., are being addressed within the scope of this standard. We have submitted several contributions to this standardization initiative since its inception. Some of the ideas presented in the thesis are adapted as contributions to the standard. They are highlighted as they occur in subsequent chapters.

As network slicing is one of the essential features of 5G networks, several initiatives related to network slicing are being carried out within 3GPP. At present, 3GPP is not only working on defining the format and sequence of slice-related information exchanged during system procedures but is also working towards defining specifications for the management of slice instances. For example, work on specifications related to slice management [48], architectures [49] and provisioning [50] is presently underway in 3GPP.

2.3 Open Challenges

Based on the discussion so far, we can obtain several insights into the open issues in the control and management of multi-RAT networks. These are listed below.

- Although we have discussed prior works related to SDN based control of cellular networks [35,37,51], the existing literature does not provide logical centralization of RAN control. Centralized RAN control leads to better system performance for most applications, especially for procedures requiring coordination across base stations, e.g., mobility, load balancing. Although 3GPP 5G now defines an architecture that centralizes RAN control within a 5G NR base station, i.e., gNodeB (gNB), control between two gNBs, and as well as end-to-end network control is still distributed in nature.

- Secondly, there are several challenges in SDN based multi-RAT RAN control. Due to the intrinsic nature of protocols used, some of the solutions, e.g., OpenRoads [23] are not easily extensible for the control of other RATs. Much of the available literature lacks precise details on aspects related to slicing and/or functional split.

- There is also a conspicuous gap in the availability of standards for SDN based control and management of networks such as WLANs.

- It is also important to note that the available prior-art does not take cognizance of the fact that there are fundamental similarities in functionalities across RATs. By identifying/dis-aggregating these functionalities and modeling multi-RAT networks as a conglomerate of NFs (composed of these functionalities), network flexibility can be enhanced.

- Another vital aspect is that the available proposals related to slicing are very specific to the architecture [35, 37]. Although a few of the works [23] are aimed at slicing multi-RAT networks, they are not easily extensible to other RATs. Hence, there is a distinct lack of a generalized framework for slicing multi-RAT RANs.

- Finally, although 3GPP has specified several message exchanges relating to slicing, it does not specify mechanisms for slicing the RAN. Moreover, the present standards recommend homogeneous slice deployments over a Registration Area to support UE slice mobility. The cost associated with such implementations may be prohibitive and necessitates the exploration of alternative mechanisms for ensuring slice mobility.

This thesis attempts to address the above challenges in the course of the forthcoming chapters. We begin our investigations in the next chapter by studying the problem of centralization of RAN control in the 3GPP 5G RAN.

Chapter 3

SDN based Centralized Architecture for the 3GPP 5G Network

As discussed in Chapter 2, logical centralization of RAN control is important for providing better system performance. In this chapter, we seek to resolve this challenge by proposing a novel method for adapting the 3GPP 5G architecture to provide centralized RAN control while adhering to SDN principles[1]. To better understand the rationale behind the enhancements and their impact on the 5G network architecture, we provide a brief description of the 3GPP 5G architectural blocks and their functionalities before arriving at the proposal.

3.1 3GPP 5G Architecture - A brief Introduction

The system architecture for the 3GPP 5G cellular network marks a departure from 4G LTE architecture by incorporating SDN and NFV as key technology enablers in the architecture [21]. Therefore, both access network and 5GC networks are re-architected based on SDN principles. Although SDN has been introduced in LTE Release 14 through a feature referred to as Control and User Plane Separation (CUPS), it is limited to the 4G core network, i.e., EPC [34]. CUPS proposes the separation of the data plane and control

[1]The work presented in this chapter has been published in [52] and was supported by the Ministry of Electronics and Information Technology (MeitY), Government of India through a sponsored project grant on 5G.

plane functionalities of 4G core network elements into separate control and plane entities connected by a well-defined interface. For example, the SGW is truncated into SGW-c and SGW-u representing control and user plane functionalities, respectively. SGW-c and SGW-u are connected through a newly defined interface Sxa. However, the 4G RAN continues to have tightly integrated control and data plane functions. The 5G network architecture as defined by 3GPP is shown in Figure 3.1.

As seen in the figure, the 5G system comprises a core network, i.e., 5GC and an access network. The 5GC consists of network functions that have either control or data plane functions. The functionalities of the various NFs within the core network are summarized below. A detailed list of functionalities is available in [21] and [53].

- **Access and Mobility Management Function (AMF)**: AMF is a control plane function that acts as the termination point for Non Access Stratum (NAS) signaling with the UE. It is also responsible for controlling both NAS and access stratum signaling security. The primary function of AMF is to manage intra and inter-system UE mobility. In order to perform this task, it supports mobility related signaling between 5GC nodes and also performs mobility control in terms of subscription and policies. It also supports network slicing.

- **User Plane Function (UPF)**: UPF is the only network function in the 5GC with forwarding plane functionality. It performs data packet routing and forwarding, reports traffic usage, and handles data plane QoS. It also performs packet inspection and also enforces policies related to the data plane.

- **Session Management Function (SMF)**: SMF is a control plane function in 5GC. It performs session management and configures traffic steering at UPF for routing traffic to the correct destination. It is responsible for enforcing policies on the control plane. It also allocates and manages UE IP addresses.

- **Authentication Server Function (AUSF)**: AUSF supports authentication for 3GPP access and untrusted non-3GPP access.

- **Network Exposure Function (NEF)**: NEF performs tasks such as receiving information from other network functions based on the capabilities exposed by them.

It stores the received information as structured data using a standardized interface to a Unified Data Repository (UDR). This information can also be exposed by NEF to other functions for purposes such as analytics.

- **Network Repository Function (NRF)**: NRF is responsible for supporting service discovery. It receives/provides the information regarding Discovery Request/discovered instances from/to network function instance(s) or the Service Communication Proxy (SCP).

- **Network Slice Selection Function (NSSF)**: NSSF is responsible for selecting the set of network slice instances serving a UE. It determines the allowed Network Slice Selection Assistance Informations (NSSAIs), configured NSSAIs and maps to subscribed S-NSSAIs. It also determines the AMF set to be used to serve a UE, or, based on configuration, a list of candidate AMF(s) by querying the NRF.

- **Policy Control Function (PCF)**: PCF is a control plane function that performs policy related tasks in 5GC. It provides policy rules to control plane functions for enforcement. It supports a unified policy framework to govern network behavior.

- **(UDM)**: Unified Data Management (UDM) is responsible for handling subscription management, user-identification, generation of authentication credentials etc. It also performs tasks related to service/session continuity, e.g., maintaining the information related to SMF assignment for ongoing sessions.

- **Application Function (AF)**: AF interacts with other 5GC network functions to perform application related tasks such as influencing traffic routing, interacting with the policy framework for policy control. Based on deployment, AFs are considered trusted/untrusted by operators, and the interaction mechanism can be direct with the concerned function/indirect via the NEF, respectively.

The 5GC architecture is designed to function as a “service-based architecture” wherein the capabilities/services offered by control plane network functions within the core network are shared to other authorized functions through a service based interface [54]. For example, ‘Namf’ is the service based interface offered by AMF. However, the network functions within 5GC interact with the access network using reference point interfaces.

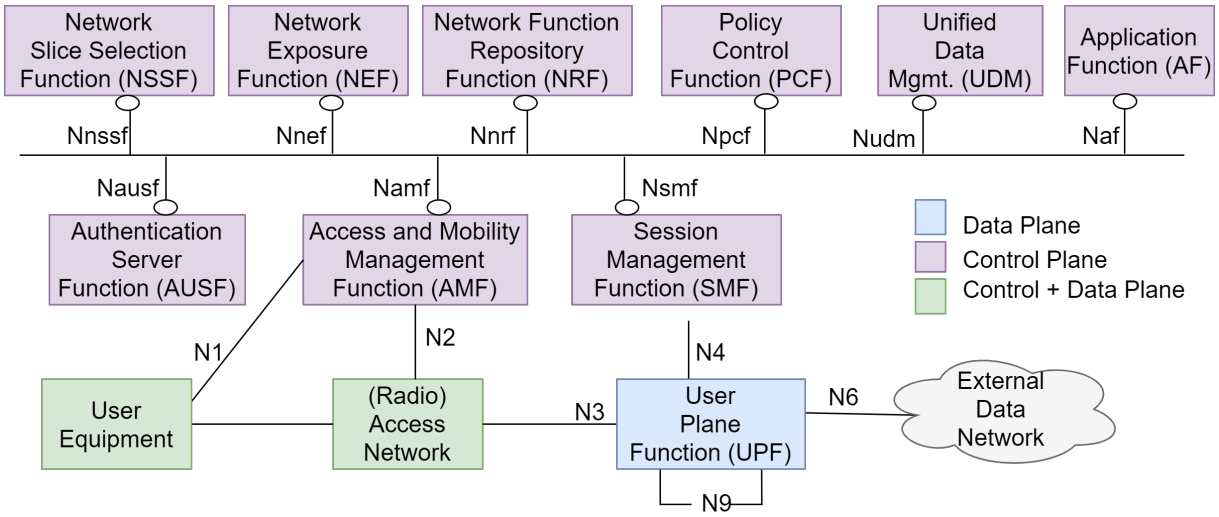


Figure 3.1: 3GPP 5G network architecture (courtesy [21])

For example, control signaling between AMF and access network is communicated over 'N2' interface.

The 3GPP 5G network supports various access connectivity types such as fixed, wireless, and satellite connectivity. The RAN can comprise 3GPP NR base stations i.e., gNBs, evolved LTE base stations e.g., Next generation eNodeBs (ng-eNBs), or WLAN APs with a suitable inter-working function. In this chapter, we limit the discussion of the access network to 5G NR RAN. 3GPP 5G standard supports both collapsed as well as the dis-aggregated architecture for the 5G gNB. In the collapsed architecture, the full protocol stack of the gNB is co-located at a given site. This is typical for non-centralized deployments or co-located deployments with other RATs [55].

In the dis-aggregated architecture type, the gNB is split into two logical nodes, viz., the gNB Central Unit (gNB-CU) and the gNB Distributed Unit (gNB-DU) as shown in Figure 3.2. These nodes are interconnected with one another over a data plane interface known as F1-U and the control plane interface known as F1-C [47]. The F1 Application protocol (F1AP) runs over the F1-C interface. The F1AP is used to carry messages for configuring the gNB-DU. The gNB-CU is further composed of control plane entity gNB-CU Control Plane (gNB-CU-CP) and gNB-CU User Plane (gNB-CU-UP). gNB-CU-CP hosts Radio Resource Control (RRC) and control part of Packet Data Convergence Protocol control (PDCP-C), whereas gNB-CU-UP hosts Service Data Adaptation

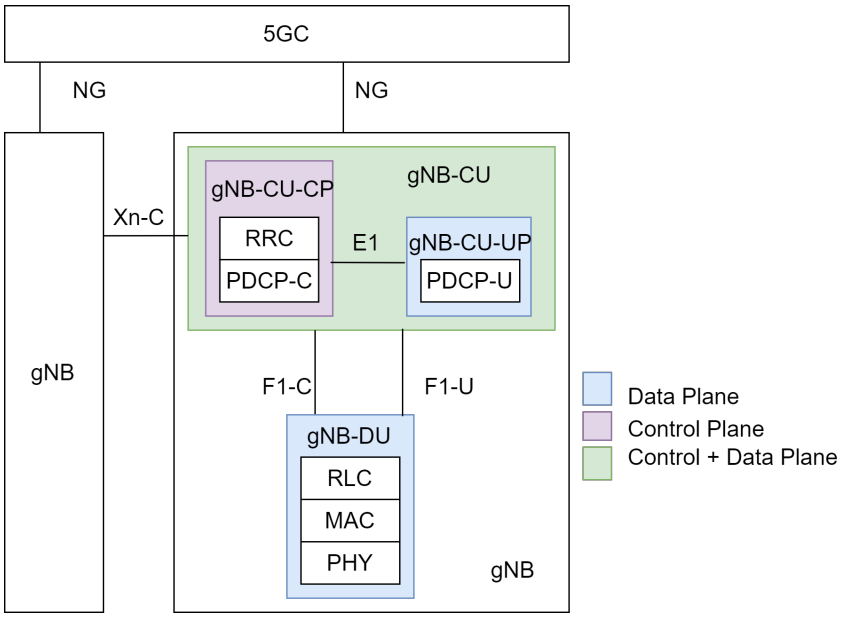


Figure 3.2: 3GPP 5G NR dis-aggregated architecture (courtesy [21])

Protocol (SDAP) along with PDCP User Plane (PDCP-U). Packet Data Convergence Protocol (PDCP) performs functions such as ciphering and deciphering, header compression and decompression, re-ordering and in-order delivery, etc. The functions of SDAP are primarily to perform a mapping between a QoS flow and a data radio bearer. gNB-CU-CP configures gNB-CU-UP through an interface known as the 'E1' interface. The RRC layer, along with the RRM functions, is responsible for the management of connected mode mobility, security keys, enforcement of QoS on the radio interface, radio bearer control, and radio admission control.

gNB-CU controls the operation of one or more gNB-DUs. The gNB-DU consists of Radio Link Control (RLC), MAC, and PHY layers. The functions of RLC may include the transfer of upper later protocol data units, error correction through Automatic Request Repeat (ARQ), sequence numbering, etc. The MAC layer performs functions such as multiplexing/demultiplexing MAC service data units. It performs error correction through Hybrid Automatic Request Repeat (HARQ) and priority handling of UEs. PHY layer performs the actual transmission/reception of data over the air interface. The gNB-CU and gNB-DU, together, appear as a unified logical entity (gNB) to the core network. The signaling procedures of the 5G cellular network are similar to that of LTE, since the standard considers co-deployment scenarios for the LTE EPC and the 5GC. The

5G cellular network provides backward compatibility with the Evolved Terrestrial Radio Access Network (E-UTRAN), by using an enhanced LTE base station, known as the ng-eNB [21].

3.2 Proposed Architecture for Centralized RAN Control

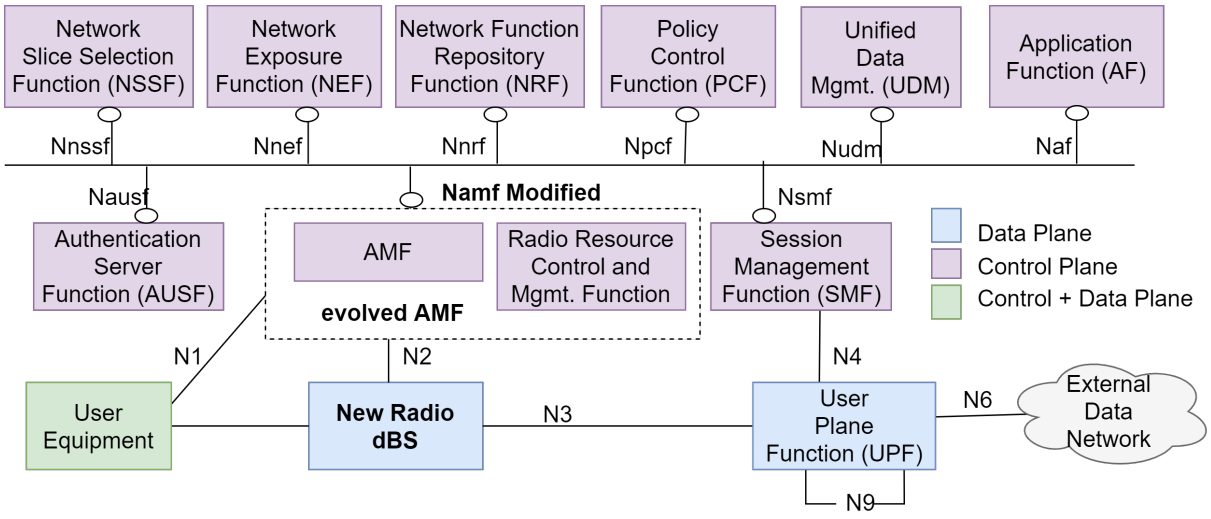


Figure 3.3: Proposed 5G network architecture

Based on the functionality of the 3GPP 5G system introduced earlier, we propose an architecture to centralize RAN control. In this architecture, the control functionality of gNB, i.e., the RRC layer along with the RRM function, is removed from the gNB and placed in the core network. We refer to the new gNB, devoid of control plane functionality and comprising of only data plane functionality, as the NR data plane Base Station (dBS). The RRC protocol layer and RRM functionality, together with the AMF, constitute a new network function located in the core network, hereinafter referred to as the enhanced AMF (eAMF).

In addition, the F1AP, which is used by gNB-CU to configure the gNB-DU in the 3GPP defined architecture, is modified and used by eAMF to control and manage the dBSs. As a result, network control gets centralized, and a well-defined separation between control and data planes in the end-to-end network is achieved. Although we consider the gNB as the reference base station in our architecture, this proposal is also valid for the ng-eNB [21].

At the time of the proposal, newer works such as [56, 57] have been proposed in-addition to the works mentioned in Chapter 2. For example, authors in [56] define a centralized system architecture for efficient resource management in LTE. They suggest decoupling certain key radio resource functionalities, e.g., handover functionality from the eNB and placing them in a centralized SDN controller. However, their proposed framework preserves the control plane interface from the eNB towards the MME intact and, consequently, may not reduce the processing time for control signals. In another work, authors advocate two tier approaches by using separate controllers for the RAN and the core network [57]. The RAN controller is responsible for mobility and interference management, whereas the core network controller regulates routing and policy. In contrast, our proposal provides the benefits of both the approaches, i.e., signaling reduction as well as a single tier approach.

Our proposal also has similarities to the Option (1A like split) considered by 3GPP for division of the protocols between gNB-CU and gNB-DU [22]. In this option, RRC is present in gNB-CU whereas all the lower protocol layers are present in gNB-DU. However, at the time of acceptance of this work for publication in [52], details of the control plane and data plane split within the gNB were not publicly available. Moreover, the relevant 3GPP specification does not provide a detailed analysis of the impact of this split. To the best of our knowledge, our solution was the first work to have proposed and evaluated a centralized SDN controller architecture for the 3GPP 5G network.

3.2.1 Signaling Procedures within the Proposed Architecture

The transposition of RRC and RRM from the RAN into the eAMF results in a leaner protocol stack. This also results in signaling cost reduction across 5GC and RAN. To understand this, we compare the 3GPP 5G control plane protocol stack with that of the proposed architecture. Figure 3.4 depicts the control protocol stack for the 3GPP defined 5G network. As shown in Figure 3.4, the gNB has a UE facing protocol stack consisting of RRC, PDCP, RLC, MAC, and PHY layers. The protocol stack of the gNB that interfaces with the core network consists of the Next Generation Application Protocol (NGAP), Stream Control Transmission Protocol (SCTP), IP, Layer 2 (L2) and Layer 1 (L1) protocols. NGAP is an application layer protocol that provides the signaling service between

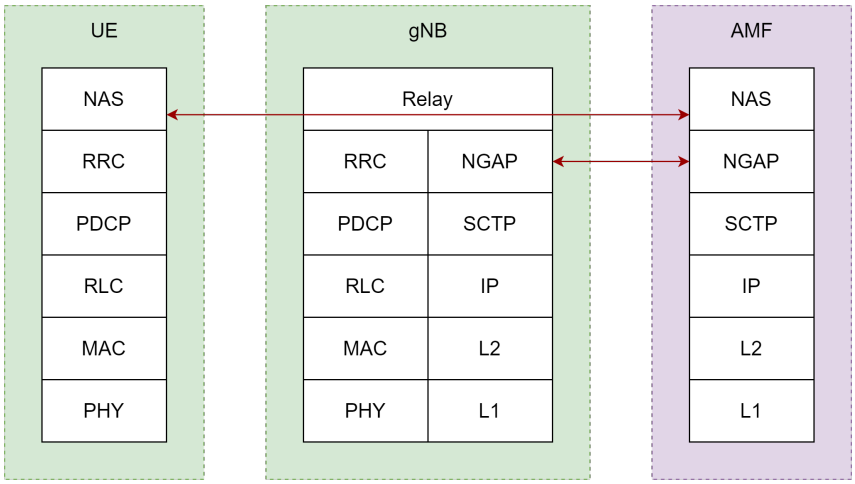


Figure 3.4: 3GPP defined 5G control plane stack (courtesy [21])

gNB and AMF for performing functions such as UE context management, mobility management, NAS transport function, Protocol Data Unit (PDU) session management, etc. SCTP is a transport protocol built on IP transport and is used for the reliable transport of signaling messages [47].

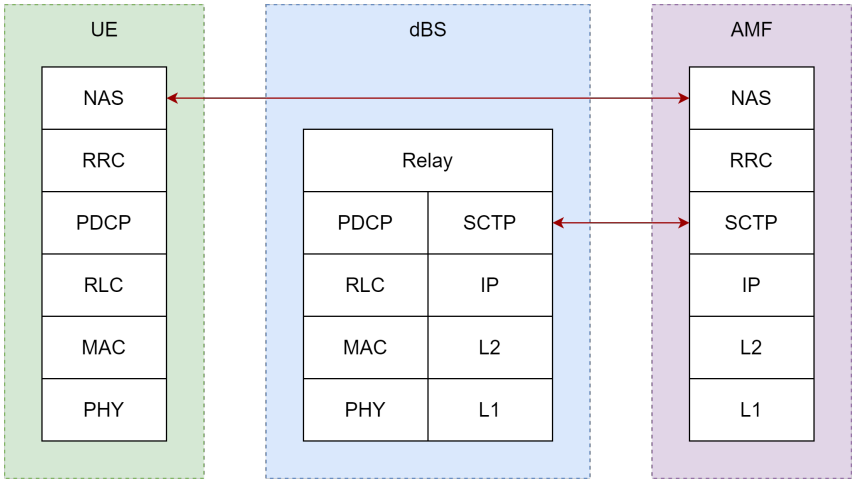


Figure 3.5: Control plane stack for the proposed architecture

In the 3GPP 5G network, the RRC layer, along with the RRM functions in the gNB perform radio resource allocation. Since both gNB and AMF possess control plane functionality, the NGAP is needed for signaling exchanges between gNB and the 5GC, e.g., to carry UE specific signaling. As a result of transposing RRC along with the RRM functionality into the AMF, the control functionality is completely transferred to the core

network. The NGAP is no longer required to carry UE specific signaling between gNB and AMF and can thus be eliminated. The resultant protocol stack for the proposed architecture is shown in Figure 3.5. To demonstrate the advantages of the proposed architecture with respect to the 3GPP defined 5G network, we study the call flows for registration and handover for both the architectures.

Registration is a procedure by which a UE attempts to access the cellular network for the first time. The details of the procedure for 3GPP defined 5G and the proposed networks have been illustrated in Figures 3.6 and 3.7, respectively. In the 3GPP 5G cellular network, the registration procedure mainly involves control message exchanges between UE, gNB, and AMF. NAS is the highest stratum of the control plane between UE and AMF. NAS messages transport signaling related mobility, authentication, session management, etc. UE exchanges NAS messages with AMF by encapsulating them using RRC protocol and transmitting them to the gNB. The gNB decodes the received messages and sends them further to the AMF with the help of NGAP. As a result, every message exchanged between the UE and AMF is processed twice.

In order to distinguish between messages encoded using RRC and NGAP in Figure 3.6, we have shown them as being encoded in RRC and NGAP containers, respectively. Additionally, a few signaling messages are exchanged between the gNB and AMF to setup flow contexts on the gNB for data transfer to a particular UE. On completion of the above signaling exchanges, the data flow may be initiated in the network. The call flow for the registration procedure in the proposed architecture is illustrated in Figure 3.7. Every NAS and RRC message is exchanged between the UE and eAMF via the dBS. These messages are encoded/decoded using the RRC protocol. An additional 'Create Flow' message is introduced. The message is used by eAMF to instruct the dBS to create a new data flow. This message is sent over a modified F1AP to configure the dBS in accordance with the flow requirements. On comparing both call flows, we can make the following observations.

The signaling required for an equivalent procedure in the proposed architecture is considerably reduced.

For example, some of the signaling messages that used to require NGAP encoding,

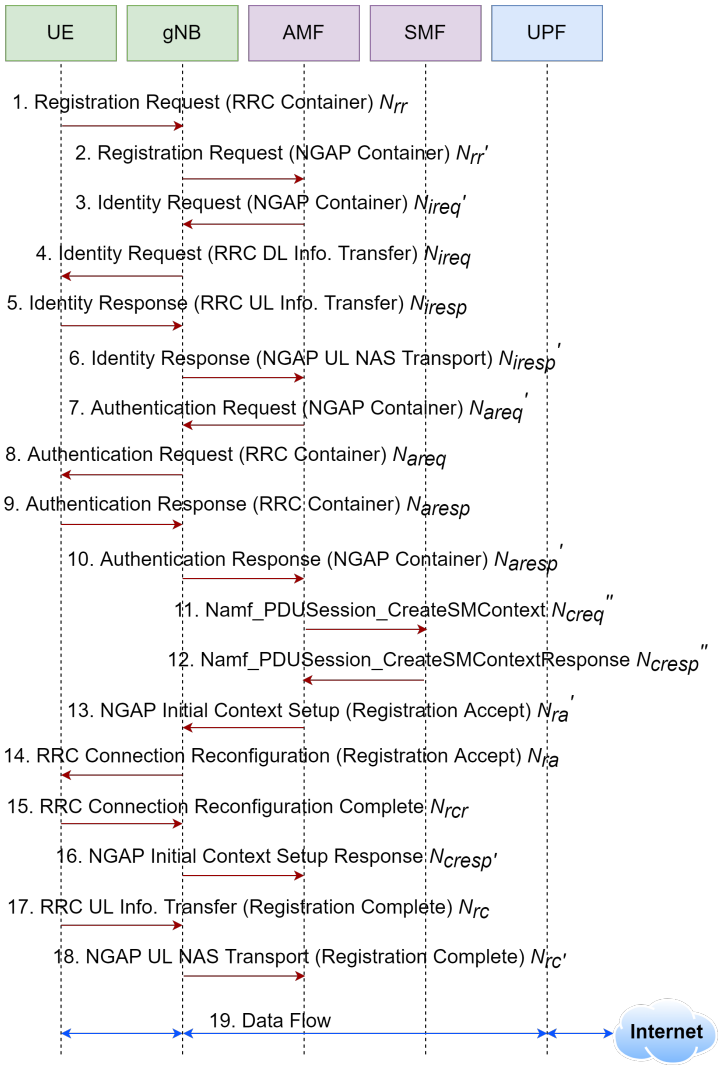


Figure 3.6: Registration procedure for the 3GPP defined 5G architecture (courtesy [21])

e.g., Initial UE message, etc., are no longer necessary due to the removal of NGAP. Moreover, the number of encoding and decoding steps for identity verification, authentication, etc., are reduced, as RRC messages are directly transmitted to eAMF without being processed at dBS. Further, due to the centralization of decision making, handshake messages can be eliminated. For example, an NGAP message 'Initial Context Setup Response' is sent by gNB to AMF in response to the Initial Context Setup Request message in the standard 5G network. Such response messages are no longer required.

Similar conclusions can be obtained for the handover procedure, which has been shown in Figures 3.8 and 3.9, respectively. In the 3GPP defined 5G cellular architecture, the gNB receives RRC measurement reports from a UE and sends Handover Required

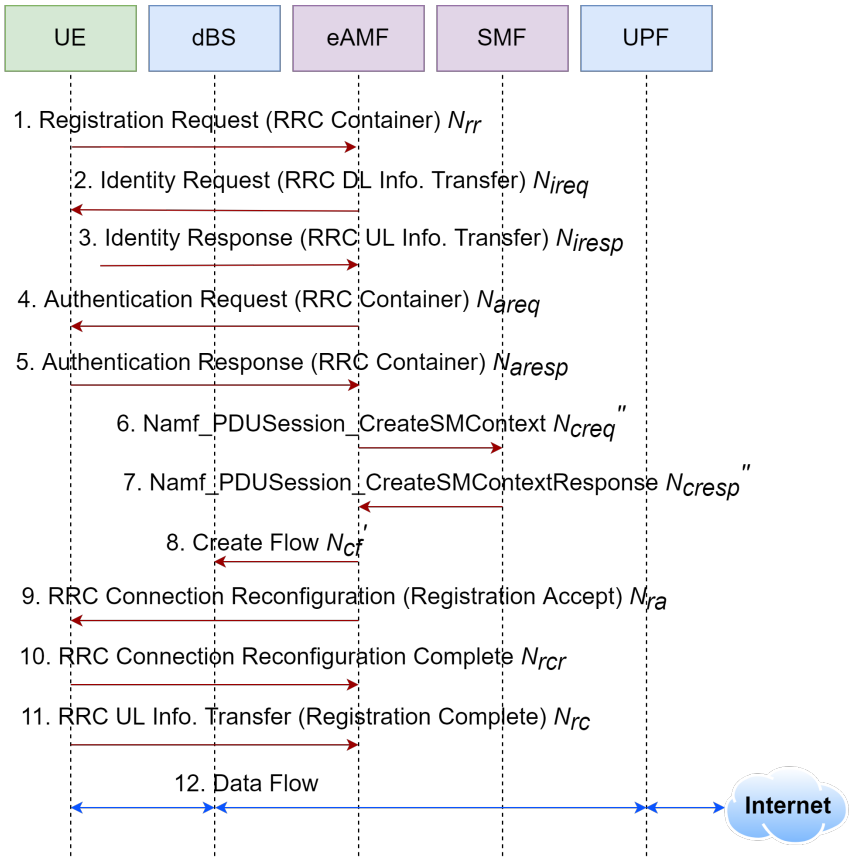


Figure 3.7: Registration procedure for the proposed architecture

message to the AMF for handover initiation, whenever required. The AMF transmits a Handover Request message to the prospective target gNB, which responds with Handover Request Acknowledgement if it is able to admit the UE.

The AMF then issues a Handover Command message to the source gNB to handover the UE to the chosen target. The source gNB sends an RRC Connection Reconfiguration message to the UE to indicate the same. The UE then sends the Handover Confirm message to the target gNB. Following this, the Handover Notify message is sent from the target gNB to the AMF. Once these steps are completed, session setup is carried out in the core network. At the time of publication of this solution, this part of the procedure was under discussion in the 3GPP working group [58]. We have illustrated this step only for the sake of completion, and it does not affect our analysis as the message exchanges for the session setup are within the 5GC and not across the core network and the NG-RAN. After the completion of session setup, the older UE context is released from source gNB.

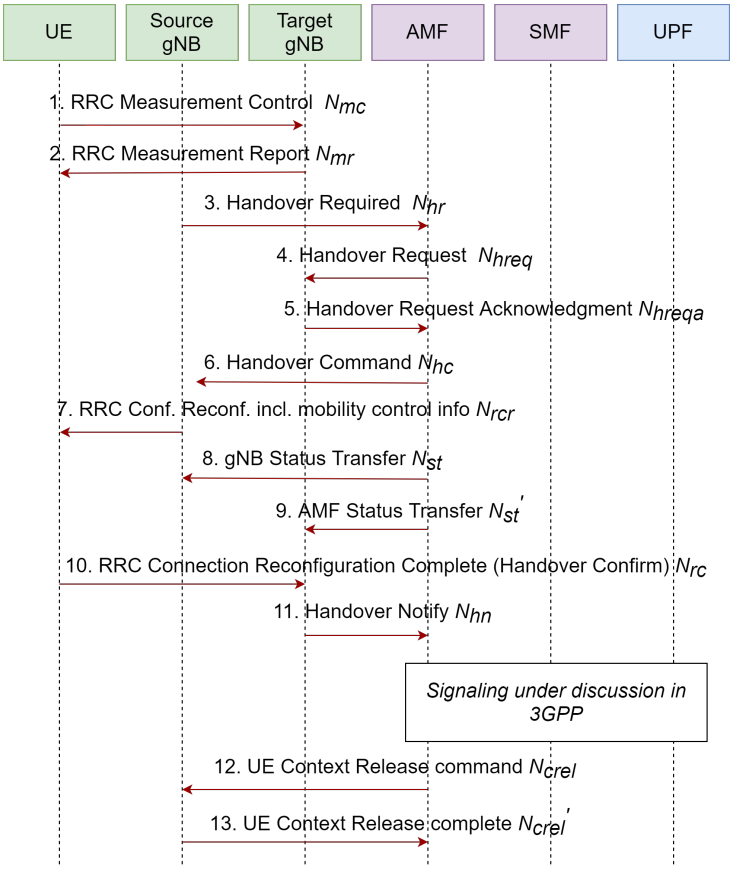


Figure 3.8: Handover in the 3GPP 5G architecture (courtesy [21])

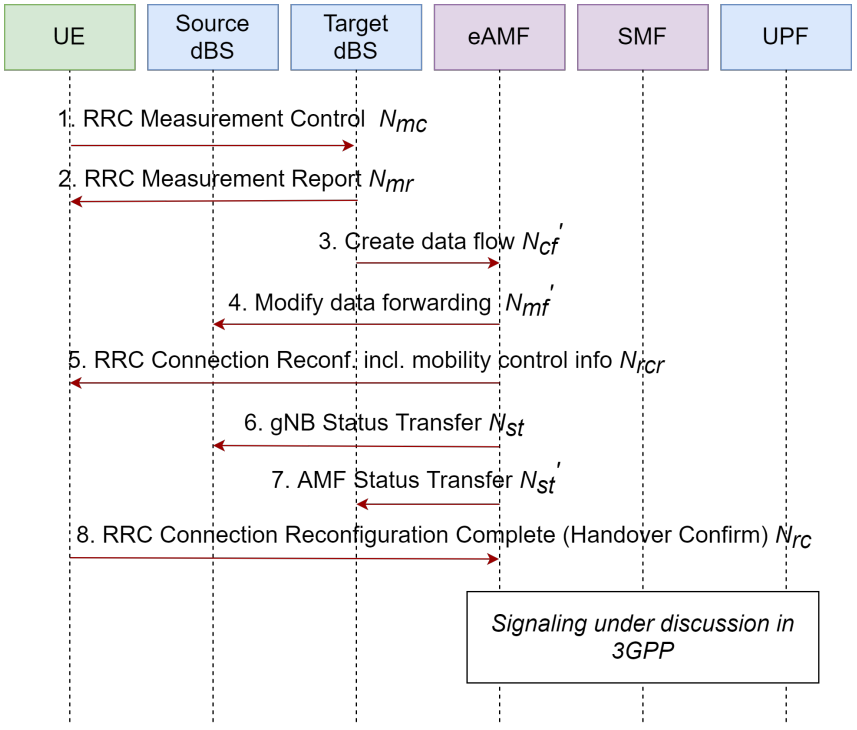


Figure 3.9: Handover in the proposed architecture

In the proposed architecture, all of the measurement reports are sent to the eAMF. The eAMF is responsible for handover decisions and transmits commands for data flow creation and modification to the gNBs, when necessary. The remainder of the call flow remains unchanged with the exception of the UE context release step, which is no longer required as the context is centrally stored in the eAMF. We observe that the handover signaling has been simplified due to the centralization of control.

An important aspect that needs to be highlighted is that the centralization of control may lead to increased signaling load at the eAMF when UE handover and registration occur frequently. However, the proposed architecture can be scaled easily by using multiple eAMFs in the network and distributing the load across them. This is in line with the working of 3GPP 5GC. Note that 3GPP 5GC provides a separation between network state and procedures. For example, the UE context is stored in the UDM and the session state are stored in the SMF [59]. The AMF is only responsible for performing connection and mobility management procedures and multiple AMFs are present in the network. For a given UE, these procedures can be carried out through any AMF interacting with the UDM hosting the UE context. Within the proposed architecture, the eAMF performs similar functions as the AMF. As a result, the proposed architecture can be scaled by distributing the load across multiple eAMFs in the network, when the signaling load due to handover/call establishment increases. To understand the impact of proposed enhancements to 3GPP 5G, we review the benefits that are provided by the proposed architecture.

3.3 Advantages of the Proposed Architecture

The proposed architecture provides several performance improvements, some of which are listed below.

- **Reduction in signaling cost due to the elimination of a protocol layer:** As a result of transposing RRC and RRM into 5GC, NGAP is no longer used for communication between the gNB and AMF. Hence, the signaling cost incurred in encoding and decoding NGAP is eliminated. The reduction in cost is quantified in Section 3.4.

- **Reduced mobility failures and faster handover**: The centralized view of network resources also aids in making better decisions for handover management. In the proposed architecture, the eAMF maintains context for the UEs and hence has access to the overall network state, e.g., traffic load at a given gNB-DU, the signal strength of various cells as observed by the UEs, UE QoS requirements, and data rates, etc. In the 3GPP 5G architecture, mobility decisions are taken at both gNBs and AMF as both UE context, and the decision making abilities are distributed. Centralization of mobility management provides a network-wide view of resources and leads to reduced handover failures as well as faster handover for the UEs.

- **Load balancing and interference management**: The proposed architecture can also facilitate better interference management and load balancing decisions with optimized algorithms, which bring about an increase in the overall system throughput. For example, where traffic distribution is non-uniform, the eAMF can take decisions to handover UEs from heavily loaded cells to the lightly loaded ones. This can be helped by strategies such as dBS transmit power control, cell-offset tuning, etc.

- **UE power saving**: The reduction in time for network access and idle mode mobility procedures result in power savings for UEs as they can now remain in the power saving idle mode for a longer time. This is due to a reduction in the time required for the UE to switch between the idle and active states, as illustrated by the registration call flows.

- **Reduced system costs:** Several studies advocate the placement of computationally intensive network control functions in the datacenter and time-sensitive data plane functions in the network infrastructure closer to the UE [60,61]. Our proposal is in alignment with this thinking and helps in reducing the costs of the gNBs, which can now be replaced with simpler devices having radio functionality.

3.4 Performance Analysis of the Proposed Architecture

We present the performance analysis for the proposed architecture in this section. The signaling cost reductions are quantitatively computed using call flows in Figures 3.6, 3.7, 3.8 and 3.9 as a reference. The call flows illustrate signaling messages that are exchanged

Table 3.1: ASN1 Processing Overhead

ASN1 Messages	**Notation**
RRC decode at gNB of message received from UE	P_{gd}
RRC encode at gNB of message sent to UE	P_{ge}
NGAP encode at gNB of message sent to AMF	$P_{ge'}$
NGAP decode at gNB of message received from AMF	$P_{gd'}$
RRC encode at eAMF of message received from UE	P_{ee}
RRC decode at eAMF of message sent to UE	P_{ed}
F1AP encode at eAMF of message sent to dBS	$P_{e'e}$
F1AP decode at eAMF of message sent from dBS	$P_{e'd}$
F1AP encode at dBS of message sent to eAMF	P_{de}
F1AP decode at dBS of message received from eAMF	P_{dd}

between various network elements during registration and handover procedures. As the initial step for signaling cost compute, we tabulate the processing overheads at a given node for both architectures in Table 3.1.

We consider that all messages have an average length of m bits, as message sizes were not standardized in the 5G specification at the time of publication of this work. Similarly, we denote time taken per bit for message exchange between any two nodes as α and processing time for any node, mentioned in Table 3.1 is equal to β. We also observe that, according to the values provided in [62] for LTE, $\alpha(\approx 1)$ms, $\beta(\approx 4)$ms, and $\alpha < \beta$. In accordance with the above information, we have calculated the time taken for signaling in the 5G architecture as well as our proposed architecture below.

1. Registration signaling cost for the 3GPP 5G architecture

$$\begin{aligned} T_{Attach} &= \alpha(N_{rr}) + 5(P_{gd} + P_{ge'}) + \alpha(Nrr') + \alpha(N_{ireq'}) + 3(P_{gd'} + P_{ge}) \\ &\quad + \alpha(N_{ireq} + N_{iresp} + N_{iresp'}) + \alpha(N_{areq'} + N_{areq} + N_{aresp} + N_{aresp'}) \\ &\quad + \alpha(N_{creq''} + N_{cresp''}) + \alpha(N_{ra'} + N_{ra} + N_{rcr} + N_{cresp'} + N_{rc} + N_{rc'}) \\ &\quad + 5P_{gd} + 3P_{ge}, \\ &= 18m\alpha + 24\beta. \end{aligned} \tag{3.1}$$

2. Registration signaling cost for the proposed architecture

$$\begin{aligned} T_{Attach'} &= 2\alpha(N_{rr} + N_{ireq} + N_{iresp} + N_{areq} + N_{aresp}) + \beta(3P_{ee} + 5P_{ed} + P_{e'e}) \\ &\quad + \alpha(N_{creq''} + N_{cresp''} + N_{cf'}) + 2\alpha(N_{ra} + N_{rcr} + N_{rc}) + P_{dd}, \\ &= 19m\alpha + 10\beta. \end{aligned} \tag{3.2}$$

3. Handover signaling cost for the 3GPP 5G architecture

$$\begin{aligned} T_{Handover} &= \alpha(N_{mc} + N_{mr} + N_{hr} + N_{hreq} + N_{hreqa} + N_{hc}) + \alpha(N_{rcr} + N_{st} + N_{st'} \\ &\quad + N_{rc} + N_{hn} + N_{crel} + N_{crel'}) + +2\alpha(N_{ra} + N_{rcr} + N_{rc}) \\ &\quad + 2P_{gd} + 2P_{ge} + 4(P_{gd'} + P_{ge}) + 5(P_{gd} + P_{ge'}), \\ &= 13m\alpha + 22\beta. \end{aligned} \tag{3.3}$$

4. Handover signaling cost for the proposed architecture

$$\begin{aligned} T_{Handover'} &= 2\alpha(N_{mc} + N_{mr} + N_{rcr} + N_{rc}) + \alpha(N_{cf'} + N_{mf'} + N_{st} + N_{st'}) \\ &\quad 2P_{ee} + 2P_{ed} + 3(P_{e'e} + P_{dd}) + P_{e'd} + P_{de}, \\ &= 12m\alpha + 12\beta. \end{aligned} \tag{3.4}$$

We observe that the registration and handover times are lower for the proposed architecture in comparison with the standard 5G cellular architecture, mainly due to the reduction in processing cost for encoding and decoding of packet headers. This would depict further improvement if the processing is moved to the core datacenter instead of the less powerful gNBs that are present in the field. Note that the proposed architecture provides the benefits of architectures such as C-RAN without the need for strict fronthaul requirements. The above observation can be quantified by using the values for α and β for LTE due to its similarity with the 5G cellular network and the non-availability of the values for the 5G cellular wireless system. Using [62], we have tabulated the calculated values of the KPIs in Table 3.2. The maximum time taken for the registration procedure for 3GPP 5G and the proposed network are obtained as 84 ms and 60 ms, respectively. Note that the registration time can vary between 74 ms to 84 ms as per the 3GPP specifications. Similarly, the time for handover procedure for standard 5G and the proposed network are 78.5 ms and 55.5 ms, respectively. We observe that the registration time is reduced by

12% to 28%, and the handover time is reduced by 29.29%.

Moreover from estimates provided in [62] for LTE, we can observe that the round trip propagation delay is limited between $4-30$ ms whereas the processing delay is larger and ranges from $47.5-60$ ms based on the configuration. 3GPP specifications also provide a base estimate of propagation delay over the air interface as 1 ms for a distance of upto 300 km [63], which is considerably lesser than the processing delay. Therefore, we can infer that the estimates in Table 3.2 remain valid for even the UEs that are faraway from the central control entity (i.e., the eAMF).

Table 3.2: Evaluated reduction in signaling time

System KPI	**3GPP 5G Architecture**	**Proposed 5G Architecture**	**Improvement**
Registration Time	$74-84$ ms	60 ms	$12\%-28\%$
Handover Time	78.5 ms	55.5 ms	29.29%

3.4.1 Simulation Setup

The comparative performance of 3GPP 5G architecture versus the proposed architecture is evaluated using the following simulation setup. For evaluation we carry out simulations with the help of the ns-3 4G LENA module [28], as there are no tools available for 5G architecture simulation at present. The performance measurements provided by the architecture are compared with the help of two scenarios. In the first scenario, we measure the reduction in signaling cost by evaluating attach times for both architectures. In the second scenario, we attempt to quantify the performance improvements in terms of system throughput due to centralized handover management in the proposed SDN based architecture. In order to perform simulations for the above scenarios, we implemented the following configuration changes/additions to the ns-3 LENA framework.

- **Changes to simulator configuration**: In order to perform measurements for attach time in the proposed architecture versus the traditional architecture, the mode of the ns-3 simulator was changed to "real-time simulation" mode for obtaining the actual system time required for processing.
- **Additions to the framework**: For evaluating the centralized handover manage-

Table 3.3: Simulation parameters for mobility management scenario

Parameter	**Value**
Path loss	$128.1 + 37.6 \log(R)$, R in kms
Tx power for LTE dBS/eNB	46 dBm
Bandwidth of the LTE dBS/eNB	10 MHz
Tx power for UE	23 dBm
LTE dBS/eNB Antenna Type	Isotropic Antenna
LTE UE speed	20 m/s
Number of LTE dBSs/eNBs	3
Number of UEs	32
Mobility Model	Constant velocity mobility model

ment, the number of UEs attached to every cell at the instant that a handover was triggered were measured. This measurement was performed using a custom callback function triggered by the 'HandoverStart' trace. The load information at each cell was used along with the RSRP information provided by the handover traces within ns-3 to perform the handover.

3.4.2 Simulation Results

As described earlier, the performance improvements provided by the proposed 5G architecture in comparison to 3GPP 5G architecture is quantified with the help of two scenarios. At first, we validate the signaling cost improvement for SDN vis-a-vis traditional LTE by measuring comparative times taken for the attach procedure (in place of Registration). We then quantify the improvement in the system throughput due to the use of a centralized algorithm for mobility management in place of traditional distributed algorithms.

3.4.2.1 Attach time evaluation

We use the real-time simulation mode of the ns-3 simulator for evaluating attach time. This evaluation was performed for a single UE by measuring the time for every signal exchanged during the attach procedure. The average attach times for the 3GPP 5G and the proposed network are observed to be 3.23 ms and 2.94 ms, respectively. From these

estimates, we can observe that the signaling time is reduced by 10%. Note that the simulator implements the S1-C interface as an abstraction. Moreover, delays due to the air interface processing are also not taken into account. As a result, the measured times are scaled down in comparison with the real world estimates but the relative performance gain remains the same. As described in the previous section, we can infer that attach time is reduced due to the reduction in processing time used for encoding and decoding.

3.4.2.2 Improved mobility management

As mentioned before, the centralized of RAN control allows for better mobility management. In order to illustrate this, we consider a scenario with three Macro eNBs, which are experiencing unequal traffic loads. Each of the eNBs has a bandwidth of 5 Mhz and is transmitting at 46 dBm. As shown in the Figure 3.10, eNB1 and eNB2 are closer to each other with a distance of 400 m and are heavily loaded. eNB3 is 500 m away from eNB1, and 300 m from eNB2. It is lightly loaded. We consider a Lognormal pathloss model in the simulation. Consider a vehicular user with a 2 Mbps connection, moving away from eNB1 towards eNB2 with a speed of 20 m/s. The parameters used for the simulation are presented in Table 3.3.

In the traditional X2-based A3 Reference Signal Received Power (RSRP) algorithm that is used in today's networks, the mobile user is handed over to eNB2 as the user received signal strength from the eNB2 is the highest. This algorithm runs in a distributed fashion and does not possess load information for all the eNBs in the network. As a result, as more and more users move away from the coverage of eNB1, they are still handed over to eNB2, and the overall system throughput starts deteriorating as the eNB becomes more and more loaded. However, the proposed SDN based architecture provides a global view of the network resources. This allows the usage of a centralized algorithm, wherein the eNB load information along with the RSRP experienced by the UE is available for the SDN controller.

This information can be used to make better mobility management decisions. For example, in this case, when UEs are reporting similar RSRPs from one or more eNBs, they can be handed over to the eNB with the lightest load. As a result, overall system

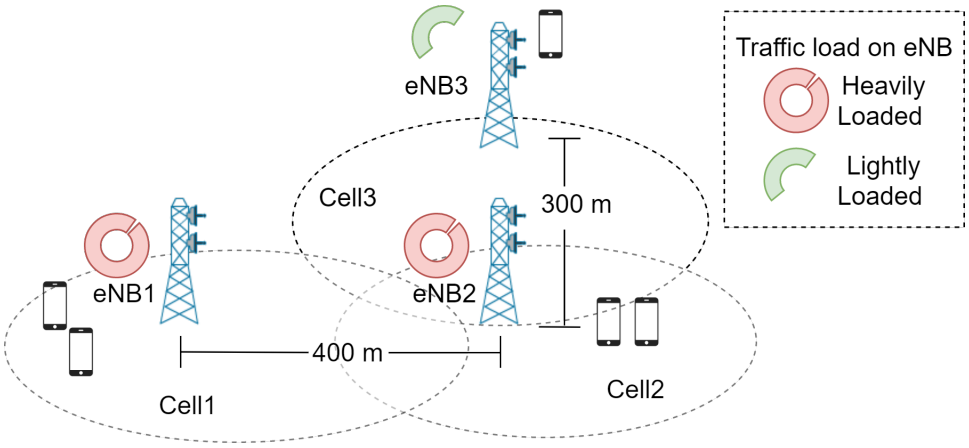


Figure 3.10: Example deployment scenario

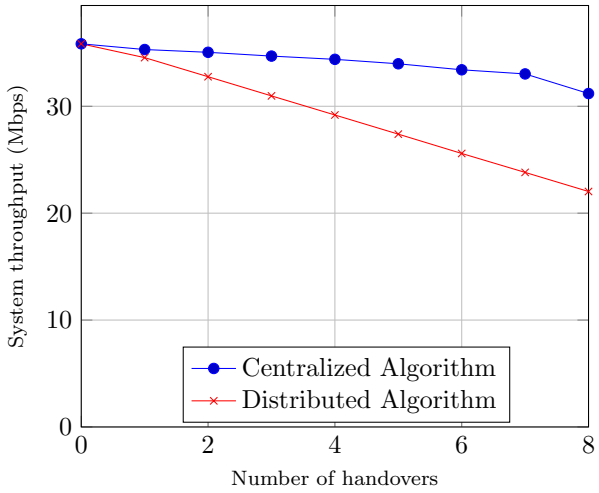


Figure 3.11: System throughput comparison for load balancing in centralized SDN versus traditional distributed LTE architectures

throughput is better due to improved load balancing, as illustrated in Figure 3.11.

Note that although we have considered a simple network with 3 cells for the evaluation, performance results for a larger network are also expected to provide similar benefits in-comparison to the present-day distributed RAN architecture. This is due to the fact even with a large network, the number of candidate target cells for UE handover are only limited to its neighboring cells (i.e., a maximum of 6 cells). Therefore the complexity of the algorithm at the centralized controller is not expected to increase greatly and it would be possible to achieve improvements in system performance.

3.5 Conclusion

In this chapter, we have proposed enhancements to the 3GPP 5G networks to centralize the RAN control functionality in line with the SDN paradigm. We have demonstrated that the movement of RRC functionality together with RRM into the core network reduces the signaling cost between the Next Generation Radio Access Network (NG-RAN) and the 5GC. Additionally, it also centralizes the control of radio resources which results in better decision making at the modified core network function (i.e., eAMF) due to the network-wide view. The elimination of the NGAP layer due to the displacement of RRC protocol from the gNB results in the reduction of processing time required for encoding and decoding of header data. We have evaluated the improvement in latency for control plane procedures, i.e., registration and handover through performance analysis for both the procedures and simulations for attach time while validating that centralization of the RRC layer and RRM functions leads to better system throughput due to improved mobility management in a dynamic environment.

However, in order to be able to perform control of multi-RAT networks comprising of LTE and WLAN, the design needs to be suitably modified. In addition, the enhancements would need to include mechanisms to render it suitable for use in multi-RAT networks that support network slicing. These considerations form the motivation for the next chapter, wherein this architecture is extended to support multi-RAT network control.

Chapter 4

SDN based Architecture for Multi-RAT Network Control

Present-day network deployments comprise a multitude of RATs inter-working with each other. As a result, the logical centralization of RAN control across RATs is necessary for optimal operation. At present, each constituent RAT within a multi-RAT network is controlled by one or more RAT-specific entities. For instance, an 'access controller' is responsible for the control and management of APs within WLAN. Similarly, entities such as eNB and MME control the 4G LTE network. Even within the 5G network, although the core network is common, radio access related decisions are taken separately within individual RATs [21]. Consequently, the control plane within a multi-RAT network is fragmented. This fragmentation prohibits a global view of the network resources, which hinders optimized allocation of network resources [64]. To resolve this, we have addressed the absence of logical centralization in 5G NR RAN control in the previous chapter. In this chapter, we extend the solution and propose a unified framework for end-to-end multi-RAT network control while ensuring scalability[1].

Researchers have attempted to solve this problem using varied approaches. In addition to the prior art described in Chapter 2, authors in [43, 67, 68] propose tiered ar-

[1]The work presented in this chapter was and was published in [65] and patented in [66]. It was supported by the Department of Telecommunications, Ministry of Communications, India as part of the indigenous 5G Test-bed project. This work has been done in collaboration with Arghyadip Roy.

chitectures for the control of multi-RAT networks. Authors in [43] propose a two-tiered approach for multi-RAT RAN control where control and management tasks related to mobility, resource allocation, and interference are handled by the core cloud whereas the edge cloud controls RAN functions. In another work [67], authors present a three-tiered architecture comprising physical infrastructure, control, and management layers. Here, the control layer performs tasks such as mobility and handover management, power control setting, access selection setting, and content management. The management layer is allocated tasks including network element discovery and monitoring, life-cycle management, and backhaul management. A three-tiered architecture for LTE and WLAN control is also presented in [68]. Here, the authors implement a mechanism for storing user plane related information and states such as UE context, etc., for the multi-RAT network in a central storage. However, all of the above solutions do not support network slicing. To the best of our knowledge, the work in this chapter is one of the first attempts to develop a unified framework for end-to-end multi-RAT network control while ensuring scalability. The working of the proposed SDN based architecture is described in the next section.

4.1 Proposed Multi-RAT Network Architecture

We propose an architecture for SDN based multi-RAT control based on the design presented in Chapter 3 while incorporating support for network slicing. Although desirable for ease of control and management, centralized control may give rise to scalability issues in large networks. To preserve scalability, we use a network slice orchestrator to split the end-to-end physical network into multiple logical networks (or network slices) based on service requirements. Network slicing, with a controller for each network slice, also brings scalability to the architecture. Each slice may comprise data plane nodes along with an associated control plane entity known as the multi-RAT controller. This controller manages the data plane nodes in a unified manner. The proposed architecture provides a framework for the deployment of RAT-agnostic control applications. It also has the flexibility to support other RATs in the future.

The system diagram of the proposed architecture is illustrated in Figure 4.1. The network comprises control entities such as multi-RAT controllers and the network slice orchestrator. The network slice orchestrator creates one or more slices on top of the

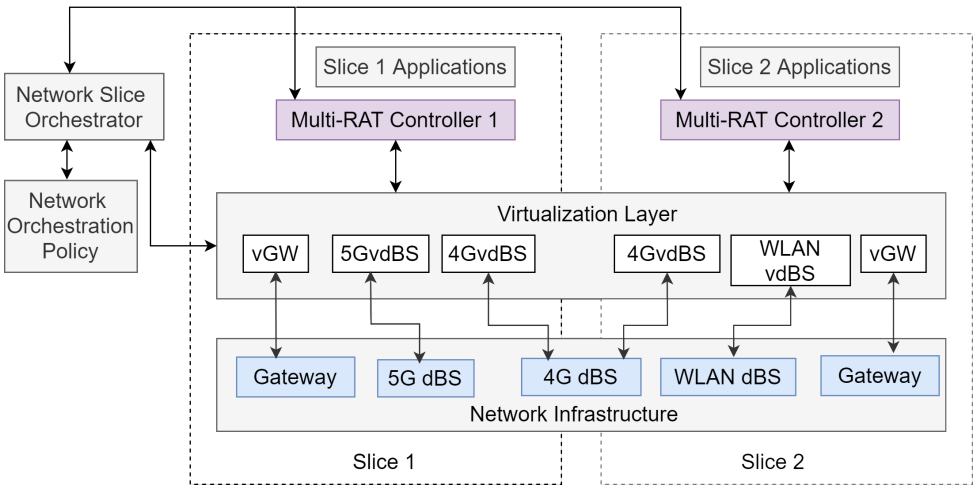


Figure 4.1: Block diagram of the proposed network architecture

network infrastructure based on the network orchestration policy.

This results in the creation of end-to-end logical networks or network slices by grouping a set of network resources based on service requirements while isolating them from each other. Each slice consists of a subset of resources from different data plane entities. It may also comprise a control plane entity that controls these resources, i.e., multi-RAT SDN controller. The size of a given network slice can be increased or decreased by regrouping the physical resources.

The data plane entities within the network include dBSs and Gateways (GWs). The dBSs are RAT-specific data plane entities that are created by eliminating the control functionalities such as radio resource control and management, mobility management from the respective RAT-specific base stations. They typically comprise forwarding plane functionality of the base station together with an optional virtualization layer. The virtualization layer provides an abstract resource view to the controller, which can be based on a virtualization policy. This view comprises virtual data plane entities which are managed by the controller. For example, as illustrated in Figure 4.2, an LTE dBS consists of only the forwarding plane of the base station viz., PDCP, RLC, MAC, PHY and an optional virtualization layer. The UE-specific dedicated (radio) resource control functionality is moved out of the base stations and placed in the controller. dBSs are responsible for carrying UE-specific control and data.

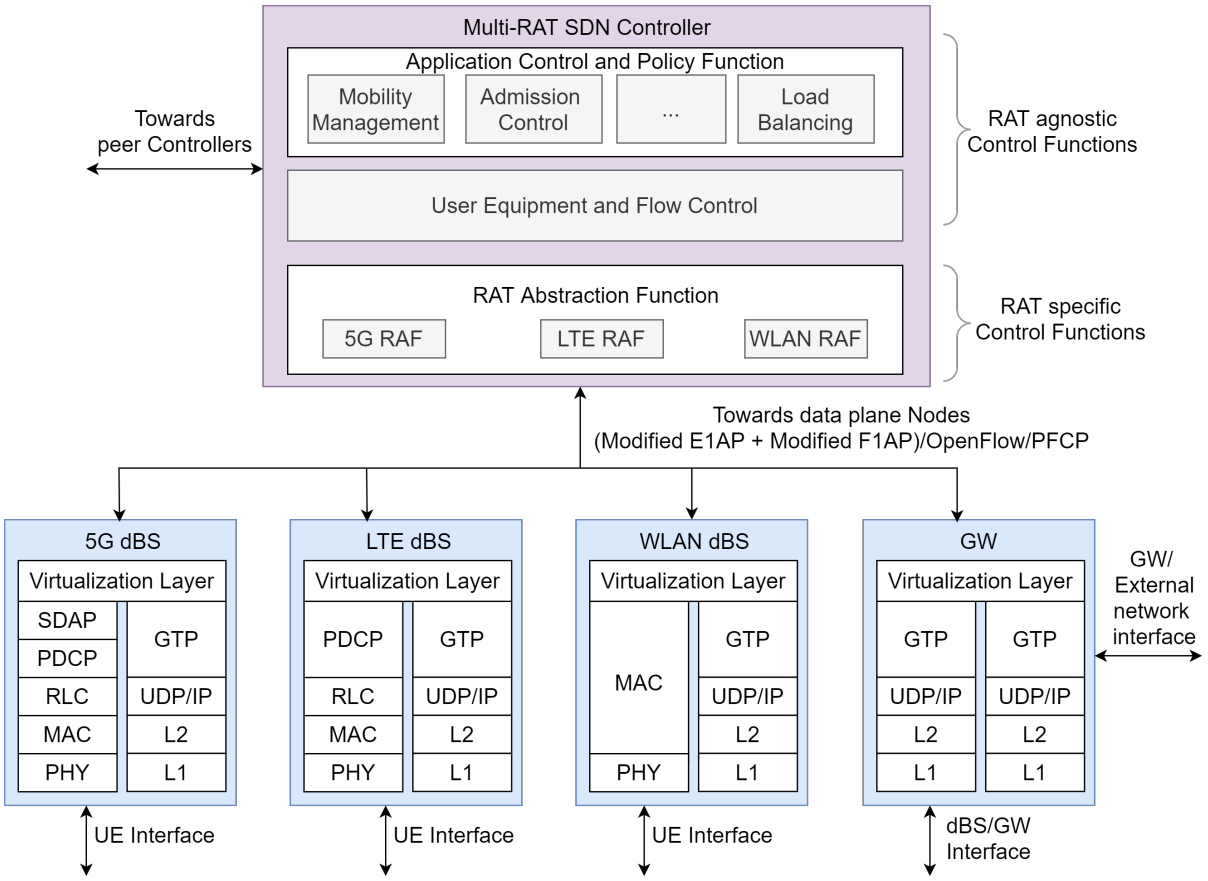


Figure 4.2: Proposed multi-RAT controller architecture and interfaces

Control messages are exchanged with the slice-specific controller, and data is forwarded to the core network according to the path configured by the controller. The cell RRC functionality, which is common to UEs within the cell, e.g., provision of configuration for cell broadcast, is pushed to the virtualization layer. The virtualization layer on an LTE dBS can manifest multiple virtual dBSs (vdBSs) on top of a single physical dBS. This is achieved by partitioning the PRBs available at the LTE dBS and allocating them to the individual virtual dBSs. Similarly, WLAN dBSs may consist of PHY and MAC layers, along with the virtualization layer.

GWs are generic data plane nodes responsible for forwarding user plane data towards other GWs/external data networks on the uplink and other GWs/dBSs on the downlink. A GW supports data forwarding for all types of UEs and all sorts of RATs. The virtualization layer may also be present at the GWs, where it manifests Virtual Gateways (vGWs).

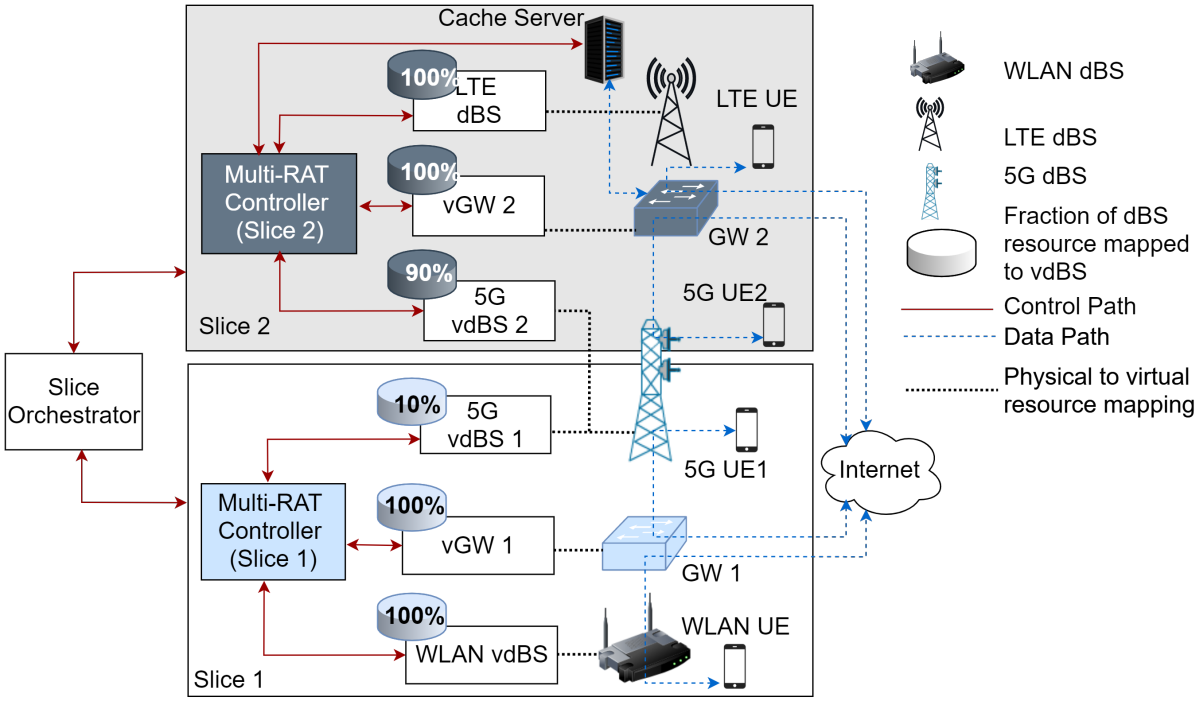


Figure 4.3: An example deployment within the proposed architecture

The virtualization layer can also be deployed within the network as a separate entity between the controller and the data plane nodes. Each of these virtual data plane entities, e.g., vdBSs or vGWs is a part of a network slice or a logical network. dBSs are responsible for forwarding user plane data exchanged between UEs and external data networks either directly (e.g., when connected to a local cache server) or via the GWs. They are also responsible for forwarding signaling/control plane messages exchanged between the UEs and the controller.

Each slice is mapped to a particular multi-RAT controller. The controller controls and manages the virtual data plane entities within the slice and provides data flow configurations to them. It is also responsible for exchanging control plane messages with the UEs. It may also exchange control plane messages with controllers, which are a part of other slices. Communication between slice controllers may be required when a single UE is communicating over multiple slices.

An example deployment of the proposed architecture is illustrated in Figure 4.3. The multi-RAT network illustrated in the figure comprises 5G, LTE, and WLAN RATs. As shown in the figure, the network supports two slices, viz., 'slice 1' and 'slice 2'. The slice

orchestrator distributes network resources across slices as per the specified policy. In this example, 5G dBS is split into two virtual dBSs with 90% of its resources allocated to 5G vdBS1 and the remaining 10% of the resources to 5G vdBS2. Each of these vdBSs is then allocated to different slices. In this example, slice 1 and slice 2 have individual controllers which control the vdBSs and vGWs belonging to that slice. However, controllers may also be shared across slices. Each slice may be governed by specific policies for resource management. Note that the proposed architecture is scalable due to the presence of multiple slices and controllers. The architecture and working of the multi-RAT controller is explained in the next section.

4.1.1 Multi-RAT Controller Architecture

Figure 4.2 illustrates the architecture of the multi-RAT SDN controller and the data plane nodes within the network. To control multiple RATs in a unified manner, functions such as UE authentication, UE mobility management, and flow control can be handled in a RAT-agnostic manner. As a result, the controller comprises functionalities for providing RAT-agnostic control and RAT-specific control. These functions are listed below.

- **RAT Abstraction Function (RAF)**: This function is responsible for handling RAT-specific functionalities within the network. There may exist a separate RAF for every supported RAT. It also manages RAT-specific control plane communication with the UEs. RAF possesses both management and control functionalities and is used to translate generic configuration provided by higher layer functions into RAT-specific configuration to be supplied to a dBS. For example, the 3GPP LTE RAF translates generic flow configuration parameters provided by the layer above into radio bearer parameters to be supplied to an LTE dBS. It also supports the RAT specific NAS and RRC layers. These layers are responsible for signaling message exchanges with the UE. The rest of the controller modules are RAT-agnostic.

- **UE and Flow Control Function (UFCF)**: UFCF maintains the context for every UE and associates flow(s) with a UE. It is responsible for setting up/handover of flows on dBSs and GWs with the desired QoS requirements. UFCF also provides a RAT-independent interface to the layer above, which may contain RAT-agnostic

control algorithms. UFCF maintains a unified list of abstract attributes such as QoS parameters, UE ID for each connected UE, and its associated data flows.

- **Application Control and Policy Function (ACPF)**: ACPF comprises slice-specific control and policy applications. Operators can introduce new application-s/policies, e.g., admission control, load balancing, etc., into a specific slice without affecting other network slices. A RAT-independent interface between the ACPF and the UFCF enables third-party vendors to implement new algorithms.

The southbound interface at the controller can be specified using various protocols that are used to configure the data plane nodes. For example, a modified version of the OpenFlow protocol [16] can be used to configure the GW. The 3GPP Packet Forwarding Control Protocol (PFCP) [69] may also be used in place of OpenFlow. Similarly, modified versions of E1 Application protocol (E1AP) [70] and F1AP [71] can be used to configure dBSs of 5G, LTE, and WLAN RATs.

4.1.2 Signaling Procedures within the Proposed Network

The architecture of the proposed system is designed such that signaling message exchanges between UE and the proposed system are similar to the exchanges experienced by the UE in the present-day network. This allows for the migration of the network to the proposed architecture without requiring changes in UE. An example call flow for handover within the proposed architecture is illustrated in Figure 4.4. The call flow uses 5G NR, 4G LTE, and WLAN as the reference RATs. The signaling for 5G NR and 4G LTE is similar to a large extent. Hence, we use a common representation for 4G and 5G dBSs as '4G/5G dBS' for the ease of illustration. The decision to perform handover for a UE is taken by the mobility management function of the ACPF within the controller. The measurement reports from the UE are forwarded to the controller to assist in the handover decision. A handover command is sent by the multi-RAT controller to initiate the handover. After the handover, UE is associated with the 4G/5G dBS. Since the UE context is maintained at the controller, re-authentication may not be required. As a result, the decision-making at multiple individual nodes such as the source and target dBSs, as done in the existing wireless networks, is no longer needed.

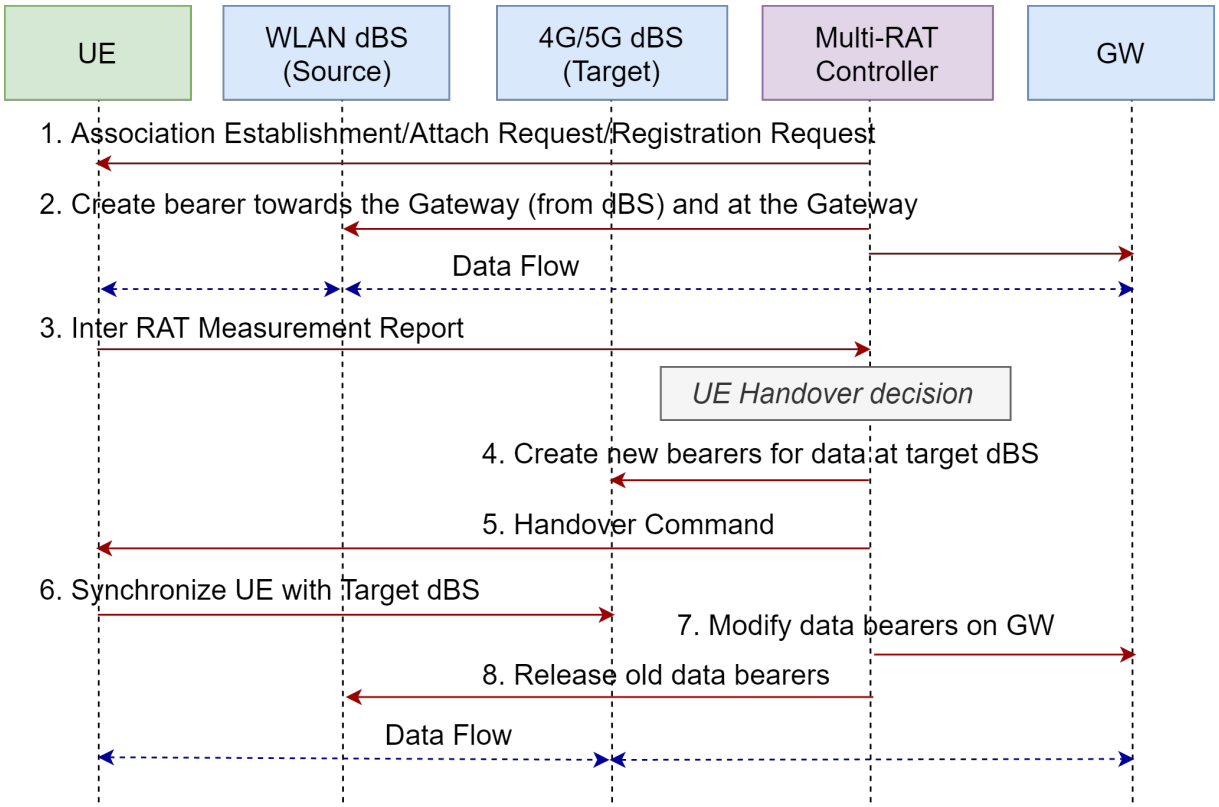


Figure 4.4: UE handover (WLAN to 3GPP 4G/5G) call flow within the proposed architecture

Unlike in today's networks, multiple handshaking signaling exchanges that are required between the dBSs and the core network used for choosing the target dBS can be eliminated in the proposed architecture. This results in reduced control signaling cost as well as reduced latency. Moreover, due to the availability of additional information such as load information at the central entity (i.e., SDN Controller) more intelligent algorithms that use information such as the present load on a cell in addition to the UE radio measurements may be used to perform the handover. This information can be easily obtained for a practical 5G communication system comprising of LTE, NR and WLAN RATs, over the management interface. For example, within LTE 'Average number of active UEs' information element specified in [72] may be used. Similarly for 5G networks, the 'number of active UEs per cell' can be obtained as part of performance measurements over the management interference [73].

Note that the call flows introduced in Chapter 3 for 5G registration and handover (Figure 3.7 and Figure 3.9, respectively) are applicable for 5G RAT control in this design.

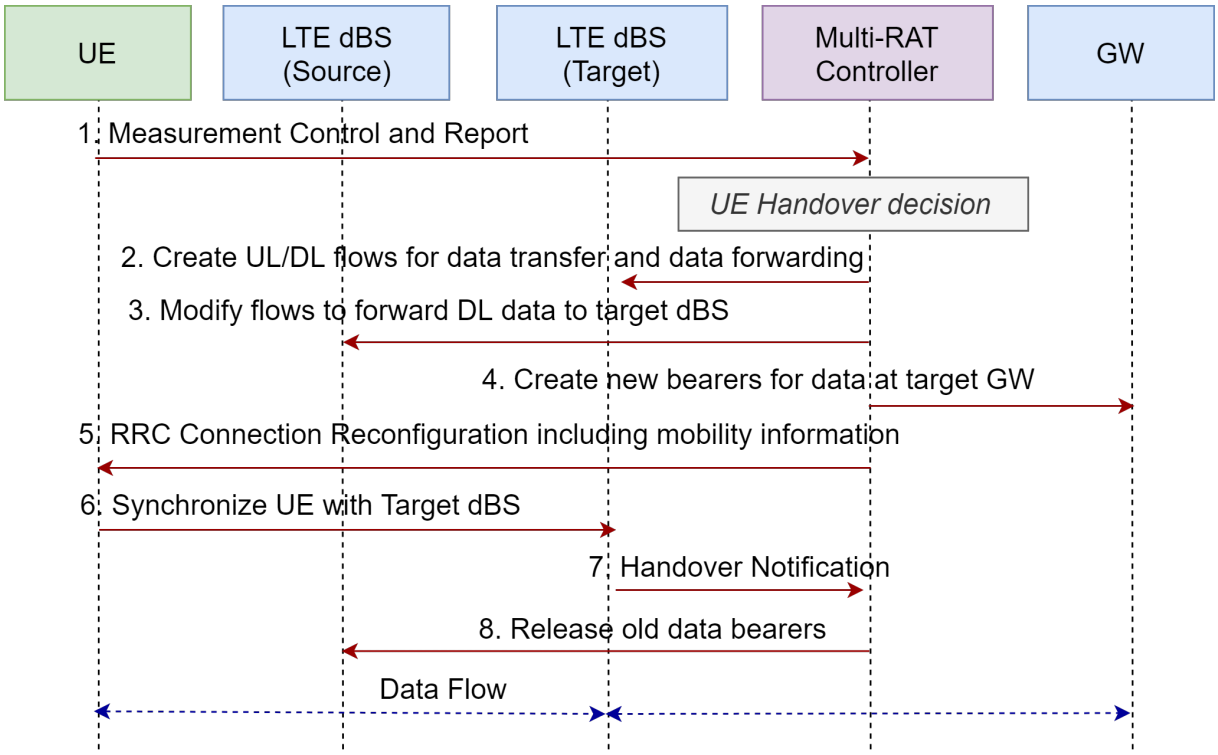


Figure 4.5: Call flow for UE handover within LTE RAT in the proposed architecture

The procedures for Handover and Registration within 5G/LTE RAT would be similar and can be obtained by replacing 'eAMF' with the proposed multi-RAT controller.

Similarly, handover within LTE RAT within the proposed architecture can be carried out using the steps illustrated in Figure 4.5. As illustrated in the figure, measurement report sent by a UE is forwarded to the multi-RAT controller. The decision to perform handover for a UE is taken by the mobility management function of the ACPF within the controller. Once the target LTE dBS and GWs are identified by the Controller, it configures rules to enable data forwarding towards the target dBS and the GW. These rules are configured at the source dBS, target dBS and the GW(s) involved in the handover. On completing the path setup, the multi-RAT controller sends an 'RRC Connection Reconfiguration' message to the UE. On receipt of this message, the UE attempts to synchronize with the target dBS. If the synchronization is successful, the target dBS notifies the multi-RAT controller regarding the completion of the handover through 'Handover Notification' message. The controller then releases the data bearers that were present on the source dBS.

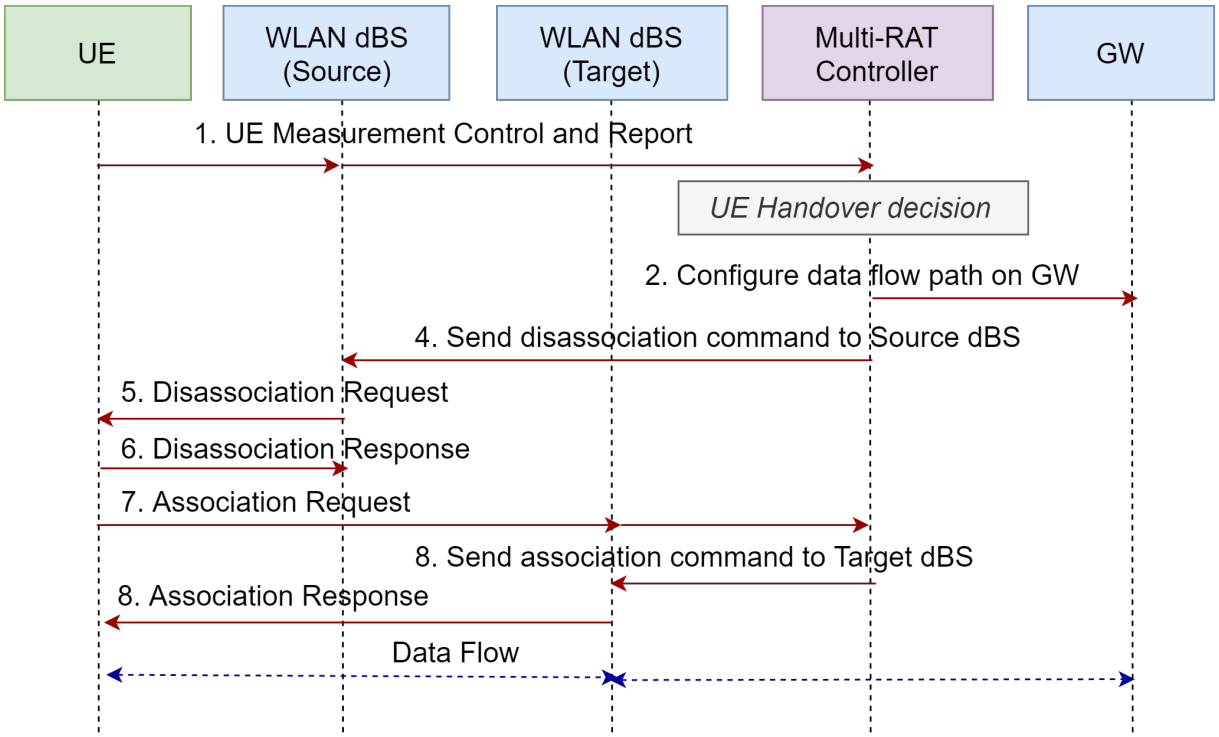


Figure 4.6: Call flow for UE handover within WLAN RAT in the proposed architecture

The procedure for handover within WLAN is illustrated in Figure 4.6. Similar to the case in LTE, measurement report from the UE is forwarded to the multi-RAT controller. Once the handover decision is taken by the controller, it sends a 'Dissociation command' to the source WLAN dBS. On receiving this command the WLAN dBS sends a 'Disassociation Request' message to the UE. This is due to the fact that handovers within WLAN are carried out by first disassociating the UE from the source dBS and then creating a new association with the target dBS. The UE then tries to associate with a another BS. The association request from the UE is forwarded to the multi-RAT Controller through the recipient BS. If the request was forwarded from the dBS that was chosen as the target dBS, the controller responds by sending an 'Association command' with a positive response. The WLAN dBS then accepts the connection by sending an 'Association response' and completes the handover procedure. If the 'Association Request' message was received from a BS other than the chosen target BS, the request is rejected and the UE would try to associate with other suitable BSs until its request is accepted.

4.1.3 Advantages of the Proposed Architecture

In comparison to the present-day network architecture, we can obtain the following advantages from the proposed architecture.

- **Unified authentication and security**: The authentication and security procedures are handled by the controller. Authentication, which is carried out in a unified manner, prevents the need for authenticating the UE every time it connects to a different RAT. This also enables seamless handovers.

- **Simplified signaling procedures**: Procedures that require coordination between multiple entities both within a RAT (e.g., intra-RAT handover) and across multiple RATs (e.g., inter-RAT handover, multi-connectivity) become simpler due to the unified framework for decision making.

- **Reduced risk of handover failures**: Although it has not been illustrated in the call flow in Figure 4.4, a dual connection from the source and target dBSs can be maintained towards the UE during handover. This prevents "ping-pong" handovers as well as reduces the risk of handover failures while ensuring session continuity.

- **Energy efficiency and power control**: Unlike in present-day multi-RAT networks, the SDN controller can regulate power levels for the entire system, thus reducing the overall interference in the RAN. This unified interference management may result in better system throughput. Some dBSs can even be turned off during periods of low traffic by re-distributing the load to the active base stations for increased energy saving.

- **Content caching and delivery**: By inspecting packets at the controller, data requests for popular content can be retrieved from locations near the dBSs instead of the external network through the GW. This results in reduced content retrieval time as well as efficient backhaul usage. Additionally, the source dBS may itself act as an anchor point and continue to serve the UE even after its handover to another dBS.

4.2 Performance Evaluation of the Proposed Architecture

We evaluate the performance offered by the proposed architecture by comparing it with that of the existing network. The simulation setup and the tools used for evaluation are described in this section. To evaluate the proposed multi-RAT architecture, we develop an ns-3 based evaluation platform in accordance with the proposal.

4.2.1 SDN based multi-RAT Evaluation Platform

The multi-RAT evaluation platform[2] has been developed on top of stock ns-3. It uses LTE and WLAN RATs as reference RATs. However, the results for 5G NR RAT are expected to be similar (but scaled) in comparison to LTE due to the similarity in throughput behavior.

The multi-RAT controller in the simulation platform is a logically centralized node comprising LTE controller and WLAN controller connected over IP. In accordance with the proposed architecture, an LTE dBS is created by transposing the control plane of the LTE eNB (i.e., RRC and RRM) into the LTE Controller. Communication between the dBS and LTE controller takes place using User Datagram Protocol (UDP) through an 'agent' (application) at the eNB and a control application at the controller. The agent forwards signaling messages such as RRC messages to the controller. This is done by intercepting control packets at the PDCP layer and forwarding them to the LTE controller. The control application at the LTE controller is coupled with the MME. Similar to the mechanism in stock ns-3, data traffic is routed from the dBS to the remote host through the combined GW.

The WLAN dBS is developed by moving the control functionality from the WLAN AP, such as admission control decisions, to the WLAN controller. The WLAN AP forwards the control messages, e.g., 'association request' to the controller over Transmission Control Protocol (TCP). The data plane packets are routed as per the routes configured by the controller. The platform also enables the creation of network slices, which are

[2]The ns-3 based simulation platform has been developed by Ashish Sharma, Rohan Kharade, and Abhishek Dandekar. The source code for the evaluation platform is available at `https://infonetsdn@bitbucket.org/infonetsdn/multirat_sdn.git`

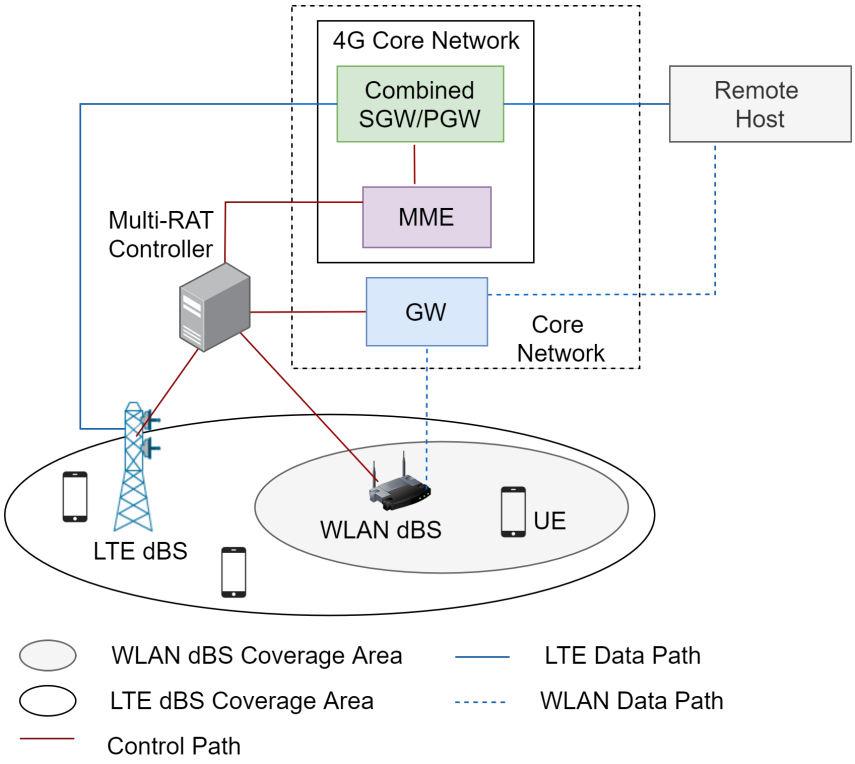


Figure 4.7: SDN based multi-RAT controller simulation setup

isolated from each other, as required by 3GPP specifications.

4.2.2 Simulation Setup

We evaluate the performance improvements provided by the proposed network in comparison with that of the present-day network for different traffic types. For this purpose, we create two network topologies, one for SDN based multi-RAT network and the other for the present-day multi-RAT. For simulating the present-day multi-RAT network, we use the stock ns-3 code to create a RAN consisting of an LTE eNB and WLAN AP with overlapping coverage areas. UEs in this network exchange data traffic with a remote host through GWs in the core network. The simulation topology for the proposed architecture is created using the evaluation platform and is illustrated in Figure 4.7. In this case, the multi-RAT controller is present as an additional node between the core network and RAN. Additionally, we create a network slice in both the networks. The slice for SDN topology consists of a multi-RAT controller that manages an LTE dBS, a WLAN dBS inside the coverage area of the LTE dBS, LTE core network, and a GW.

To simulate a slice within present-day multi-RAT, we construct a similar topology with LTE eNBs, WLAN APs, LTE core network, and GWs. We also assume that the GW has enough capacity to support all the best-effort users and hence, does not create a bottleneck. We believe that the users can be associated with either the LTE dBS/eNB or the WLAN dBS/AP. However, we associate the users present outside the coverage of WLAN dBS/AP with the LTE dBS/eNB. We also assume that the slice consists of a certain fraction of the total capacity of the LTE dBS/eNB as specified by the network orchestration policy and a WLAN dBS/AP. Let C_L be the maximum capacity (in Mbps) provided on the LTE dBS/eNB for the best-effort slice. The resources in LTE are equally divided among the data users. However, the data rate of individual users in LTE is also limited by the policy in the network (D_A, say). This assumption is in line with the restrictions posed by UE Aggregate Maximum Bit Rate (AMBR) [21]. If there are j data users in LTE, let the total throughput obtained in LTE be denoted by $D_L(j)$. Therefore,

$$D_L(j) = \min\{jD_A, C_L\}. \tag{4.1}$$

The throughput obtained in LTE increases linearly with the number of users. However, it is limited by C_L. In WLAN, we consider the saturation throughput model [74] for data users. This model accurately characterizes the maximum throughput of the system. It is calculated based on packet length, channel idle time, and the contention amongst users. If there are k users in WLAN, let the per-user throughput (in Mbps) be denoted by $D_W(k)$. The simulation parameters for WLAN and LTE RATs are described in Table 4.1 and have been obtained from [75]. The data user arrivals are assumed to follow Poisson processes, and service times are exponentially distributed. We consider $D_A = 5\,\text{Mbps}$ and $C_L = 50\,\text{Mbps}$.

We validate the performance of the proposed architecture with the help of a simple association algorithm for a slice supporting best-effort data users. The optimal solution for user-association, which maximizes the total throughput for the best-effort slice, can be obtained using the well-known value iteration algorithm [76] under suitable assumptions on the arrival process and service time of data users. However, it is known to have an exponential worst-case computational complexity. This motivates us to propose a simple

and greedy solution having a polynomial computational complexity. The algorithm is kept simple as it is used only as an illustrative tool to evaluate the capabilities of the architecture, which provides a network-wide view of the resources. Other sophisticated approaches for user-association can also be devised in the future. Also, the proposed algorithm does not require the knowledge of the statistics of the distribution of user arrivals and can hence, be implemented in real-time.

4.2.3 Algorithm for User Association

Algorithm 1 User association algorithm for best-effort slice in the proposed multi-RAT architecture

```
1: Initialize j ← 0, k ← 0 and B ← 0 .
2: procedure USER–ASSOCIATION
3:     for each arrival of data users do
4:         Evaluate B using Equation (4.2).
5:         if B == 1 then
6:             Associate user with LTE dBS.
7:             k ← k + 1.
8:         else
9:             Associate user with WLAN dBS.
10:            j ← j + 1.
11:        end if
12:    end for
13: end procedure
```

The steps for the proposed algorithm are described in Algorithm 1. As described, whenever a user sends an association request to the multi-RAT controller via the LTE dBS/WLAN dBS, the multi-RAT controller views the number of active best-effort users in LTE and WLAN, viz., j and k. To evaluate the preferred RAT for the association, we evaluate the following boolean variable B.

$$B = I_{\{D_L(j+1)+kD_W(k)>D_L(j)+(k+1)D_W(k+1)\}}, \tag{4.2}$$

Table 4.1: Simulation parameters for the multi-RAT network

Parameter	**Value**
Mean service time for user	60s
Path loss	$128.1 + 37.6\log(R)$, R in kms
WLAN channel bit rate	54 Mbps
Tx power for LTE dBS/eNB	46 dBm
Bandwidth of the LTE dBS/eNB	10 MHz
Tx power for WLAN dBS/AP	23 dBm
Tx power for UE	23 dBm
Antenna Type (LTE and WLAN)	Isotropic Antenna

where $I_{\{.\}}$ denotes the indicator variable. After evaluating B using Equation (4.2), if we observe that $B = 1$, then the user is associated with LTE, else WLAN is selected. The association which provides a better throughput is chosen based on the current load of both LTE and WLAN dBSs. For example, consider that $D_A = 5$ Mbps, $C_L = 50$ Mbps, and we consider 802.11g [77] WLAN dBS. Calculation [74] reveals that $D_W(1) > D_A$. Therefore, when the system is empty, it is better to associate an incoming user with a WLAN dBS. However, the greedy scheme dictates that- if a new user arrives prior to the departure of an already associated WLAN user, it is better to associate the new user with LTE since $2D_W(2) - D_W(1) < D_A$ and so on.

Note that the algorithm assumes the availability of timely and accurate load information at the controller for making decisions regarding handover and association. If the available information is stale or incorrect, the handover/association decisions made by the controller would be inaccurate and may result in degrading the system throughput. However, in a practical 5G multi-RAT system, timely and accurate multi-RAT load information is available to the central controlling entity as part of performance measurements over the management interface. These measurements are available periodically with periodicity ranging from an integer value of a few seconds (streaming based reporting) to 5 minutes (file based reporting), based on the mode of performance reporting [78].

We demonstrate the performance improvement provided by our architecture vis-a-vis existing network architectures for different traffic types in the succeeding section.

4.2.4 Simulation Results

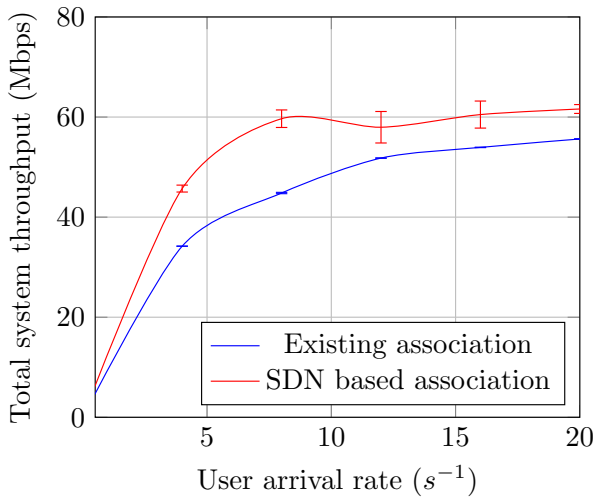
We measure the slice throughput and end-to-end data transfer latency per byte for different types of data traffic in both the networks. In existing multi-RAT networks consisting of overlapping LTE and WLAN coverage areas, an incoming user is always associated with WLAN until the WLAN AP denies association. All incoming users are then associated with the LTE eNB. We implement Algorithm 1 for RAT selection in the proposed network architecture. Note that this algorithm can be easily implemented in the proposed architecture as a global view of network resources is available at the controller. However, these types of algorithms cannot be easily implemented in present-day networks as they possess a fragmented view of the resources at every RAT.

4.2.4.1 Network Slice Supporting Best-effort Traffic

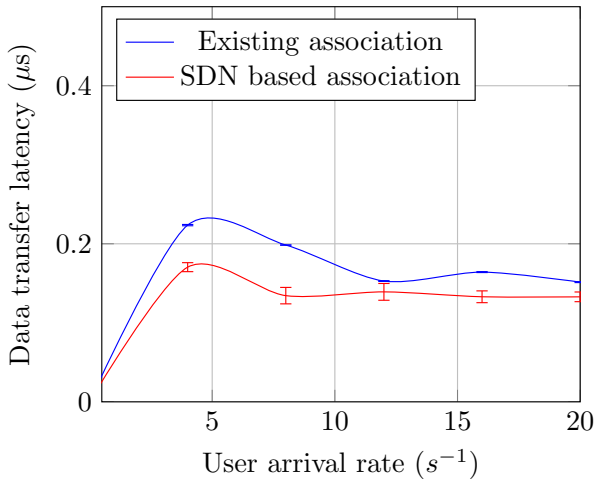
As illustrated in Figure 4.8(a), the slice throughput for the SDN based multi-RAT network is consistently better than that of existing networks for best-effort data traffic. This is due to the fact that in the existing multi-RAT networks, a given RAT may not possess the load information of other RATs. In the proposed multi-RAT architecture, the presence of load information of all the constituent RATs at the controller improves user association decisions, leading to an improvement in the total slice throughput. We also observe that the improvement obtained for end-to-end latency per byte of data (illustrated in Figure 4.8(b)) is higher with increased arrival rates in the SDN framework in comparison to the existing network. In existing multi-RAT networks, load information of the LTE system is not considered. Users are associated with WLAN until its capacity is reached. This results in increased packet transfer delays and reduced slice throughput as the contention in WLAN increases with increased user arrival rates.

4.2.4.2 Network slice Supporting Constant Bit Rate (CBR) Traffic

In this scenario, we evaluate the slice throughput and the data transfer latency for a network slice that supports traffic requiring CBR for each user, such as video, Voice over Internet Protocol (VoIP). In existing networks, users are blocked if the RAT (that the UE is connected to) is unable to provide QoS guarantees for traffic requests due to lack of capacity. However, for the SDN based multi-RAT network, we assume a load threshold



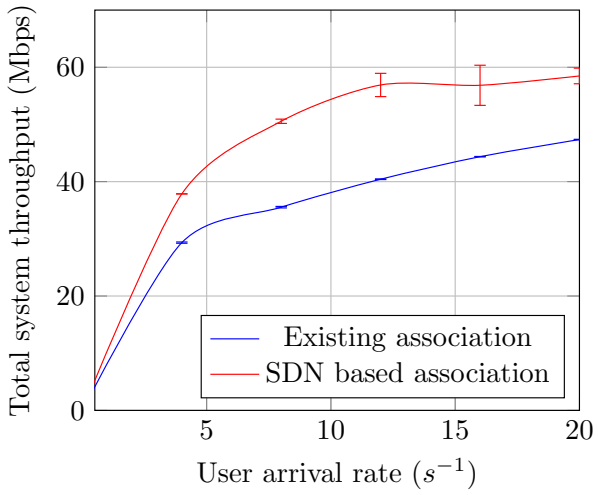
(a) Slice throughput v/s best-effort user arrival rate



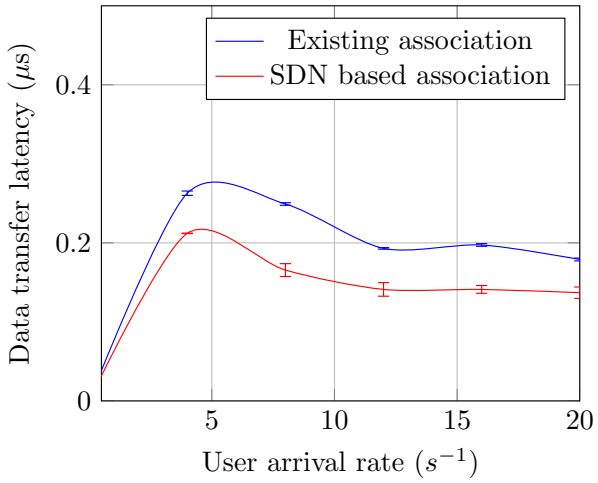
(b) Data transfer latency v/s best-effort user arrival rate

Figure 4.8: Performance of best-effort traffic slice

configured as per the network orchestration policy for LTE and block all users after the threshold. We also consider a threshold on the number of WLAN data users to guarantee that the per-user throughput is always above a certain data rate. The load threshold for WLAN is calculated based on [74]. As a result, we now block users in WLAN and LTE whenever their respective load thresholds are reached. Below the load thresholds, user associations are handled as specified in Algorithm 1. For this scenario, we assume that the constant bit rate for a user in LTE is 3 Mbps. The throughput for the proposed network architecture, as illustrated in Figure 4.9(a) is better for CBR traffic. Similarly, the latency for data transfer, as shown in Figure 4.9(b) is also lower in comparison to



(a) Slice throughput v/s CBR user arrival rate



(b) Data transfer latency v/s CBR user arrival rate

Figure 4.9: Performance of CBR traffic slice

that of the existing network. This is because the proposed algorithm always associates the user with the best RAT, given the current load conditions of the system.

4.3 Conclusion

In this chapter, we have proposed an SDN based network architecture for unified control and management of multi-RAT networks. With the help of this architecture, we are able to create multiple network slices over a shared physical infrastructure. End-to-end network control is provided through the use of slice-specific controllers. This not only ensures scalability but is also able to support diverse services, as each slice can be con-

stituted by slice-specific controllers and slice-specific data plane functions. The proposed multi-RAT controller architecture provides RAT-agnostic interfaces to applications and a virtualized view of the network resources. This virtualized view enables simplified control and management, especially for features that requiring interaction between RATs and also within multiple nodes of the same RAT. The performance improvements provided by this architecture have been demonstrated through call flows and simulation results.

In conclusion, we have elucidated scalable mechanisms for enabling multi-RAT control within sliced networks. However, this work does not provide an evaluation of the granularity at which network resources are to be virtualized for optimal network performance. In the forthcoming chapter, we investigate this problem and also develop a generalized framework for slicing the multi-RAT RAN.

Chapter 5

SDN/NFV based Framework for 5G RAN Slicing

5G networks are expected to support different verticals and services having different latency and throughput requirements over a shared infrastructure. As a result, network slicing, which allows for the creation of end-to-end logical networks has been put forth as an important 5G requirement by 3GPP. Network slices are to be deployed in a manner that allows operators to manage and orchestrate each slice independently, without affecting the operations of other slices [63]. This allows service providers to deploy newer services over a common network infrastructure without affecting existing services. Moreover, multiple instances of a slice type (isolated from each other) can also be deployed within a network. Although the signaling exchange for slice creation has been specified by 3GPP, the choice of network functions for a specific slice within the RAN is at present deemed to be implementation dependent [53]. Also, many of the existing solutions for network virtualization and slicing are tailored for a specific network architecture and do not provide insights into important factors for slicing other architectures [31].

In this chapter, we endeavor to bridge this gap by proposing 'Virtualized RAN (VirtRAN)', which is a framework for slicing the 5G multi-RAT RAN[1]. Prior to introduc-

[1]This work has been published in [79] and also been submitted as a standards contribution to IEEE [27]. This work was supported by Department of Telecommunications, Ministry of Communications, Government of India as part of the Indigenous 5G Test Bed project.

ing the proposal, we analyze the available mechanisms for RAN slicing and then proceed to outline the characteristics that are essential for a good slicing solution.

5.1 Observations on RAN Slicing

As indicated in Chapter 2, OpenFlow, in its present form, is not suitable for usage in software defined wireless network environment. This is due to the fact that the concept of flows may not be able to capture some of the distinctive characteristics of wireless networks. With a flow level interface, the allocation of the underlying resources, e.g., bandwidth, compute, or storage to each of the network slices (individual flow spaces) is not visible to the SDN controller and the controller is unable to allocate these resources directly to the network slices. Even though the controller configures flows on the data plane nodes, the responsibility of actual resource allocation to individual flows is with the data plane. While this issue does not occur in wired networks as there is minimal variation in the link quality across time and users, this may not be the case in wireless networks. Due to time and user-specific variation in radio channels, the number of underlying resources allocated to individual flows and therefore, to the network slices (flow ensemble) varies over time, which may defeat the concept of slice separation.

To illustrate this problem, we consider an SDN based LTE network where two users are accessing a Youtube video over a smartphone using LTE. In this network, one of the users (user 1) is present in the center of an LTE cell near the base station while the other user (user 2) is at the cell edge, far from the base station. Let us assume that the SDN controller configures the network into two slices and configures a policy for providing equal resources to them. Also, we assume that user 1 is accessing slice 1 and user 2 accesses slice 2. However, the user at the cell edge (user 2) is likely to experience a poor radio link as compared to the user at the cell center (user 1). Even though the policy was to provide equal resources to the flows, due to adverse channel conditions of user 2, the base station would allocate a larger amount of resources to user 2 for compensating for the channel conditions. Further, resource allocation may vary over time, especially if users are mobile. If the resources present in slice 2 are inadequate for compensating for the channel conditions, some of the resources may be drawn from slice 1, thus affecting slice isolation. As described above, the network policy proves ineffective. It also demonstrates

that granular control may not be achieved only through flow-configuration in Wireless networks.

Another important factor for ensuring optimal system performance in wireless networks is interference management. In order to achieve this, operators not only have to optimize throughput by configuring flow paths but also manage radio parameters such as transmit power, etc. OpenFlow does not have provisions for configuring radio-related parameters and may need to be extended or supplemented by another protocol such as Network Configuration Protocol (NETCONF), SNMP, etc. Although some work has been done to address this issue for WLANs [80], it still remains a challenge in cellular networks such as LTE and 5G-NR.

One of the recommendations that is available within the specifications is the use of SDN based network architectures for slicing networks [20]. SDN architectures can provide features such as virtualization and recursion etc., making them ideal for the implementation of network slicing vis-a-vis traditional networks. An example recursive framework for SDN based networks has been defined by Open Networking Foundation (ONF) [81]. This framework allows SDN controllers to be placed in a recursive or hierarchical fashion for better scalability. Within a recursive framework, a higher-level controller, say at level $n + 1$, appears to the lower level controller n as an application. Similarly, the controller at level $n - 1$ appears as data plane to the controller at level n. Recursion allows for the creation of applications that provide finer-grained services by combining multiple applications. Recursive placement may also be used to provide different security levels within the network.

The division of the flow-space into smaller sub-spaces using OpenFlow can be utilized in recursive network architecture. In a recursive architecture, SDN controllers are organized into a recursive hierarchy. A lower-level controller (the controller closer to the data plane) is responsible for subdividing the flow-space into sub-spaces and mapping these individual sub-spaces to independent virtual networks. Each one of these virtual networks (sub-spaces) may further be exposed to separate higher-level controllers for management and control purposes. The virtual network controllers can manipulate the corresponding virtual networks through OpenFlow protocol. However, with a recursive architecture as

proposed in ONF [81] and OpenRoads [23], the data plane may not have an awareness of network slices, as the division of flow space into different network slices is visible only to the controller. Therefore it may be challenging to adhere to any resource separation at the slice level in such a scenario.

A widely used element to aid network slicing is the hypervisor. A hypervisor is used for abstracting the physical network to create one or more isolated virtual networks that can be controlled individually. The creation of multiple networks can be achieved by placing the hypervisor between the data plane and the controller plane. Hypervisors are used for reducing the complexity of network control by providing a simpler view of the network in terms of topology, network resources, etc. A detailed analysis of the various types of hypervisor for SDN networks is available in [82].

5.2 Desirable Characteristics of Slicing Frameworks

Motivated by the above discussion, we identify the characteristics for designing better frameworks for slicing. We recommend that a slicing framework should possess the following characteristics-

- The architecture should support the concept of abstract network resources, which can be used for virtualization and network slicing.
- The abstract network resources should not be wholly unconnected from the underlying resources that are being represented and allocated. For example, the concepts of traffic flow, as defined in OpenFlow and used in OpenRoads [23], is a simple and abstract resource, but it may not be capable of accurately representing the radio resources in wireless access networks.
- The architecture should provide a flexible virtualization scheme so that different mechanisms (simpler to more complex) can be used for virtualization depending on the use case/requirements to be supported.
- The architecture should support the virtualization of resources at multiple levels as resources can be defined and grouped at different levels. For example, the LTE physical layer takes a frequency band and represents it as PRBs to higher layers.

Similarly, the MAC layer may take the PRBs and represent them as Virtual Resource Blocks (VRBs). The architecture should have the flexibility to work with such different types of abstractions.

- It should support an SDN based architecture with separate control and data plane functions with an open interface between the two.

- The architecture should support a clean separation between the control plane and the data plane and allow for data plane virtualization at multiple levels without the presence of the control plane functions in the data path. This is different from the SDN based architecture as proposed by ONF.

We use the characteristics listed here as a guideline to design 'VirtRAN', a recursive framework for multi-RAT RAN slicing. The details of the proposed framework are described in the next section.

5.3 VirtRAN - Proposed Framework for SDN based Multi-RAT Slicing

VirtRAN enables us to virtualize the network at multiple layers in order for better slicing the network. The proposed framework is illustrated in Figure 5.1. As illustrated, VirtRAN has a well-defined separation of control and data planes. The data plane is further subdivided into multiple sub-planes, wherein each sub-plane performs a part of the data plane functionality and utilizes a set of resources. There may be an optional 'Virtualization Layer (VL)' over each one of these sub-planes. The VL at a given level creates an abstract view of the underlying resources and provides these resources/groups of resources to the layer above for use. It is also responsible for ensuring isolation across resource groups. It not only virtualizes radio resources but may also virtualize compute and storage resources for higher sub-planes. As illustrated in Figure 5.1, each network slice (resource group) at a sub-plane level may be controlled/managed by a separate slice-specific SDN Controller. VL acts under the control of an orchestration entity known as the orchestrator. The role of the orchestration entity may also be played by the SDN controller responsible for the control of the underlying sub-plane (below the VL).

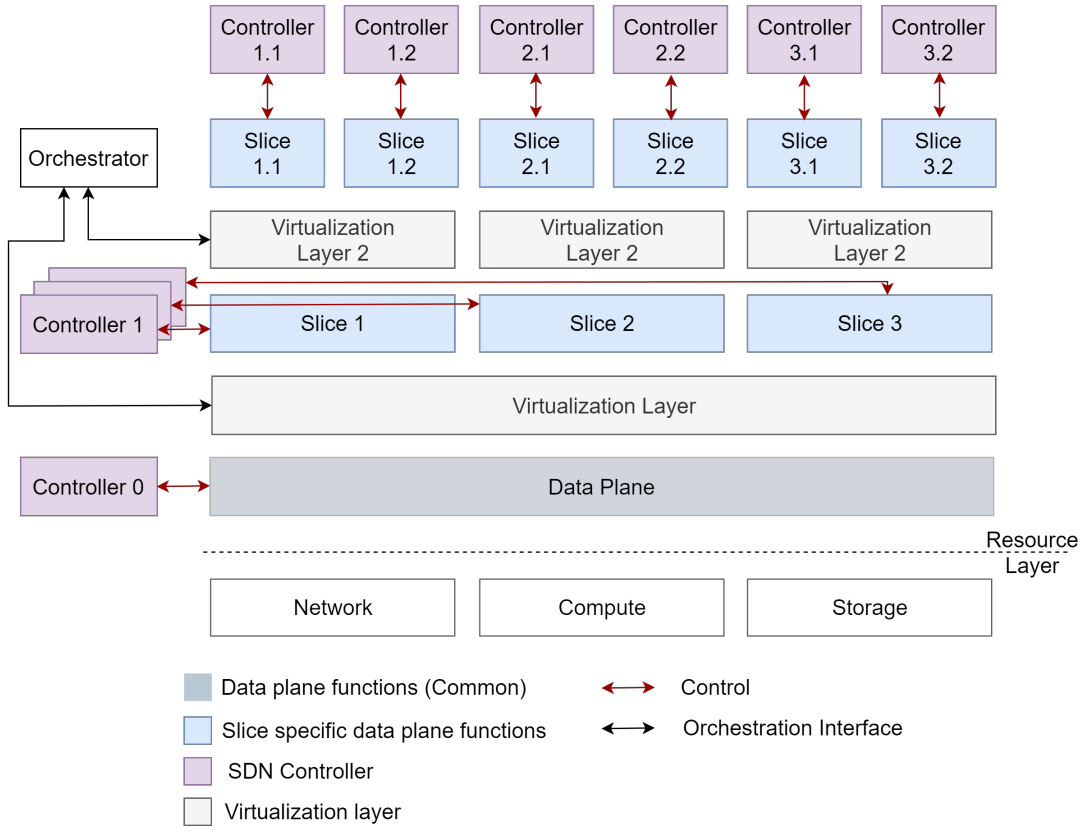


Figure 5.1: Proposed recursive architecture for slicing

In this case, the SDN controller manages/controls both the virtualization layer and the underlying sub-plane.

We discuss further details of the proposed architecture with the help of an example of a network consisting of LTE/5G NR technology, illustrated in Figure 5.2 considering the eMBB use case. In this case, a possible division of the radio interface in sub-planes is to group the radio protocol layers and put them into separate sub-planes, e.g., PHY and MAC Layer are part of a single sub-plane whereas RLC, PDCP (and SDAP in case of 5G) layers form a part of another sub-plane. The sub-plane containing the RLC/PDCP layers also contain other layers such as GPRS Tunneling Protocol - User Plane (GTP-U) etc., which are part of the core network interface. In this case, a VL can be placed over the PHY and MAC layers and would be responsible for virtualizing the underlying physical radio resources into subsets of (virtual) radio resources. The VL is also used to divide the virtual resources into multiple subgroups and allocate each of the resource subgroups to a virtual network or network slice.

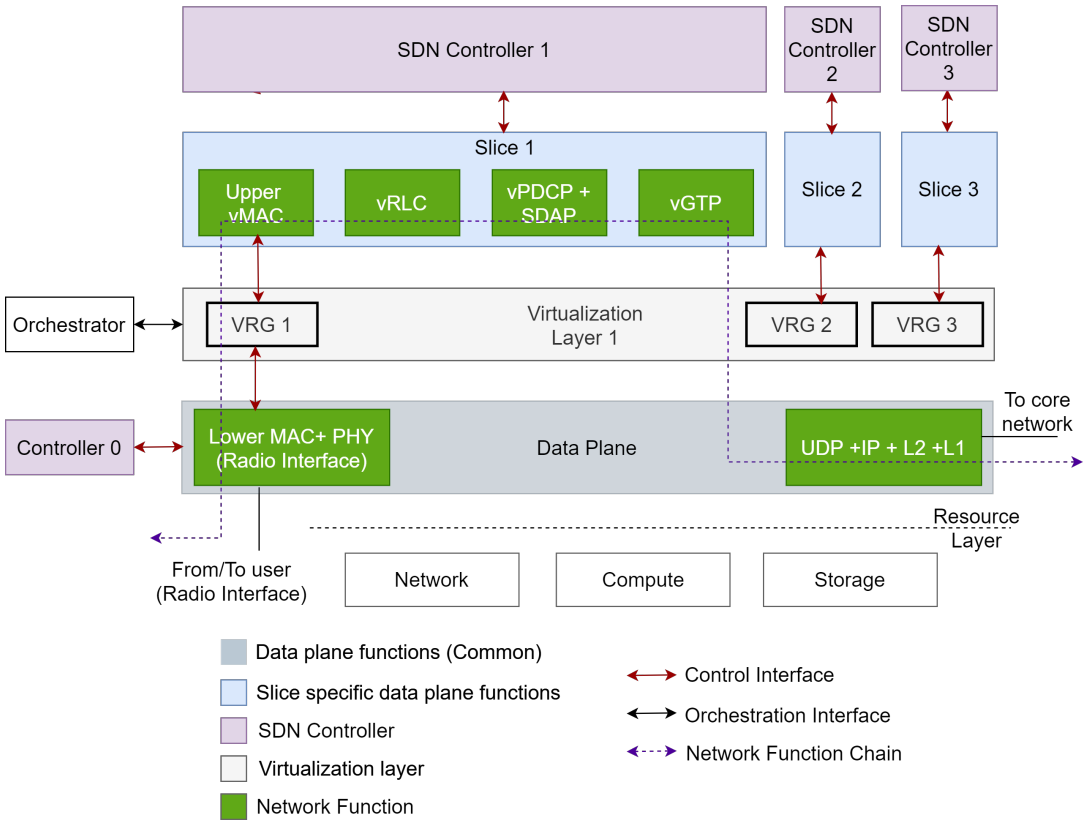


Figure 5.2: Data Path illustration for 5G NR in the proposed recursive architecture

It would be possible to have multiple network slices comprising individual RLC and PDCP layers. Here a slice-specific MAC layer functionality would also be required in the higher sub-plane to allocate virtual radio resources to slice-specific users. To implement MAC layer slicing in a practical 5G system, the resources available from layer below it (i.e., the PRBs present in the physical layer) would be grouped at periodic intervals into one or more VRBs. The number of vRB groups within a period are equal to the number of slices required to be supported and resources within an individual VRB are allocated to a specific slice. The grouping of global radio resources into VRBs is performed by a virtualization layer (placed over the PHY layer) at a granularity ranging from the duration of a Transmission Time Interval (TTI) i.e., 1 ms to a larger duration, based on the implementation. The number of PRBs present in a VRB is decided based on a slicing algorithm, which is implementation dependent.

The resources within a given VRB are then scheduled using a slice specific scheduler. The scheduling information along with other information that common across slices (i.e.,

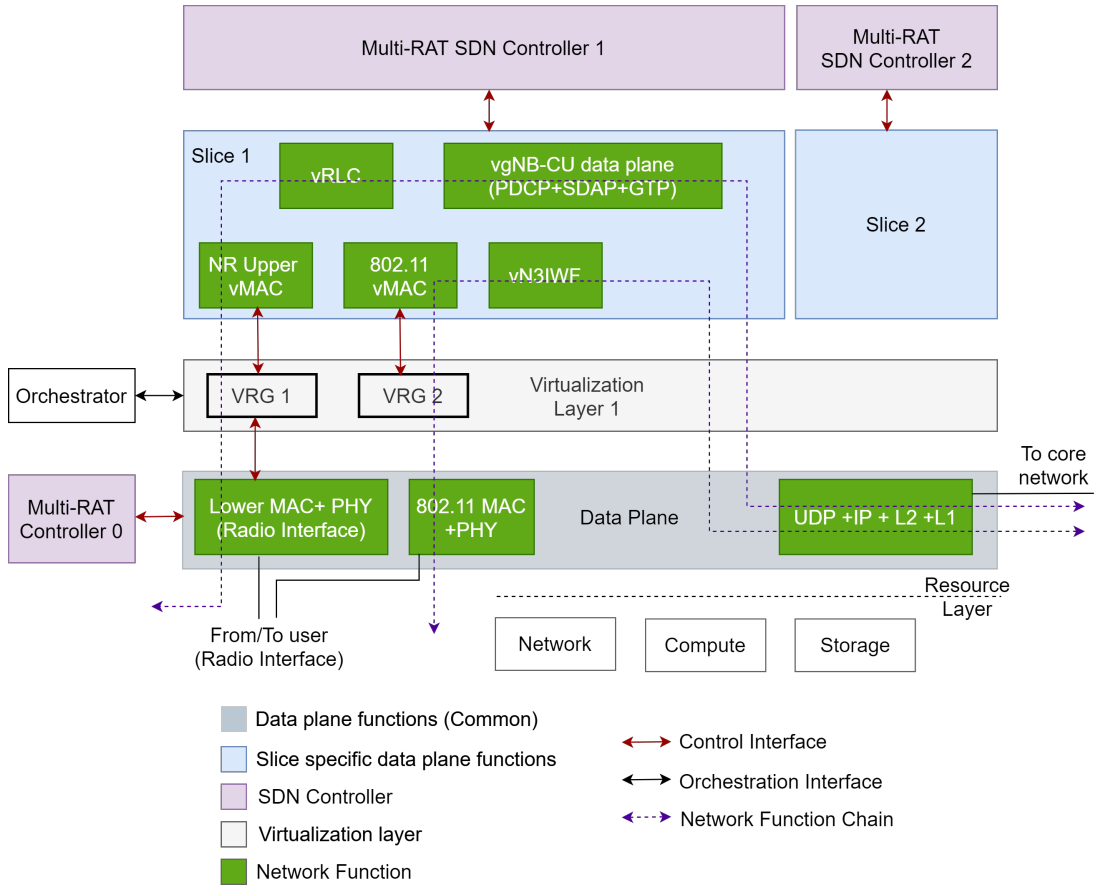


Figure 5.3: Example of virtualization of the proposed architecture for 5G NR and WLAN

reference signals, control channel information etc.) is combined into a global map and then transmitted on the downlink. On the uplink, the virtualization layer would be responsible for mapping UE requests to the allocated slice specific resources.

A slice may also contain other layers, such as SDAP and GTP-U. Each of the slices utilizes one of the resource subgroups. When this concept is extended to the multi-RAT RAN as in illustrated Figure 5.3, we observe that there are slice-specific virtual MAC, RLC, PDCP, SDAP, and GTP functions.

The hypervisor over the MAC layer divides the underlying radio resources into three different groups. Each group is mapped to a corresponding virtual resource group and presented as a slice. The slice-specific virtual MAC function is mapped to one of these resource groups by the virtualization layer. However, there is no limitation on the number of slices and virtual resource groups, and the VL may be configured to support any number of slices/resource groups.

It is possible for the proposed architecture to support a unified Multi-RAT RAN. The VL situated over LTE MAC and the IEEE 802.11 MAC layer unify the underlying RAT-specific physical radio resources through an abstract/virtual resource view created for the higher sub-planes. This would enable the higher sub-plane data functions, e.g., RLC, PDCP layers, to utilize either or both of the underlying LTE or 802.11 MAC layers seamlessly. The division of data plane into multiple sub-planes, virtualizing the resources at multiple levels (at each sub-plane level) and putting each sub-plane under sub-plane specific SDN controller enables granular control over the resources being used in the network as opposed to the architectures such as the one proposed by ONF, wherein there may be a single level of abstraction used over the data plane resources.

As shown in Figure 5.1, it is a recursive architecture wherein there may be a VL over each of the slice-specific sub-planes. The VL over each slice would further divide the underlying slice-specific resources into smaller resource groups and export them for manipulation to another higher level slice-specific function. Although most of the discussions here have been in the context of RAN, it is evident that the concepts explored here are generic and can be applied to the core network as well. The performance improvement provided by this proposal is evaluated by implementing a rate-limiting mechanism for VirtRAN using ns-3. This evaluation and the mechanism are detailed in the succeeding section.

5.4 Simulation of VirtRAN

In this section, we provide simulation results for VirtRAN compatible LTE RAN. The simulation is carried out for LTE-RAN using ns-3 [28] LENA module. We use LTE for simulation as a full stack 5G NR simulator is unavailable, to the best of our knowledge. The VirtRAN concepts are illustrated through the creation of network slices in LTE RAN. In order to perform the simulations, the following additions to the ns-3 LENA simulator were made. This additions are described in Section 5.4.1.

5.4.1 Additions to ns-3 LENA framework

For realizing recursive slicing with ns3-LENA, we adopted the approach of slicing the PDCP and MAC layers. The following changes were made to achieve recursive slicing.

- For virtualizing the PDCP layer, we have defined a PDCP resource in terms of the achievable downlink data rate. Therefore, a slice on the PDCP layer has been rate-limited to provide a maximum data rate of say 'n' Mbps. The downlink rate limit has been enforced by first estimating the number of PDCP PDUs required to provide the peak-throughput of 'n' Mbps and then enforcing that the number of actual transmitted PDUs are within the estimated range with the help of a counter. The counter has been configured to drop all packets destined for a given slice, when the number of transmitted PDUs exceeded the allowed limits for the slice.

- At the MAC layer, slicing has been carried out over a time window of duration 10 ms (i.e., MAC frame duration). Each frame comprises 10 subframes of duration 1 ms. For slicing MAC resources (over the 10 ms window) in a given ratio say $n : 10-n$, we have statically allocated the resources in the first n subframes to slice 1 and the resources in the next $10 - n$ subframes to slice 2. We have used a slice-specific scheduler to schedule the UEs within a slice. In our simulations, we have demonstrated MAC layer virtualization on top of the Proportional Fair scheduler.

5.4.2 Simulation Setup

The simulation setup consists of a macro eNB (Release 8) with a single transmit and receive antenna. Hence, the eNB has a maximum capacity of 75 Mbps in each direction. Users are uniformly distributed within a cell radius of 100 m and access services over slice 1 or slice 2. Each user accesses an application at a data rate of 10 Mbps. We also assume that the number of users for each slice is different and that every user has full buffer traffic. The rest of the parameters used in the simulations are detailed in Table 5.1. We perform rate-limiting at PDCP to achieve a distribution of resources in the ratio 3 : 2 across the slices. To achieve this, we perform the rate-limiting using two different mechanisms. For both the mechanisms, we classify downlink traffic at the PDCP layer as belonging to either slice 1 or slice 2. Traffic flow statistics are monitored for every slice, and once the threshold for a slice is attained, the remaining slice users are not serviced. As a result, the maximum rate for each slice is limited.

As our proposal aims to restrict the maximum data rate for the slice only on the downlink, uplink PDCP traffic should remain unaltered and is therefore isolated.

Table 5.1: Simulation parameters

Parameter	**Value**
Path loss	$128.1 + 37.6 \log(R)$, R in kms
Tx power for LTE dBS	46 dBm
Bandwidth of the LTE dBS	10 MHz
Tx power for UE	23 dBm
Antenna Type LTE	Isotropic Antenna

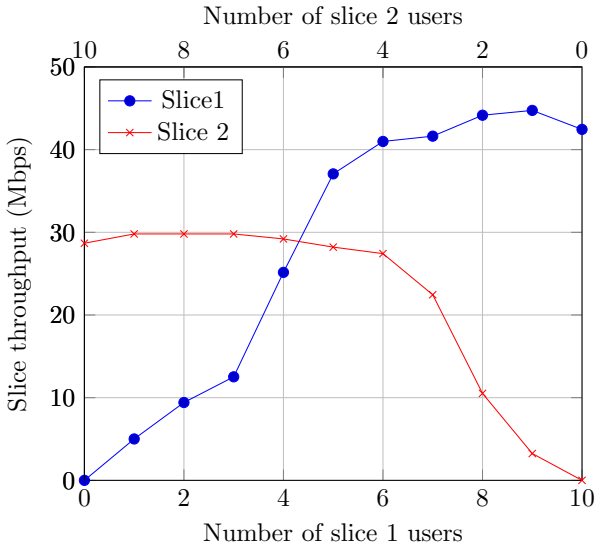


Figure 5.4: Virtualization based on system capacity

5.4.3 Mechanism 1: Resource Division based on System Capacity

In this mechanism, we enforce a hard limit on the resource division based on system capacity. The limiting throughput is set to 45 Mbps for slice 1 and 30 Mbps for slice 2. The resulting slice throughput for various user distributions is illustrated in Figure 5.4. As expected, we observe that the throughput for both slices does not exceed the limit, even when a particular slice is loaded.

5.4.4 Mechanism 2: Resource Division based on Offered Load

In the second mechanism, we divide resources in the prescribed ratio based on the offered load. If we consider that there are 10 users within the cell, the offered load is 100 Mbps. Therefore, we set a limiting throughput rate of 60 Mbps for slice 1 and 40 Mbps for slice 2. As illustrated in Figure 5.5, this type of allocation provides higher throughput for a slice,

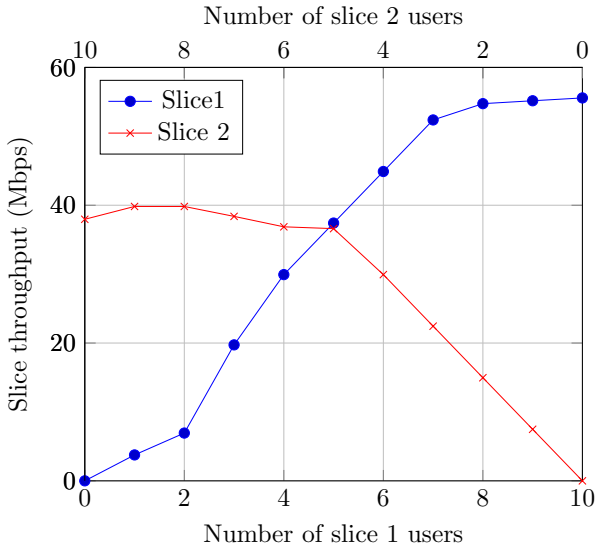


Figure 5.5: Virtualization based on offered load

especially when the system is not loaded to its capacity, while ensuring resource division at the PDCP level. This also proves that virtualization at higher layers of the protocol stack provides a simpler way to achieve network slicing at lower system loads and better channel conditions.

From Figures 5.4 and 5.5, we observe that throughput for both slice 1 and slice 2 is within the required range. Figure 5.6 also illustrates that load based resource allocation provides better system throughput in comparison to the system capacity based fixed resource allocation. These examples demonstrate that network virtualization and slicing implemented at the PDCP layer can achieve the desired goals. This is especially true when the users associated with each network slice are experiencing similar channel conditions and accessing similar types of services. However, there may be certain limitations to the implementation of network slicing at the PDCP layer.

This becomes apparent, especially when the network consists of slices providing services of different traffic priorities, and users associated with a slice providing high priority traffic are consistently experiencing poor radio channel conditions. In such scenarios, virtualization at a lower layer (sub-plane), e.g., at the MAC layer, would be necessary to improve resource sharing.

To illustrate this fact, we increase the number of users within the cell upto 50. We

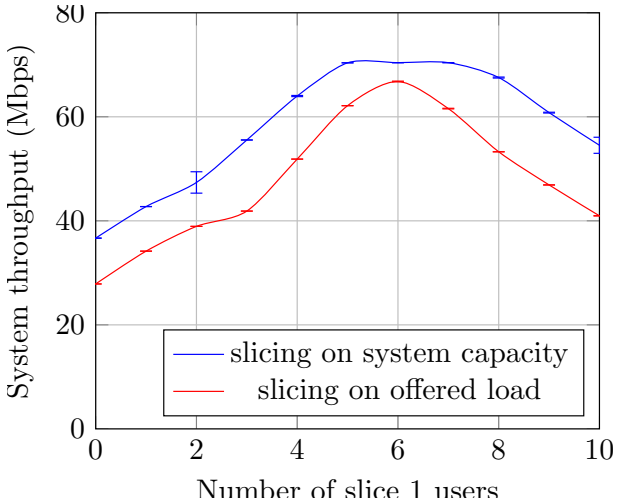
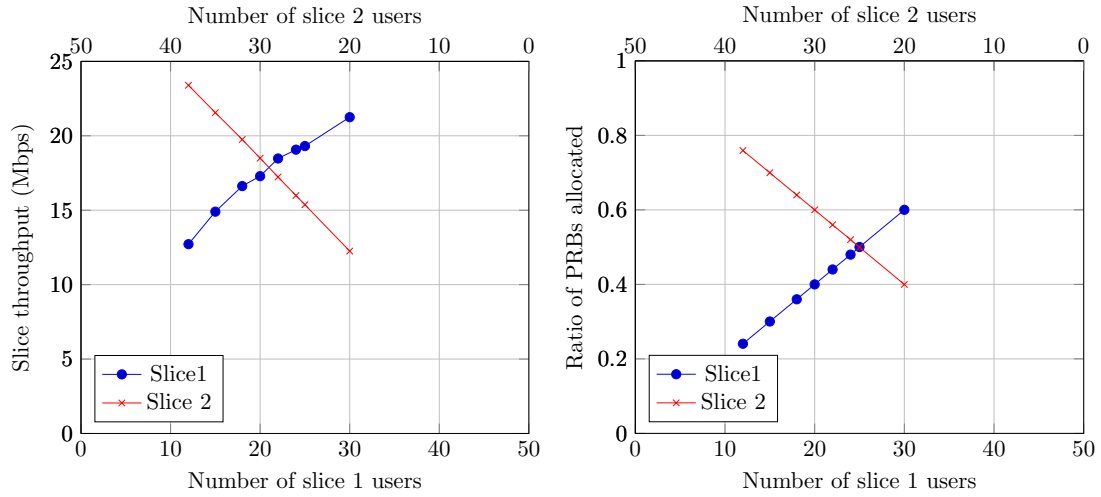


Figure 5.6: Comparison of system throughput for resouce division based on system capacity and offered load

statically allocate 30% of the resources at PDCP to slice 1 users and 70% of the resources to slice 2 users. Among them, slice 1 is providing Guaranteed Bit Rate (GBR) type traffic, and slice 2 is serving non-GBR users. It is assumed that slice 2 users are at the edge of the cell. We calculate the slice throughput and the ratio of the PRBs allocated to each cell for this scenario in the range of 12 to 40 users. The range is chosen such as to ensure that each slice has sufficient traffic and is saturated. We observe that, if slice 1 users are generally experiencing poor channel conditions, a larger number of radio resources are granted to them to maintain the QoS. This can be observed from Figures 5.7(b) and 5.7(a) where the ratio of PRBs granted to slice1 is consistently greater than the 30% allocated by the PDCP layer and there is a large variation in the throughput for each slice. In order to alleviate this, isolation and allocation of resources could also be performed at lower layers such as MAC, based on the network conditions.

Note that for this particular scenario, it would be possible to estimate the results without the help of the simulator. In the above scenario, we have added resource constraints only at the PDCP level ((30% for slice 1 and 70% for slice 2) and no rules have been enforced at the MAC level. The rule at the PDCP level only results in the limitation of the system throughput and does not allow for precise control of radio resources. Together with this, the use of proportional fair scheduler at the MAC layer attempts to provide a similar rate to all users, thus dividing the capacity among the users.



(a) Throughput for slices when slicing is performed at PDCP layer

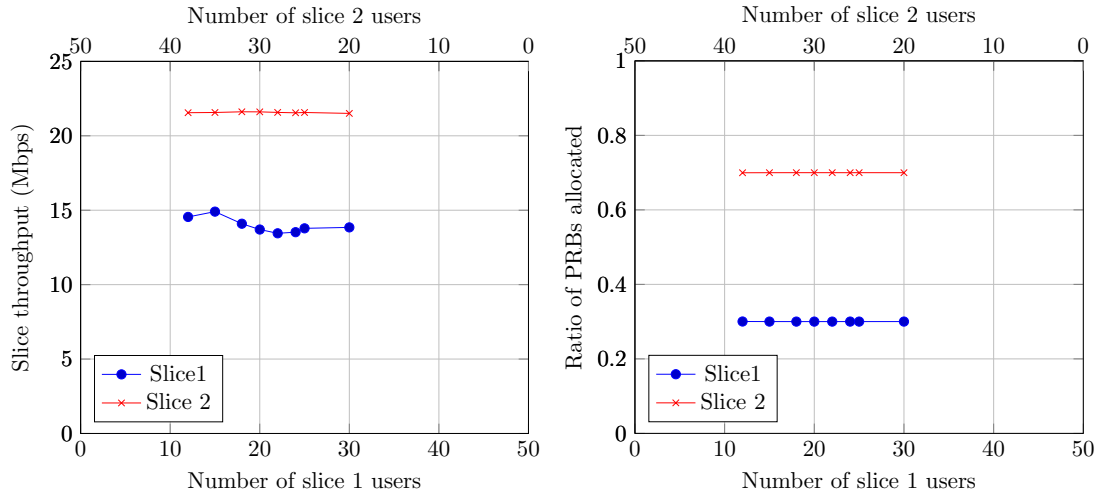
(b) Ratios of PRBs allocated when slicing is performed at PDCP layer

Figure 5.7: Throughput and PRB allocation ratio for slicing at PDCP layer

VirtRAN allows for recursive virtualization across layers for achieving the fine-grained RAN control as desired in this scenario. This kind of recursive virtualization can be performed in situations where slicing has already been performed at a higher layer, and the operator wishes to further enforce a finer-grained QoS at a later point in time. The results for this configuration are illustrated in Figures 5.8(b) and 5.8(a). Here, we can observe that the ratio of PRBs granted at the ratio is nearly equal to the pre-defined ratio of 30% to slice 1 and 70% to slice 2. As a result, we are able to ensure that the slice with poor radio conditions is not resource-starved. This is shown through a smaller variation in slice throughput in Figure 5.8(a). Note that the slice throughput and hence system throughput obtained through slicing at both MAC and PDCP layers is lesser than that obtained through MAC layer slicing. This is due to the fact that slicing allocations have just been created to divide the system resources and have not been devised to optimize system parameters such as throughput, delay, etc., and is not a result of slicing at a lower layer. Based on the discussion so far, we can observe that VirtRAN provides several advantages over existing proposals. They are summarized in Section 5.5.

5.5 Advantages of the Proposed Architecture

A summary of the advantages offered by VirtRAN are mentioned below-



(a) Throughput for slices when slicing is performed at PDCP and MAC layers

(b) Ratios of PRBs allocated when slicing is performed at PDCP layer

Figure 5.8: Throughput and PRB allocation ratio for slicing at PDCP and MAC layers

- The framework does not suffer from the flaws proposed by schemes like OpenRoads [23], which uses a single abstraction mechanism, i.e., flow space for the underlying data plane. The mechanism proposed here can be used to specify data plane abstractions at multiple levels enabling more granular control over resources in the network. The individual data sub-plane abstractions used by the controller to control/manage the individual sub-planes are dependent on the resources being managed by these sub-planes. For example, the sub-plane containing the GTP and PDCP function is controlled through the usage of tunnels/bearers/data rates, whereas the sub-plane containing the MAC may be controlled through manipulation of radio resources. Further, placing a VL over the LTE MAC layer and virtualizing the underlying radio resources and allocating them to different slices enables better slice level control over radio resources than an architecture like OpenRoads, which uses the abstraction of flow spaces to manage the radio resources also. Due to better granular control over resources at multiple levels and the flexibility of virtualization, the proposed architecture is better suited to mobile networks than other schemes, as discussed in the paper.

- By using VirtRAN, we can bring a Multi-RAT network comprising LTE, 5G NR and, IEEE 802.11 under a unified SDN based control and management framework.

We can also integrate IEEE 802.11 technologies with LTE/5G NR in a simpler manner. This can be achieved through the VL, which virtualizes the underlying 5G NR, LTE, and 802.11 PHY and MAC layer level resources in a unified manner for the higher level sub-plane functions like RLC and PDCP. This is not possible in the existing 3GPP LTE or 5G NR radio access network.

- The introduction of a VL between sub-planes (containing individual data plane layers) and associating the sub-planes through these virtualization layers does not require any deviation from the 3GPP specifications. The resource abstractions are built over these layers by the VL as an additional mechanism without impacting their functionality. This differs from some of the schemes like Radiovisor, which may require changes in the protocol stack. Moreover, the suggested changes do not impact UE protocol, and therefore, no changes are required in the UE to communicate with 5G/LTE/WLAN networks.

- VirtRAN offers a flexible mechanism for network virtualization and slicing, wherein virtualization (and consequently the slicing) can be performed at one or more different data plane layers (planes). Therefore, it is possible for a vendor/operator to choose the layers (planes) at which virtualization should be performed depending on the use cases that need to be supported. As shown through simulations, for the scenario depicted in Figure 5.7, virtualization and slicing can be done at the PDCP layer. An example of such a scenario is when an LTE network is to be used by two different operators to provide similar types of services. In such a scenario, virtualization can be done at the PDCP layer, and there may not be a need to virtualize it at a lower layer. Whereas for a different scenario, when the network slices are created to support different types of services, say uRLLC and eMBB (similar to the case in Figure 5.8, virtualization may be done at the MAC layer. By using VirtRAN, we can also virtualize the RAN node at multiple layers (planes) as shown in Figure 5.2.

5.6 Conclusion

In this chapter, we provide key insights into the requirements for slicing multi-RAN architectures. To address the prevalent gaps in slicing 5G multi-RAT RAN, we propose and evaluate VirtRAN, an SDN/NFV based architecture for slicing multi-RAT RANs. VirtRAN supports a unified framework for virtualization and slicing of multi-RAT wireless access networks. To enable an integrated virtualization framework for multi-RAT networks, we utilize the concept of abstract network resources. These abstract network resources can be defined uniformly for different RATs and enable their integration under a single framework. Network slicing under VirtRAN can be performed at one or more different data plane layers (planes), and depending on the requirement, one can choose the layers at which the network slicing is to be done. The architecture also provides a framework for recursive virtualization, which is essential for achieving policy-based resource allocation.

VirtRAN addresses an important problem by providing a generalized solution for slicing 5G multi-RAT RANs. However, with regards to practical slice deployments, we also need to be cognizant of the fact that the present 3GPP specifications are restrictive, requiring homogeneous slice deployments within a geographic area for ensuring slice mobility. Thus, to achieve complete flexibility in slicing the 5G multi-RAT RAN, frameworks like VirtRAN must be supplemented by additional protocols/mechanisms that enable the relaxation of slice deployment requirements while providing a similar level of performance. In the forthcoming chapter, we propose enhancements to an existing 3GPP protocol for bringing in more flexibility in slice deployment requirements, thus completing the solution.

Chapter 6

Protocols for Enabling Flexible Slice Deployments

As previously discussed, 3GPP 5G networks employ network slicing to implement multiple service types corresponding to diverse business needs. A UE can be connected simultaneously to multiple slices for accessing various services [21]. In this chapter, we identify the drawbacks of the requirements identified in 3GPP specifications with regards to slice deployment in multi-RAT environments [53] and propose mechanisms for enabling flexible slice deployments[1]. At the outset, we review the requirements mentioned in 3GPP and understand their implications.

6.1 Slice Deployment Specifications in 3GPP

At present, the 3GPP specifications allow UE to concurrently connect to 8 slices. Separate PDU sessions are created for the UE for each of the connected slices. Every slice can be configured to support a different service and is identified using S-NSSAI. S-NSSAI specifies the slice behavior in terms of services and features as well as comprises information that helps in differentiating multiple slices. For ensuring service availability across the network, network slices must be deployed on a sufficient number of network nodes.

[1]The work presented in this chapter has been published in [83, 84] and patented in [85]. This work was supported by the Department of Telecommunications, Ministry of Communications, Government of India as part of the indigenous 5G Test Bed project.

Deployment of a network slice requires two types of resources viz., link resources with constraints such as bandwidth, latency, packet loss, etc., and node resources such as compute and storage at each node [86]. 3GPP specifications state that slices must be deployed homogeneously over a Registration Area (RA) for providing adequate coverage [53]. As a result, all supported slices need to be deployed on each network node within the UE RA. This is a challenging proposition for twofold reasons. Firstly, RAN may comprise nodes with varying hardware/software capabilities, e.g., macro Base Stations (BSs) with a larger pool of resources and small cell BSs typically with limited resources. It may not be easy to support slices with diverse service requirements on every network node, especially on small cell BSs. Secondly, 3GPP 5G networks may support multiple RATs such as 5G NR, LTE, and WLANs connected to a common 5G core network. As the underlying characteristics of these RATs are different, the achievable latency and throughput values for each of these RATs differ from one another. Therefore, not all slices may be supported on a RAN node. At present, this is avoided by designating separate RAs for 3GPP and non-3GPP RATs, enabling the deployment of different sets of slices over them. However, this may lead to the unavailability of certain slices for a UE when it moves across RAs.

Also, studies show that network costs increase sharply with an increase in the number of slice instances at the nodes in the network [87]. Therefore, deploying slices over an entire RA would lead to an increase in the network CAPEX and Operating Expense (OPEX). Moreover, the 3GPP slice deployment specifications place a lot of additional demands on the capabilities of 5G base stations (gNBs) when an increased number of slices are to be deployed [87]. Due to all the above reasons, it is important to allow network slices to be deployed in a more flexible manner. For enabling flexible slice deployments, we suggest the use of 3GPP multi-connectivity [88]. We demonstrate that with a few enhancements, multi-connectivity can help in improving UE slice mobility and availability for UEs connected to multiple slices, thereby allowing flexible slice deployments.

Several studies that independently explore UE slice mobility management and multi-connectivity are available in the literature. A summary of the prior art is provided in Section 6.2.

6.2 Related Work

In some of the studies on multi-connectivity [89, 90], the authors focus on reducing handovers in Heterogeneous Network (HetNet) scenarios while improving mobility robustness by maintaining connections to multiple BSs. The authors in [89] propose the use of dual connectivity in 5G NR networks for improving handover performance by tracking UE channel quality on multiple links. In the proposal, the authors use a local coordinator located in close proximity to the cell to enable rapid path switching in the event of link failure. They also demonstrate that the path switching mechanism using dual-connectivity is faster than a handover. In [90], the authors propose a network slicing based mobility management model for UE mobility.

Some of the research on UE mobility in slice-based 5G networks [43,91] is focused on defining SDN based architectures and defining handover procedures within those frameworks. Authors in [43] propose an architecture for a 5G system with a two-tiered RAN made up of macro and femto base stations, which share the spectrum. In this work, the authors also illustrate a slice-based handover procedure within the architecture and propose a scheme for allocating power and sub-channels for the network slice. In another work [92], authors aim to reduce handover signaling cost in dense networks using multi-connectivity. This is attained by using a split control plane and data plane architecture and maintaining connections to multiple APs at a time so that the service remains uninterrupted. In another work [93], performance improvements obtained by using multi-connectivity for URLLC in HetNet scenarios are evaluated. Authors in [94] focus on inter-macro base station handovers for SDN based networks in a high-speed railway scenario. They define a scheme to reduce handover failures through coordination between two macro base stations by proposing changes on the protocol stack and by replicating RRC signaling on both the base stations.

In [90], the authors exploit the property of localization of slices to certain areas within the network given that for certain types of slices, UEs may remain stationary. They propose protocols for slice mobility with the aim to reduce control signaling by eliminating the location tracking functionality for stationary UEs and reduce location update frequency and paging frequency.

In [95], the authors propose an architecture for slice-based mobility management for HetNets. They propose a scheme for offloading flows within the same slice across different RATs. The authors in [96] propose an architecture based on SDN and NFV for managing network slices and their associated resources dynamically. The authors propose to use different elements for handling intra-slice and inter-slice mobility. In [97], authors propose the design of a flexible SDN/NFV based 5G architecture with a focus on network slicing. They analyze and differentiate scenarios where UE mobility management is required to be performed within and across slices and illustrate the same through call flows.

Unlike our proposal, studies on service availability through multi-connectivity appear to be limited to a single service. Although these works explore and provide solutions for slice mobility and multi-connectivity separately, they do not look at slice mobility and slice-specific service delivery cohesively. To the best of our knowledge, our work is the first work to propose the use of multi-connectivity as a protocol to facilitate UE slice mobility. To better understand the nature of the proposed enhancements, we first review multi-connectivity as defined by the 3GPP 5G network and then proceed to highlight the changes proposed.

6.3 Multi-connectivity in the 3GPP 5G Network

Multi-connectivity is a protocol that enables multi-mode UEs to connect to two heterogeneous RAN nodes, such as 5G NR femto gNB and a 5G NR macro gNB at the same time [88]. Traffic corresponding to a PDU session is sent/received over these multiple radio connections. Although 3GPP defines four variants of connectivity based on the core network node and RAN node type, e.g., Evolved Universal Terrestrial Radio Access (E-UTRA)-NR dual connectivity for 4G-5G node dual-connectivity, NR-NR Dual Connectivity (NR-DC) for 5G-5G node dual-connectivity, we focus our discussion on NR-DC. Figure 6.1 illustrates the connections for a particular UE, which is dual connected to two 5G gNBs. As illustrated in the figure, the dual-connected system consists of two gNBs, one of which is known as the Master Node (MN) and connected to the core network through the NG interface. The MN connects to the UE through the air interface (Uu). The signaling to the core network may be transferred through the MN.

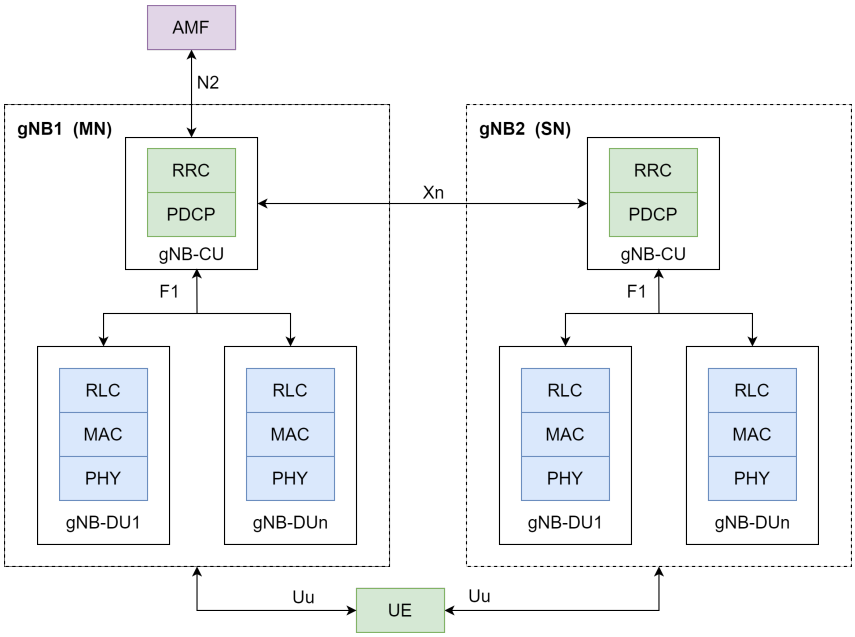


Figure 6.1: Block diagram of 3GPP NR multi-connectivity architecture

The other gNB known as the Secondary Node (SN) is connected to the MN over the Xn interface. The MN behaves as the master for the UE connection and is responsible for setting up PDU sessions in the RAN. The SN is also connected to the UE through Uu interface, and it can send radio control related configurations to the UE using the Signaling Radio Bearer 3 (SRB3) message. Also, as illustrated, each gNB consists of units known as the gNB-CU and one or more Distributed Units (gNB-DUs) under the control of a single gNB-CU. Typically, multi-connectivity is used in HetNet deployments where the MN is a larger coverage macro-gNB, and the SN is a lower power gNB with a smaller coverage area that overlaps with the MN's coverage area. Also, protocols such as LTE WLAN Aggregation (LWA) [98] and LTE WLAN integration over IPSec (LWIP) were defined for multi-connectivity to other RATs in previous 3GPP releases. However, at present, multi-connectivity between 3GPP 5G NR and WLAN has not been defined in 3GPP specifications [88].

We propose a few enhancements in the multi-connectivity protocol that not only remedy the above-mentioned issues but also enhance UE slice mobility and slice availability. The proposed changes are described in the subsequent section.

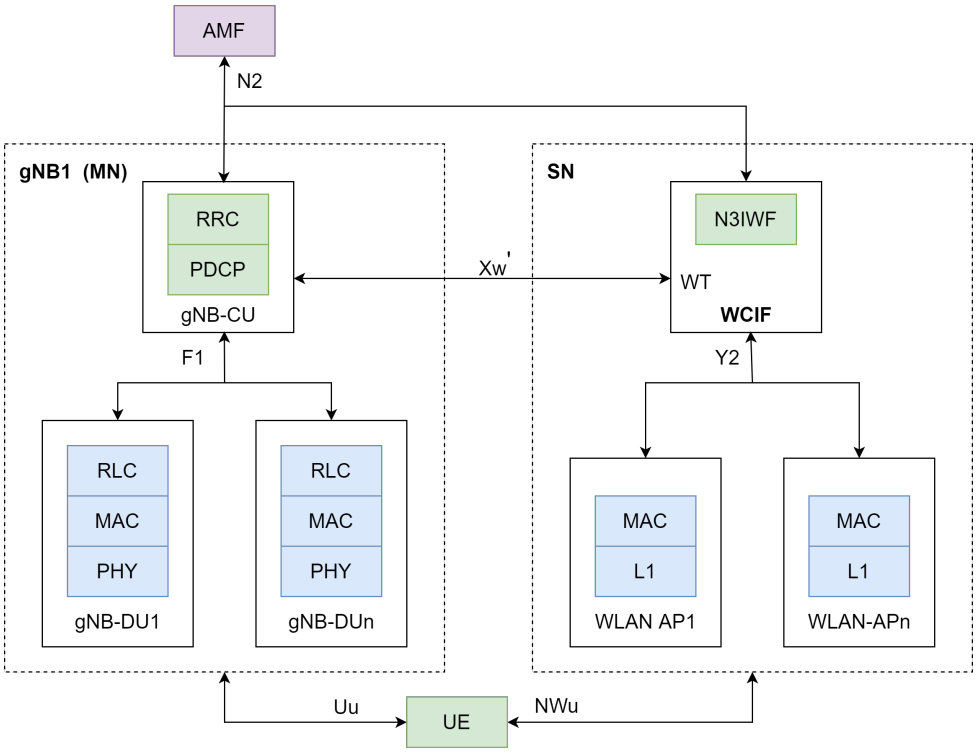


Figure 6.2: Block diagram of proposed NR-WLAN multi-connectivity architecture

6.4 Proposed Enhancements to 3GPP Multi-connectivity

We propose that the multi-connectivity protocol be extended to allow UE to connect to two (or more) BSs having similar cell sizes (such as two femto gNBs) or those having differing cell sizes (such as a macro gNB and a femto gNB). They may even belong to different RATs such as a macro gNB and a WLAN AP. By doing this, slice availability can be improved, especially when BSs support disparate slices. The procedure for connecting 3GPP NR nodes such as two femto gNBs is the same as that used for 3GPP NR multi-connectivity.

The proposed system diagram for enabling multi-connectivity between NR gNBs and WLAN APs, with coordination in RAN, is illustrated in Figure 6.2. The proposed protocol is adapted from LWA [98] and multi-connectivity specifications [88]. 3GPP defines a Wireless Termination (WT) point as part of LWA for connecting LTE with a non-3GPP network, the placement of which is implementation-dependent. Similar to the 3GPP defined WT point for LWA, we define a WLAN Controller (WC). It is responsible for managing WLAN APs from a multi-connectivity perspective. It can be placed together

with the Non 3GPP InterWorking Function (N3IWF), a function defined by 3GPP to connect WLAN APs to the 5G core network. This combined (WC+N3IWF) function, referred to as "WLAN Control Interface Function (WCIF)", is connected to gNB-CUs over an interface (Xw′), akin to the 3GPP LWA Xw interface. The procedures for slice mobility and slice availability for the enhanced system are presented in Section 6.5.

6.5 Procedures for Slice Mobility and Availability

In this section, we describe the call flows for handover and PDU session establishment procedure for illustrating slice mobility and availability in the enhanced system. The call flows are also used to describe the benefits obtained in a qualitative manner.

6.5.1 Slice Handover using Multi-connectivity

Firstly, we describe slice handovers in the enhanced system for a scenario where UE handover occurs within the 5G NR RAN. Then we proceed to describe the UE handover procedure between 5G NR and WLAN.

6.5.1.1 Scenario 1: Slice Handover within 5G NR

Figure 6.3 illustrates the call flow for UE slice mobility considering slice handover support as defined by 3GPP [53]. As illustrated in the figure, the UE supports 'n' slices with 'm' active PDU sessions. The active PDU sessions correspond to the number of slices that the UE is connected to at present, with one PDU session per slice. When the received signal strength of the serving cell at the UE falls below a given threshold, the source gNB initiates a handover to a possible target cell. Note that different algorithms for target cell selection and different criteria (e.g., signal strength, support for slices) may be used by the network provider at the source gNB to select the target cell/gNB. However, after the target selection, the proposed 3GPP protocol ensures that the UE is handed over to a single target cell/gNB.

As illustrated in Figure 6.3, handover is initiated towards the chosen target cell (gNB) by sending a "Handover Request" message. The target gNB can only admit the PDU sessions mapped to the supported slices (S-NSSAIs), and the remaining sessions are rejected.

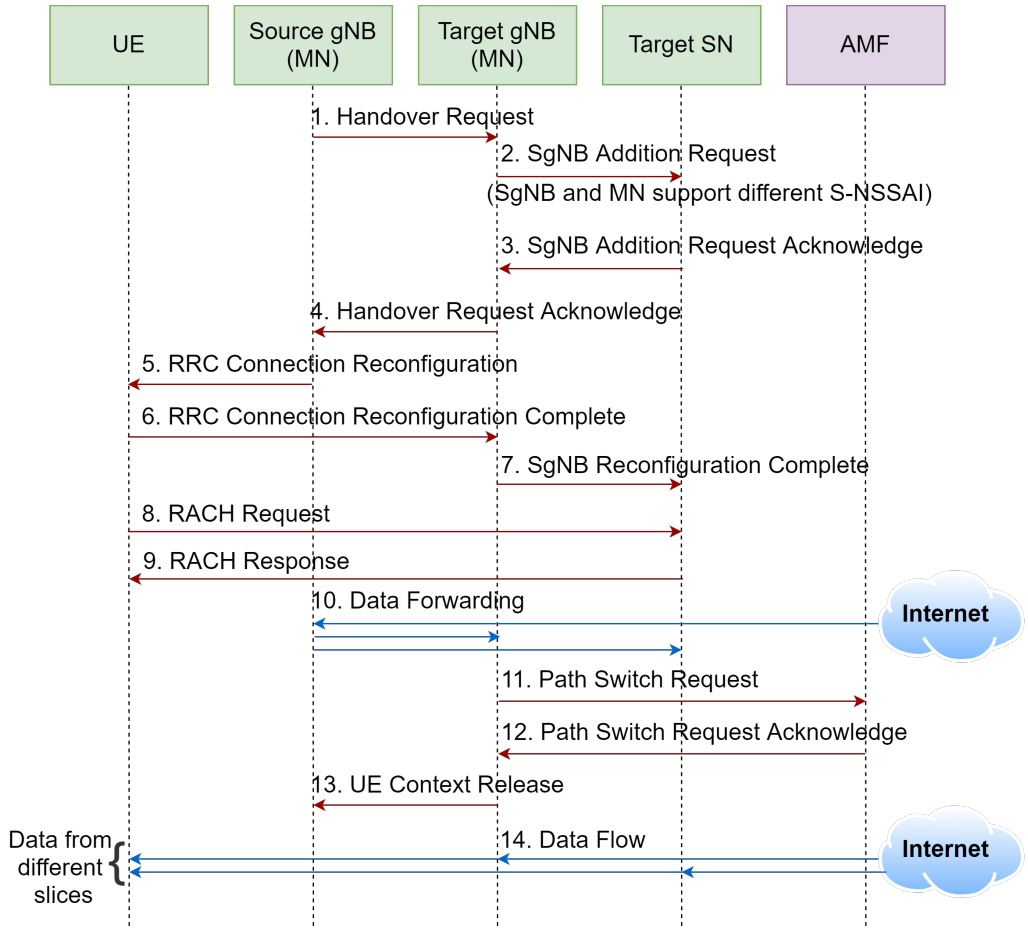


Figure 6.3: 3GPP 5G slice mobility call flow (courtesy [53])

The information of the admitted sessions is then sent to the source gNB using the "Handover Request Acknowledge" message. Only the admitted sessions are then handed over to the target. The handover is followed by path switch request/response signaling to indicate the new endpoints for the GPRS Tunneling Protocol (GTP) tunnels as the connected gNB is now changed. Note that if the target cell does not support a particular slice, then the PDU sessions belonging to that slice can not be handed over and leads to session discontinuity. This is especially true of practical deployments, as some of the slices are available only within certain areas of the network [53]. Also, even if all the slices (S-NSSAIs) are supported on the target gNB, handovers may also fail due to the unavailability of resources for some (or all) of the slices.

As mentioned earlier, the procedure for slice handover remains similar with the exception that UE multi-connectivity to similar types of nodes is also enabled. By using multi-connectivity, we can enable concurrent connections from a UE to multiple RAN

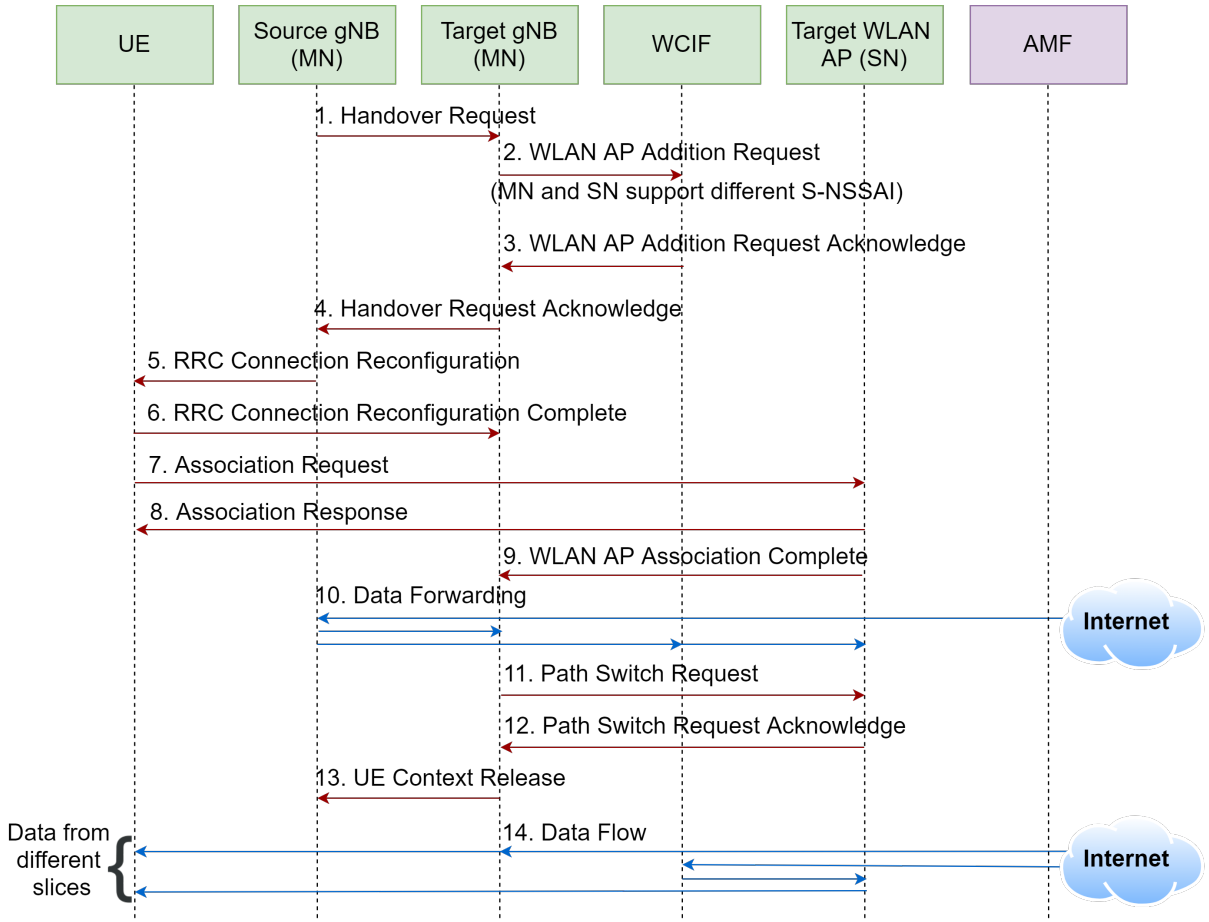


Figure 6.4: Proposed slice handover to both 5G NR and WLAN using multi-connectivity

nodes, especially when the nodes support different slices. As a result, handover failures that result when all active slices are not supported on a single RAN node are avoided.

6.5.1.2 Scenario 2: Slice Handover between 5G NR and WLAN

A similar procedure can be followed for handover within nodes comprising multiple RATs when one of the target nodes may be a WLAN AP. The signaling procedure when the target SN is a WLAN AP is illustrated in Figure 6.4. WLAN AP is added as SN once the "WLAN AP Addition Request Acknowledge" message has been received by the target gNB (MN) and is acknowledged. At this stage, UE context has been set up on both the MN and SN nodes. UE is then associated with the target SN. The target SN admits the UE by sending "Association Response" as UE context has already been setup. WLAN AP then informs the target MN about UE association through "WLAN AP Association Complete" message. The rest of the procedure consisting of data forwarding and path

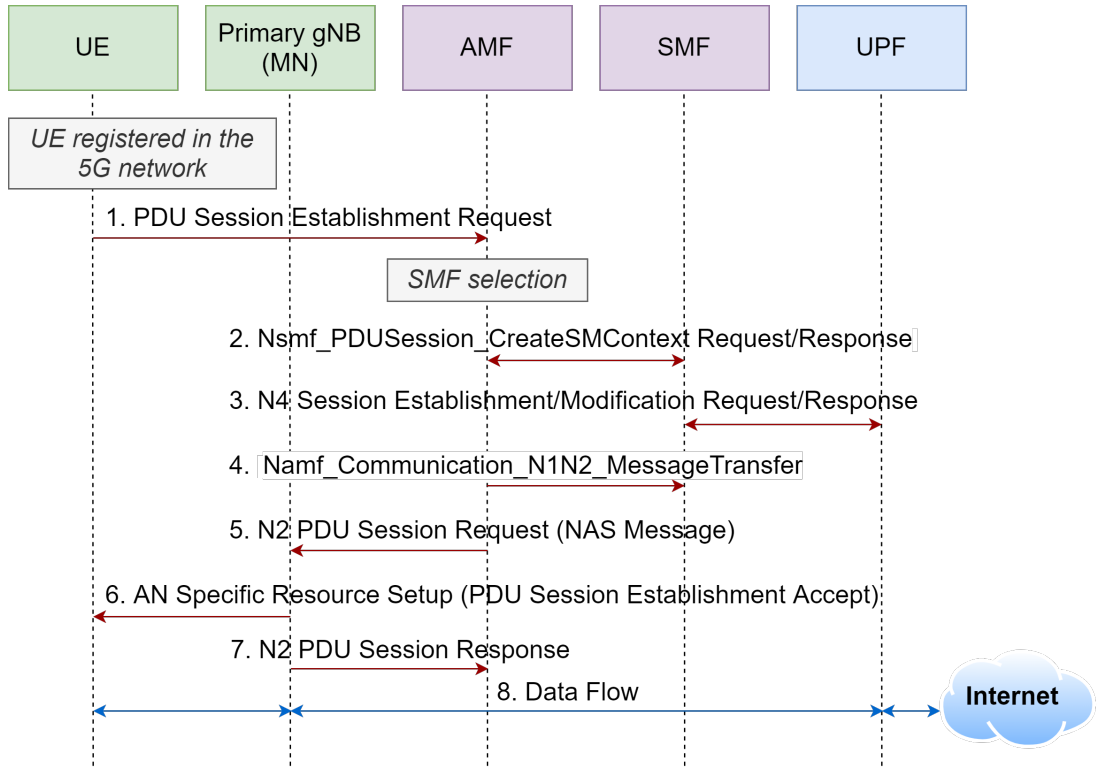


Figure 6.5: 3GPP 5G PDU session establishment (courtesy [58])

setup is similar to that of 3GPP handover.

6.5.2 Proposed Slice Availability Enhancements

Network slice availability within the 3GPP 5G network is decided during UE registration. A UE provides a list of slices that it wishes to access at the time of registration. After registration, the UE is notified of the slices that it can access by the AMF. However, the actual slices that are accessed by UE are determined during the PDU Session Establishment procedure (data session establishment), as illustrated in Figure 6.5.

Increased slice availability helps in reducing mobility failures arising due to lack of slice support on the target BS. It also improves service availability, especially in dense HetNet environments where resources may be constrained. Instead of deploying every slice over all nodes within the RA as stipulated by 3GPP, we propose to increase slice availability by suitably multi-connecting UEs. This could be performed at the time of PDU session establishment, as slice resources are actually accessed at that time. We illustrate the proposed procedure when UE is simultaneously connected to either two

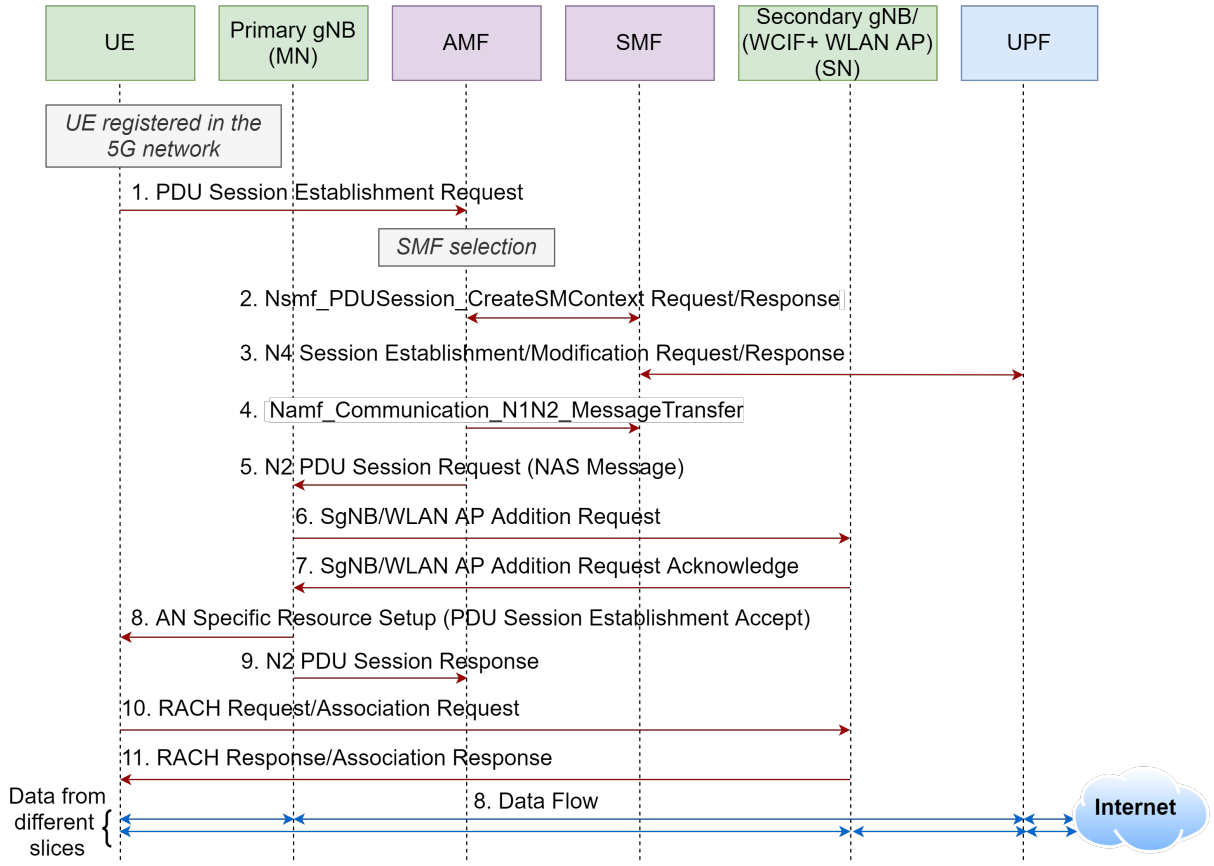


Figure 6.6: Proposed session establishment procedure

NR nodes/one NR node and a WLAN node in Figure 6.6. The proposed procedure is a modified version of the 3GPP UE Session Establishment call flow.

To improve slice availability in the enhanced system, we account for the unavailability of slices on a particular RAN node. This may occur due to lack of slice support on the RAN node or due to overloading. For example, the existing 3GPP session establishment procedure mandates the AMF to inform gNB regarding the sessions that were established for supported slices using an N2 message, "PDU Session Request". Further, if a gNB is unable to support a slice-specific session due to resource constraints, these PDU sessions are rejected. In contrast, within the proposed system, the gNB tries to establish the sessions that were unsupported by using multi-connectivity, i.e., by adding a suitable secondary radio node and then establishing these PDU sessions over the secondary gNB. The algorithm for the choice of the secondary node is left to be implementation dependent. The actual number of the data sessions that can be supported due to the secondary node addition depends on the availability of support for the required slice type(s) on

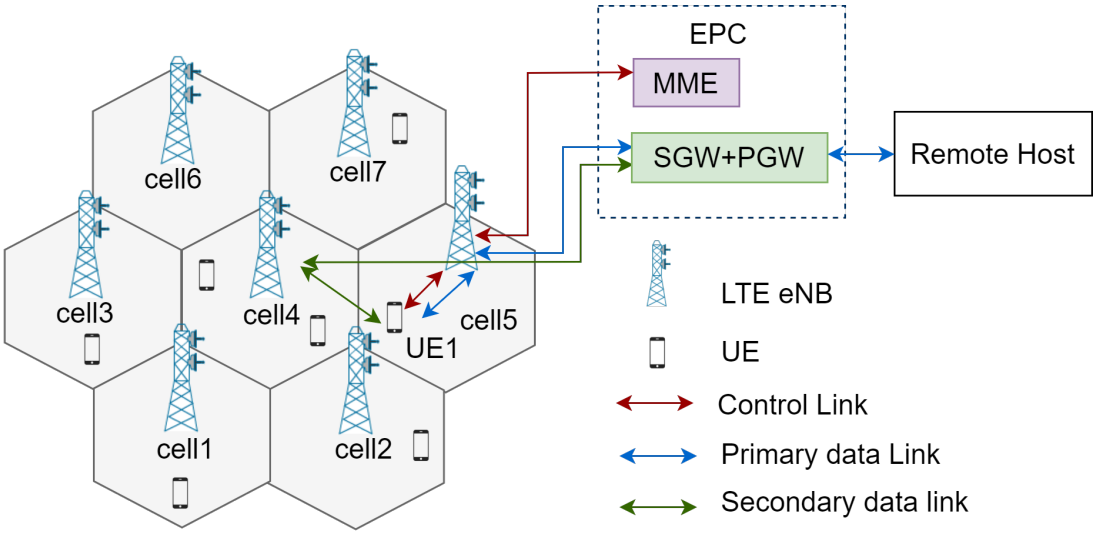


Figure 6.7: Simulation setup with LTE

the secondary nodes and the degree of multi-connectivity supported by the UE. Once a suitable secondary node is added, the gNB sends an "AN Specific Resource Setup" message with the list of available slices to UE, and the required radio configurations are performed on the UE. Confirmation is also provided by gNB to AMF in the form of an "N2 PDU Session Response" message to account for any session failures. Data for established sessions is then received at UE through multiple paths.

6.6 Evaluation of the Proposed Protocol

In order to evaluate the performance of the proposed enhancements, we evaluate metrics such as system throughput, handover failures, and slice availability. For this, we simulate a 7 cell cluster in ns-3 [28]. We use LTE as the reference RAT in the simulations, as to the best of our knowledge, an open-source simulator for 5G NR RAT is unavailable at present. The result for slice handover in LTE will be similar to that in 5G NR as the handover behavior of both these RATs is similar. This cluster, as illustrated in Figure 6.7, consists of eNBs placed in the center of the cell. The setup also consists of an EPC connected to the eNBs and a remote host with which data is exchanged. As in practical deployments, different slices are supported within different areas of the network. We consider 3 UDP applications corresponding to different slices, each with a data rate of 10Mbps. We carry out simulations for 10 different slice distribution configurations detailed in Table 6.2, where slices are flexibly distributed within the network. The simulation parameters are provided

Table 6.1: Simulation parameters for LTE and WLAN

Parameter	**Value**
Path loss	$128.1 + 37.6\log(R)$, R in kms
Tx power for LTE eNB and UE	46 dBm, 23 dBm
Antenna Type LTE eNB	Isotropic Antenna
Antenna Height of LTE eNB and AP	32 m, 2.5 m
Handover Algorithm Used	A3-RSRP
Hysteresis value	3 dB
Inter-site distance for LTE eNBs	500 m

Table 6.2: Slice distribution in LTE

	Config1	**Config2**	**Config3**	**Config4**	**Config5**	**Config6**	**Config7**	**Config8**	**Config9**	**Config10**
Cell1	1, 2	2, 3	1, 3	1, 3	1, 2, 3	1, 3	1, 2	2, 3	1, 2	1, 3
Cell2	1, 2	1, 3	2, 3	1, 2	1, 3	1, 2	1, 2, 3	1, 2	3	1, 2
Cell3	1, 3	1, 2	2, 3	2, 3	2, 3	1, 2	2, 3	1, 2, 3	2, 3	2, 3
Cell4	1, 2, 3	1, 3	2, 3	1, 3	2, 3	1, 3	1, 3	2, 3	3, 1	1
Cell5	1, 3	2, 3	1, 2	1, 2, 3	1, 2	2, 3	1, 2	1, 2	1, 2	1, 2, 3
Cell6	2, 3	1, 3	1, 2”	1, 2	1, 3	1, 2, 3	2, 3	1, 2	1, 2	2
Cell7	2, 3	1, 2	1, 2, 3”	1, 3	1, 2	2, 3	1, 2, 3	1, 2, 3	1, 2, 3	2, 3

in Table 6.1 and have been obtained from [75]. We assume that users are randomly distributed within the system and use the random way point model to characterize user mobility. The users move with a maximum velocity of 40km/hr. We also assume that the user arrivals follow a Poisson distribution with an arrival rate of λ arrivals per second and have a service rate of μ within the system.

For the multi-connectivity scenario, whenever a UE is handed over to a target cell without support for a given slice, we create another connection to a suitable secondary eNB. This is illustrated in Figure 6.7 where a UE (UE1) in cell 5 is multi-connected to the eNB in cell 5 as the MN and an eNB in cell 4 as the SN. The secondary eNB is chosen such that it has a suitable RSRP value. It should also support the slice that is required by the UE but is unsupported by the target Master eNB. Note that this secondary connection is torn down if the UE is again handed over and the new target cell supports all the required slices.

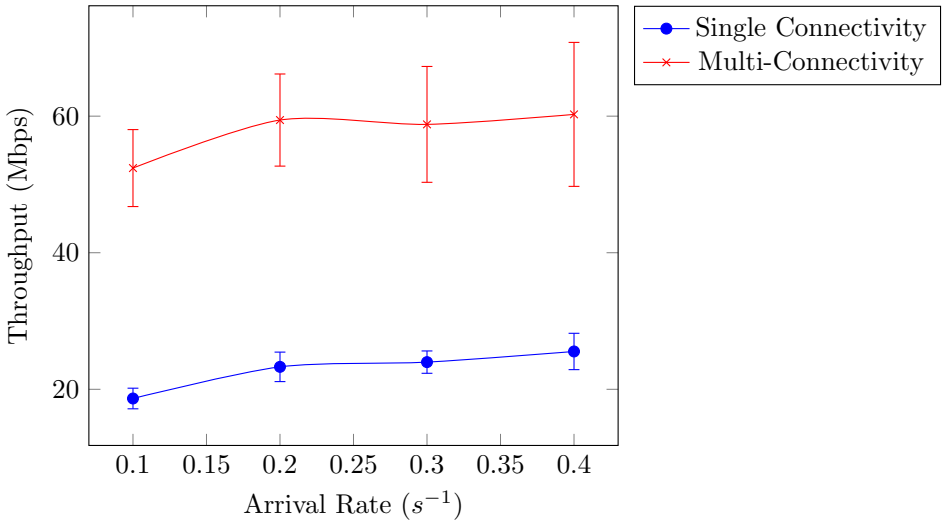


Figure 6.8: Median system throughput v/s user arrival rates for different connectivity types.

6.6.1 System Throughput

We also evaluate the system throughput for various slice configurations in LTE by varying the user arrival rate. The median system throughput for 10 different configurations with single connected users and multi-connected users have been depicted in Figure 6.8.

The confidence intervals across 10 configurations have been plotted on the same graph with a line. As shown by the results, the system throughput in a multi-connected system is higher than that of the single-connected system. This is because PDU session continuity is maintained on handover by connecting to a secondary eNB if the target eNB does not support one or more slices present on the source eNB. Other than maintaining service continuity, we can observe that the resources from multiple eNBs are utilized for the sessions, thus improving resource utilization.

6.6.2 Rate of Handover Failure

We measure the rate of handover failure for various user arrival rates (λ) for both the protocols. The rate of handover failure is defined as the ratio of the number of failed slice handovers due to the lack of slice support on the target node, i.e., eNB/WLAN AP to the total number of slice handovers. The service rate (μ) is maintained at 1 user per second. We measure the handover failure rates for single connected users and also for multi-connected users for various slice configurations. This metric is evaluated for both

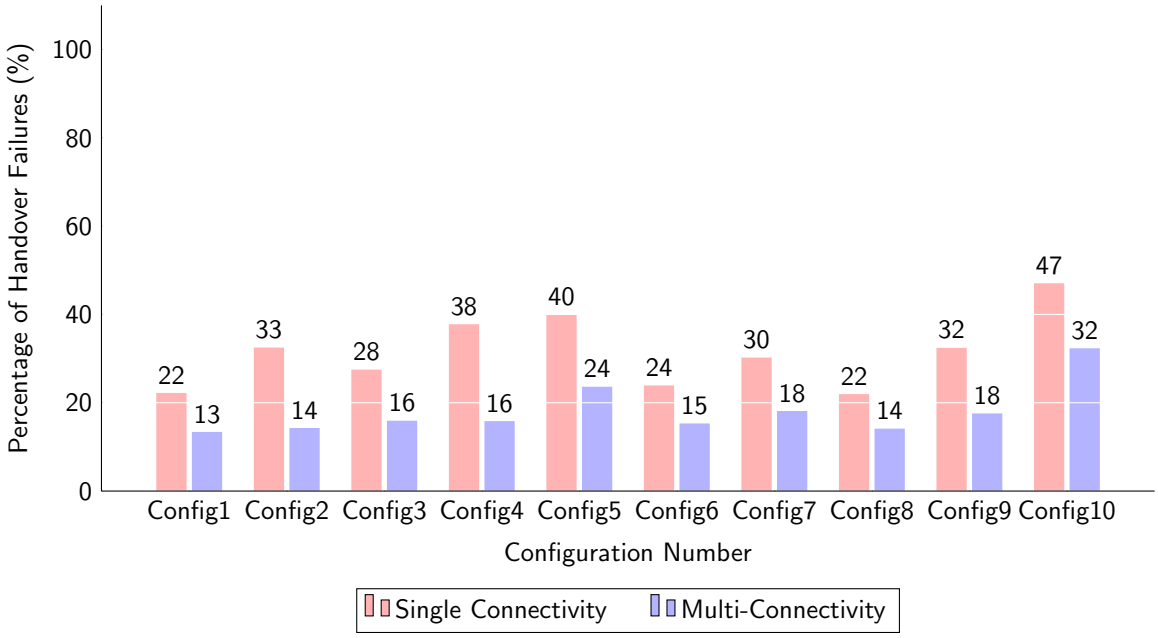


Figure 6.9: Percentage of handover failures for various slice configurations in LTE

LTE as well as a multi-RAT case with LTE eNBs and WLAN APs. The results for LTE illustrated in Figure 6.9. As seen in Figure 6.9, the rate of handover failure is reduced by using multi-connectivity, with an average improvement of 13.53% for the configurations considered in our simulations.

The multi-RAT setup consists of 32 WLAN APs uniformly distributed within the coverage area of 7 cell cluster. WLAN APs are expected to support a maximum of two service types (slices) at a time. We assume that slice support across RAN nodes is not homogeneous. In this case, we consider 3 UDP applications corresponding to different slices, each having a data rate of 10 Mbps, 5 Mbps, and 2 Mbps, respectively. We perform simulations for 10 different slice distribution configurations for LTE. These configurations are mentioned in Table 6.2. For WLAN, we provide the configurations shown in Table 6.3.

Table 6.3: Slice distribution in WLAN

AP Numbers	**Slices**
AP2, AP7, AP10, AP12, AP13, AP17, AP20, AP22, AP27, AP30, AP32	1
AP1, AP3, AP5, AP6, AP8, AP9, AP11, AP15, AP16, AP19, AP21, AP25, AP26, AP29, AP31	2
AP4, AP14, AP18, AP23, AP24, AP28	1, 2

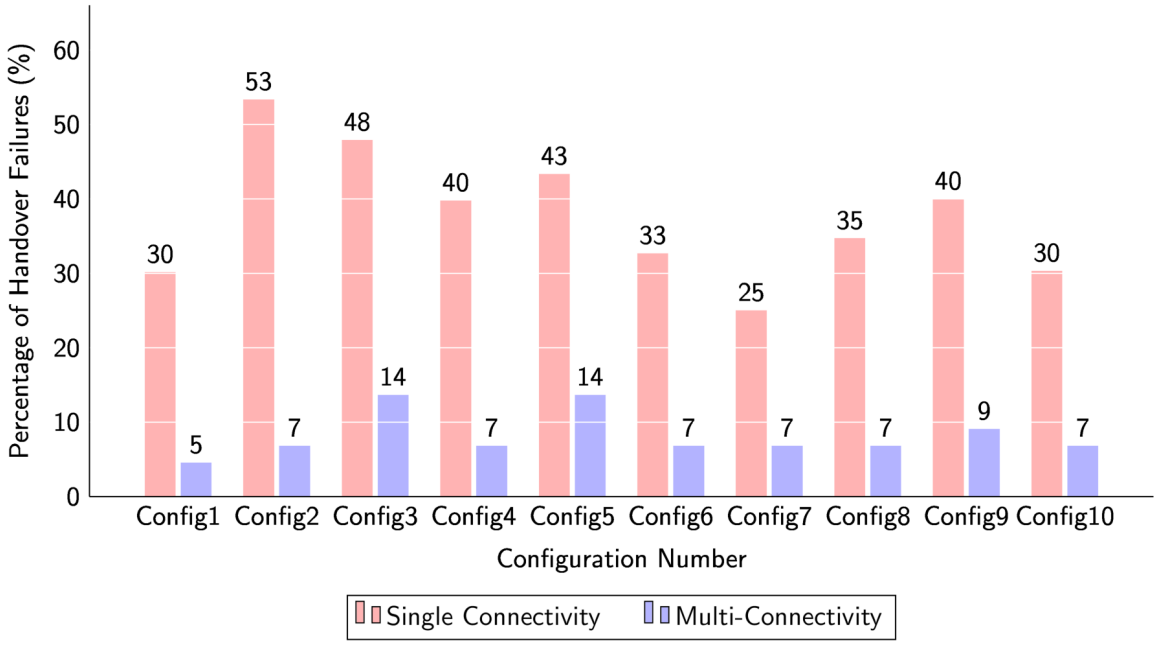


Figure 6.10: Percentage of handover failures for various slice configurations for multi-RAT network

As shown in Figure 6.10, the rate of handover failure is reduced by employing multi-connectivity, with an average improvement of between 26.7% to 35.839% for the configurations considered. As handover failures may cause data session discontinuity affecting the user Quality of Experience (QoE), a reduction in handover failure improves the QoE for a given user.

The results in Figure 6.10 demostrate a reduction in the rate of handover failure for a fixed number of WLAN APs. An increase in the density of WLAN APs is expected to further reduce the percentage of handover failures if the WLAN APs are able to support the slices requested by the UE. In the considered multi-RAT scenario, even when the required slices are not supported on the target WLAN, if the UE was handed over from LTE to WLAN, the UE would be able to maintain a secondary connection over LTE (due to the larger coverage area of LTE) for the unsupported slice, thereby reducing handover failures. However, the exact number of APs needed to reduce the handover failures to below a certain percentage needs to be studied further.

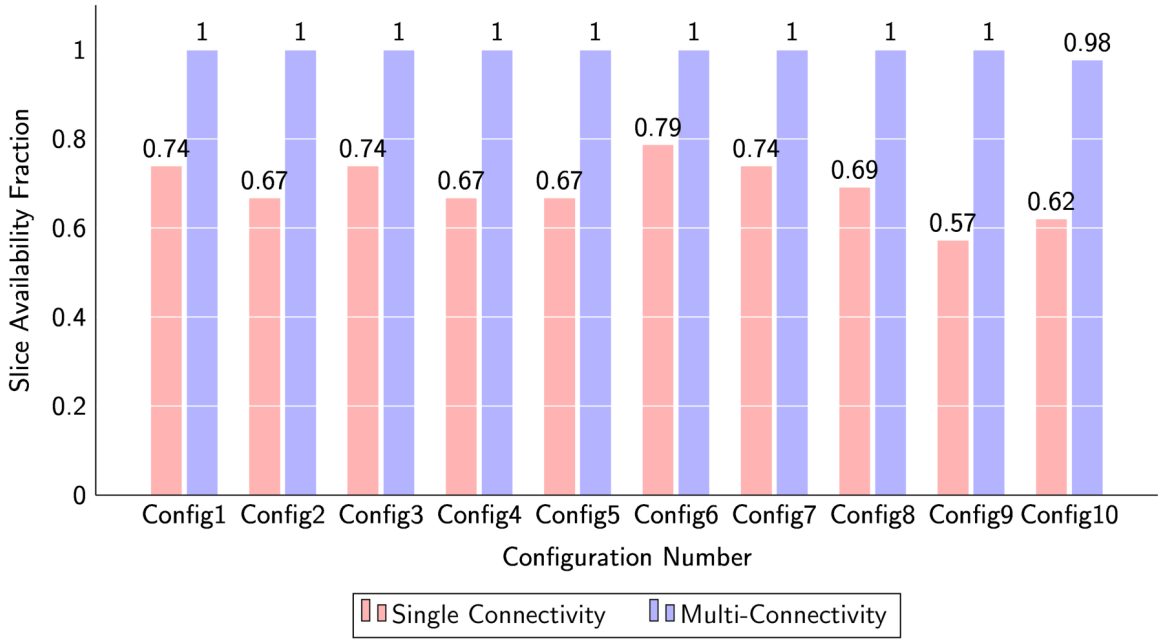


Figure 6.11: Slice availability for various configurations in multi-RAT network

6.6.3 Slice Availability Fraction

Slice availability is calculated as the ratio of the number of slices available to a UE at the time of its arrival versus the total number of slices supported by the multi-RAT network. For example, if at the time of its arrival, a UE can access 2 slices and the total number of slices supported by the network is 3, the slice availability fraction is given as $\frac{2}{3}$. The slice availability fraction for single connectivity versus multi-connectivity (with one additional radio link) is then evaluated for all UEs in a multi-RAT scenario. Availability is averaged over a time interval of 100 seconds, observing multiple user arrivals. As seen from Figure 6.11, slice availability improves considerably with multi-connectivity and is nearly equal to the desired ratio (availability fraction=1) for the considered scenarios.

6.7 Conclusion

In this chapter, we have provided an overview of the existing UE slice mobility protocols and highlighted the need for flexible network slice deployment for minimizing network costs. To this end, we have proposed enhancements to 3GPP multi-connectivity to improve slice availability and slice mobility performance in 5G multi-RAT networks. The

proposal not only results in bringing about flexibility in network slice deployments but also has the potential to reduce the capital and operating expenses for the network. Through this solution, we address the gap within the 3GPP specifications by defining a mechanism to achieve multi-connectivity between 3GPP NR and WLAN. We are also able to perform UE handover between 3GPP and WLAN through coordination between the RAN nodes, which is not supported by 3GPP specifications today. The proposal is suitable for integration into the 3GPP specifications as there are marginal changes in signaling towards UEs. It is also aligned with the concept of ultra dense network deployment in 5G and beyond networks. Finally, we have also demonstrated the flexibility brought about by this scheme through ns-3 simulations.

Chapter 7

Conclusions and Future Research Directions

We have explored several underlying themes concerning the multi-RAT nature of the present-day and next-generation cellular broadband networks in this thesis. The salient theme within the described content is to identify gaps in existing standards and propose SDN based solutions that can be integrated into the next-generation standards with minimal changes to the UE. Additionally, we have proposed a few ideas that can serve as a premise for future standardization activities. In summary, the thesis offers a holistic solution for SDN based multi-RAT network control and management, beginning from multi-RAT architecture design, slicing implementation and finally, deployment.

In this chapter, we summarize the key contributions of the thesis and discuss ideas for future exploration.

7.1 Contribution Summary

The key contributions of this thesis are summarized below.

- **Defining SDN based architectures for multi-RAT networks:** Based on our study of the present-day multi-RAT networks, we have inferred that logical centralization of control is essential for better resource allocation and network performance in multi-RAT networks. As a first step, we have proposed enhancements to the

3GPP 5G network for logically centralizing control within the 5G network. This work, described in Chapter 3 is aligned with the SDN paradigm. We have demonstrated that the proposed architecture has a better performance in terms of control signaling latency qualitatively through call flows and quantitatively through ns-3 simulations.

Later, we have extended this work to incorporate control and management of multi-RAT networks in Chapter 4. The proposed architecture has been designed to provide end-to-end control for multi-RAT networks in a unified manner. This work provides a virtualized view of the network resources, simplifying network control and management. Additionally, the proposed architecture provides RAT-agnostic interfaces to applications. More importantly, the proposed design simplified the implementation of features requiring interaction across multiple nodes, e.g., multi-connectivity compared to existing 3GPP 5G/4G networks. The design also resolves scalability issues present in most existing works while incorporating support for network slicing. We can support slices with a diverse set of requirements within the proposed architecture by implementing slice-specific controllers and slice-specific data plane functions. Lastly, we have evaluated the performance benefits provided by this architecture through a simple user-association algorithm in the multi-RAT framework.

- **Defining frameworks for slicing multi-RAT RANs:** During the course of architecture design, we also observed the absence of mechanisms for multi-RAT RAN slicing. Our study of existing prior art established that most solutions were specific to the RAT and did not provide generic guidelines/learning that could be used for slicing newer RAT types. Moreover, the present 3GPP guidelines for slice deployment are restrictive, requiring homogeneous slice deployments in a geographic area. While this requirement ensures slice mobility, it proves to be expensive due to high CAPEX and OPEX. We have addressed these problems in a two-fold manner. As the first step, we have devised a set of generic guidelines for slicing the RAN. We have also defined VirtRAN, an example recursive slicing framework based on the guidelines provided. This ensures that the framework design is customizable for future requirements. This work has also been submitted to an IEEE 5G standardization activity [27] as a plausible solution to address the current gaps in standards.

- **Defining enhancements for UE slice mobility and availability:** To enhance UE slice mobility and avoid the need to deploy slices homogeneously, we have proposed the use of UE multi-connectivity. This solution involved enhancements to the 3GPP multi-connectivity protocol can also be used to perform RAN aggregation between 5G and WLAN, which is not defined by the 3GPP 5G standards at present. We have evaluated the advantages of the system through simulation results.

- **Identification of gaps in standards and providing possible solutions:** In the process of designing SDN based architecture for multi-RAT networks, we have identified key gaps in standards, e.g., lack of inter-working mechanisms between WLAN and NR RAN, absence of slicing framework for RAN, etc. To correct these deficiencies, we have proposed viable solutions and have submitted some of them to standardization activities. We have also gained fresh perspectives on RAN disaggregation, such as viewing the network as a sum of its basic functionalities. This work is currently in progress, and an early version of the same has been submitted as a contribution to IEEE P1930.1 [27]. Further details of the same are provided in Section 7.2.

7.2 Future Research Directions

In this work, we provide a point of departure to a broader set of problems related to SDN multi-RAT architecture design and network slicing. Some of the potential extensions are listed below.

- **Architecture design with data plane optimization:**
  In Chapters 3 and 4, we have designed the end-to-end network architectures from the perspective of minimizing the control signaling overhead. Similarly, we can also attempt to optimize the data plane. For example, within the core network, an alternative transport protocol to GTP could be evaluated, as GTP has a high processing overhead.

- **Algorithm design for SDN based multi-RAT architectures:**
  The multi-RAT architecture proposed in Chapter 4 abstracts RAT-specific details

and provides a common set of parameters to the applications. By using the design as a basis, we can study the design of efficient centralized multi-RAT RRM algorithms.

- **Enhanced slicing framework design:**

  We have provided a generalized solution for slicing the multi-RAT RAN through VirtRAN. By evolving this solution to include the mobile core network, we could provide a generalized framework for end-to-end network slicing.

- **Algorithms for flexible slice deployments:**

  We have addressed the problem of slice mobility by suggesting a multi-connectivity based approach for non-homogeneous deployments in Chapter 6. However, this solution can be taken a step further by devising algorithms for flexible slice-deployments in multi-RAT networks considering various factors such as deployment costs, the number of concurrent connections allowed on the UE, availability of network resources, etc. These algorithms can also be designed to use online approaches where the requirements change dynamically in real-time.

- **Design of multi-RAT architectures for specialized use-cases:** With advancements in technology, newer applications/use-cases are introduced into the network. While we have a lot of flexibility in designing architectures to support specialized use-cases with green-field deployments, this may not be true for enhancing existing deployments. It may also be necessary to customize the existing network architecture to support newer applications for monetizing the existing infrastructure. Therefore, a practical area of research is to use a combination of existing RATs with SDN control to support advanced features/applications without the need for upgrades.

  We have explored one such architectural solution for increasing LTE network coverage for mission-critical scenarios, when Device to Device (D2D) communication is not supported by the LTE network. D2D is a feature that enables LTE UEs in proximity to communicate with each other without the intervention of the LTE eNB. However, the implementation of this feature is complex in practice.

  There are also additional limitations such as support for only single-hop relay communication and a distributed manner of relay selection which may not provide an

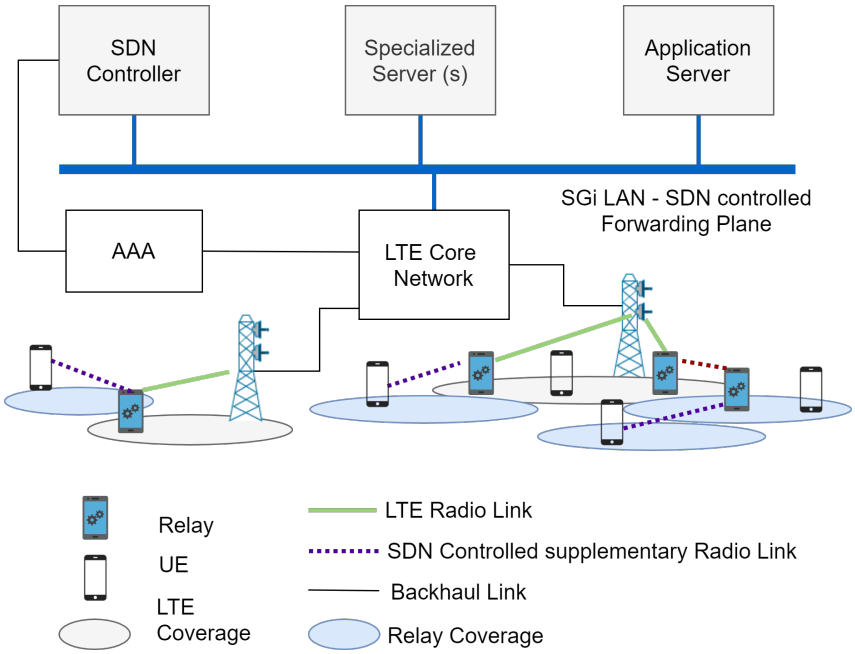


Figure 7.1: Proposed SDN based overlay network architecture

optimal path for data transfer. [1]

We have attempted to solve this problem by introducing an overlay network over the LTE network. The overlay network comprises UEs, relays (UEs augmented with additional RAT support), and other network nodes such as specialized servers, e.g., application servers required for supporting mission-critical services. Primarily, the UE accesses the application server providing the service through the LTE network. In case of loss of coverage, e.g., obstructions in buildings, etc., the UE can still communicate with the server via a relay over the supplementary radio. The relay is connected to the LTE network, either directly to an LTE eNB or through other relays. This solution is easier to implement as most UEs today are multi-mode UEs supporting RATs such as WLAN, Bluetooth in addition to LTE. The proposed architecture is illustrated in Figure 7.1. The design consists of an SDN Controller which controls the overlay network and also configures the devices to behave as relays when required. As the network is centrally controlled, it is possible to authenticate devices, manage security on the UE-relay link, and also control interference within the network. Also, unlike the existing solution, multi-hop relay communication can

[1]The solution described here is published in [99] and was supported by the Ministry of Electronics and Information Technology (MeitY), Government of India through a sponsored project grant on 5G.

be supported. This solution opens up further possibilities for research relating to relay selection, overlay network control, etc.

- **Dis-aggregating the multi-RAT RAN:** Dis-aggregated network architectures provide the ability to deploy NFs in a distributed manner. 3GPP enables RAN dis-aggregation in 5G by splitting the 5G gNB into gNB-CU and gNB-DU. However, no such approaches are available at present for multi-RAT networks. We have proposed an early-stage solution to this problem by leveraging the fact that all RATs perform similar functions, albeit with varying implementations. We have identified 6 such NFs which can be used as building blocks for any RAT. Any RAN node such as 5G gNB, WLAN AP, etc., can be represented as a combination of these functions chained in a specific order. Note that not all of these NFs may be present in every RAT. The description of these NFs is provided below.

  - **Base Station Function (BSF)**: This network function is responsible for radio transmission to (and reception from) the UE. It has functionalities for radio transmission and reception, error control and may also support related control functionality such as packet scheduling, link adaptation, and power control. Typically the BSF would comprise radio transmission/reception functionality along with PHY and MAC layer.
  - **Adaptation Function (AdpF)**: AdpF is responsible for adapting data flows for the underlying layers supported by the BSF. The nature of the adaptation is based on the service requirements of the flow and can include functionality such as handling packet loss through ARQ, packet re-ordering, and in-sequence delivery of packets.
  - **AgrF**: The Aggregation Function (AgrF) is aggregates/dis-aggregates data traffic received from or sent towards multiple BSFs. It then exchanges this traffic with internal entities such as the Inter-working Function (IWF) or external entities such as UPF.
  - **OptF**: The Optimization Function (OptF) performs optimizations on data traffic. Example optimizations include IP header compression/decompression, flow classification, metering and dropping packets, address translation, etc.

– **Security Function (SF)**: The SF is responsible for ciphering and integrity protection of user data and UE-specific control messages being sent over the wireless medium.

– **Inter-working Function (IWF)**: IWF is responsible for ensuring inter-working of RAN with other networks such as 4G/5G core network, external data network, etc.

To better explain this concept, we provide an example by dis-aggregating the gNB user plane protocol stack. As is illustrated in Figure 7.2, BSF in gNB may consist of its MAC and PHY protocol stack. Similarly, AdpF can be constructed using RLC. The PDCP layer, which performs functions such as ciphering/deciphering, can be envisioned to constitute the SF. For some services, this layer also performs optimization over the data, e.g., it performs header compression for VoIP data packets. Hence, some of the PDCP together with SDAP may constitute the OptF. The set of layers such as L1, L2, IP, UDP, and GTP used to interface to 5G UPF constitute the IWF.

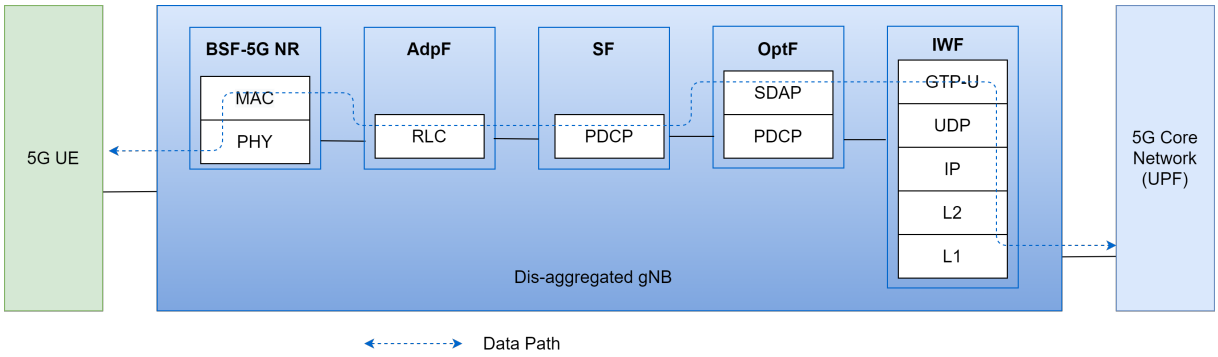


Figure 7.2: RAN dis-aggregation of the gNB user plane protocol stack

In a similar manner, we can dis-aggregate the control plane stack of the gNB as illustrated in Figure 7.3. The control functionality in the gNB, i.e., RRC together with RRM are transposed into the SDN controller. As seen from the figure, the IWF would be instantiated differently from the previous case and would comprise SCTP, IP, L2, and L1.

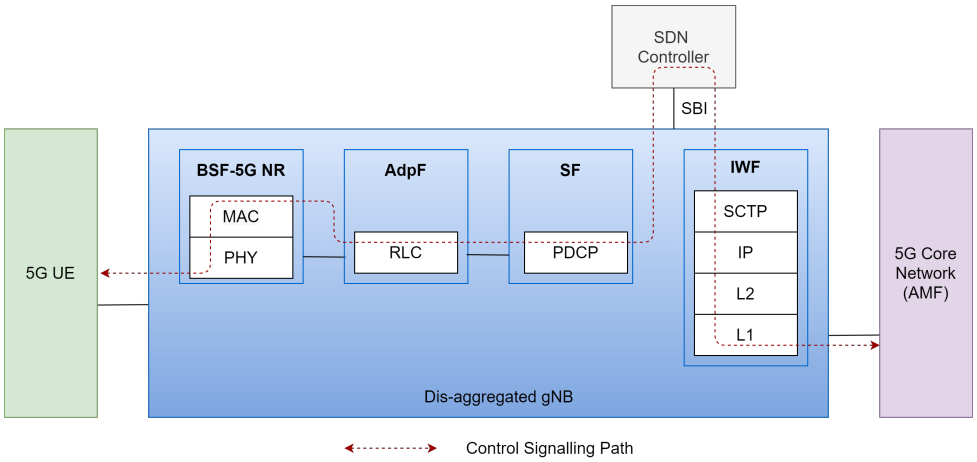


Figure 7.3: RAN dis-aggregation of the gNB control plane protocol stack

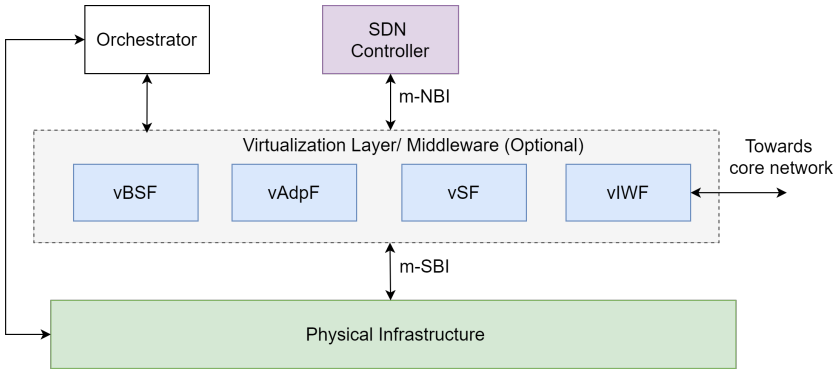


Figure 7.4: Architecture for dis-aggregating RAN nodes in existing networks

To implement dis-aggregation in existing RANs, a virtualization layer (also referred to as Middleware) is necessary. Network orchestrator, together with the virtualization layer, provides an abstract view of the network resources (in terms of the dis-aggregated NFs) to the controller. In continuation of the previous example, for the gNB, the virtualization layer may comprise E1AP, F1AP, NETCONF, and OpenFlow protocols. As evident, RAN dis-aggregation can help us architect flexible networks that can be optimized to a certain use-case.

While this initial design has been submitted to IEEE P1930.1 standardization activity [27] for consideration, a lot more efforts are necessary for achieving a complete design solution. For example, we require standardized data models for configuring the component NFs. By developing standardized data-models and defining protocols

to exchange their capabilities, we may be able to define a service-based architecture for the RAN.

In conclusion, it is evident that the SDN and NFV based approaches discussed in the thesis prove to be promising for the design and development of present-day and beyond multi-RAT networks. While we have explored some aspects of this topic, there are far more dimensions that hold significant potential for further study and research.

Bibliography

[1] "Ericsson Mobility Report," Ericsson, White Paper, [Online]. Available: https://www.ericsson.com/4adc87/assets/local/mobility-report/documents/2020/november-2020-ericsson-mobility-report.pdf, 2020, Last accessed: 27-02-2021.

[2] "Network Slicing for 5G Networks and Services," 5G Americas, White Paper, [Online]. Available: http://www.5gamericas.org/files/3214/7975/0104/5G_Americas_Network_Slicing_11.21_Final.pdf, 2016, Last accessed: 27-02-2021.

[3] "The 5G Economy," IHS Markit, White Paper, [Online]. Available: https://www.qualcomm.com/media/documents/files/ihs-5g-economic-impact-study-2019.pdf, 2019, Last accessed: 27-02-2021.

[4] "IMT Vision–Framework and overall objectives of the future development of IMT for 2020 and beyond," ITU-R, Technical Report, [Online]. Available: https://www.itu.int/dms_pubrec/itu-r/rec/m/R-REC-M.2083-0-201509-I!!PDF-E.pdf, 2015, Last accessed: 27-02-2021.

[5] 3GPP TS 38.104 V17.0.0, "NR; Base Station (BS) radio transmission and reception," 3GPP, Technical Specification, 2021, [Online]. Available: https://www.3gpp.org/ftp//Specs/archive/38_series/38.104/38104-h00.zip.

[6] 3GPP TS 38.201 V16.0.0, "NR; Physical layer; General description," 3GPP, Technical Specification, 2020, [Online]. Available: https://www.3gpp.org/ftp//Specs/archive/38_series/38.201/38201-g00.zip.

[7] A. Checko, H. L. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. S. Berger, and L. Dittmann, "Cloud RAN for Mobile Networks—A Technology Overview," *IEEE Communications surveys & tutorials*, vol. 17, no. 1, pp. 405–426, 2014.

[8] E. Pateromichelakis, F. Moggio, C. Mannweiler, P. Arnold, M. Shariat, M. Einhaus, Q. Wei, O. Bulakci, and A. De Domenico, "End-to-End Data Analytics Framework for 5G Architecture," *IEEE Access*, vol. 7, pp. 40 295–40 312, 2019.

[9] 3GPP TR 23.791 V16.2.0, "Study of enablers for Network Automation for 5G," 3GPP, Technical Report, 2019, [Online]. Available: https://www.3gpp.org/ftp//Specs/archive/23_series/23.791/23791-g20.zip.

[10] 3GPP TR 21.915 V15.0.0, "Release description; Release 15," 3GPP, Technical Report, 2019, [Online]. Available: https://www.3gpp.org/ftp//Specs/archive/21_series/21.915/21915-f00.zip.

[11] 3GPP TR 21.915 V1.0.0, "Release description; Release 16," 3GPP, Technical Report, 2020, [Online]. Available: https://ftp.3gpp.org//Specs/archive/21_series/21.916/21916-100.zip.

[12] E. Khorov, A. Kiryanov, A. Lyakhov, and G. Bianchi, "A Tutorial on IEEE 802.11ax High Efficiency WLANs," *IEEE Communications Surveys and Tutorials*, vol. 21, no. 1, pp. 197–216, 2018.

[13] Y. Ghasempour, C. R. da Silva, C. Cordeiro, and E. W. Knightly, "IEEE 802.11 ay: Next-generation 60 GHz communication for 100 Gb/s Wi-Fi," *IEEE Communications Magazine*, vol. 55, no. 12, pp. 186–192, 2017.

[14] IETF RFC 7426, "Software-Defined Networking (SDN): Layers and Architecture Terminology," IETF, Technical Specification, 2015, [Online]. Available: https://tools.ietf.org/pdf/rfc7426.pdf.

[15] D. Kreutz, F. Ramos, P. Verissimo, C. E. Rothenberg, S. Azodolmolky, and S. Uhlig, "Software Defined Networking: A Comprehensive Survey," *Proceedings of the IEEE*, vol. 103, no. 1, pp. 14–76, 2014.

[16] "OpenFlow Switch Specification V1.5.1," ONF, Technical Specification, [Online]. Available: https://opennetworking.org/wp-content/uploads/2014/10/openflow-switch-v1.5.1.pdf, 2015.

[17] N. McKeown, T. Anderson, H. Balakrishnan, G. Parulkar, L. Peterson, J. Rexford, S. Shenker, and J. Turner, "OpenFlow: Enabling Innovation in Campus Networks," *ACM SIGCOMM Computer Communication Review*, vol. 38, no. 2, pp. 69–74, 2008.

[18] ETSI GS NFV 003 V1.4.1, "Network Functions Virtualisation (NFV); Architectural Framework," ETSI, Technical Specification, 2018, [Online]. Available: https://www.etsi.org/deliver/etsi_gs/nfv/001_099/003/01.04.01_60/gs_nfv003v010401p.pdf.

[19] ETSI GS NFV 002 V1.2.1, "Network Functions Virtualisation (NFV); Architectural Framework," ETSI, Technical Specification, 2014, [Online]. Available: https://www.etsi.org/deliver/etsi_gs/NFV/001_099/002/01.02.01_60/gs_NFV002v010201p.pdf.

[20] ONF TR-526 V1.0, "Applying SDN Architecture to 5G Slicing," ONF, Technical Report, 2016, [Online]. Available: https://www.opennetworking.org/wp-content/uploads/2014/10/Applying_SDN_Architecture_to_5G_Slicing_TR-526.pdf.

[21] 3GPP TS 23.501 V16.7.0, "System architecture for the 5G System (5GS)," 3GPP, Technical Specification, 2020, [Online]. Available: https://www.3gpp.org/ftp//Specs/archive/23_series/23.501/23501-g70.zip.

[22] 3GPP TR 28.801 V15.1.0, "Telecommunication management;Study on management and orchestration of network slicing for next generation network," 3GPP, Technical Report, 2018, [Online]. Available: https://www.3gpp.org/ftp//Specs/archive/28_series/28.801/28801-f10.zip.

[23] K.-K. Yap, M. Kobayashi, R. Sherwood, T.-Y. Huang, M. Chan, N. Handigol, and N. McKeown, "OpenRoads: Empowering Research in Mobile Networks," *ACM SIGCOMM Computer Communication Review*, vol. 40, no. 1, pp. 125–126, 2010.

[24] "Global Mobile Trends: What is driving the mobile industry," GSMA Intelligence, White Paper, [Online]. Available: https://www.gsmaintelligence.com/research/?file=8535289e1005eb248a54069d82ceb824&download, 2018, Last accessed: 27-02-2021.

[25] "How do we plan for 5G NR network deployments coming in 2019?" Qualcomm, White Paper, [Online]. Available: https://www.qualcomm.com/media/documents/

files/best-practices-for-deploying-5g-nr-networks.pdf, 2018, Last accessed: 29-01-2021.

[26] V. G. Nguyen, T. X. Do, and Y. Kim, "SDN and Virtualization based LTE Mobile Network Architectures: A Comprehensive Survey," *Wireless Personal Communications*, vol. 86, no. 3, pp. 1401–1438, 2016.

[27] "P1930.1 - Recommended Practice for Software Defined Networking (SDN) based Middleware for Control and Management of Wireless Networks," IEEE, Website, [Online]. Available: https://standards.ieee.org/project/1930_1.html, 2016, Last accessed: 27-02-2021.

[28] "ns-3 Network Simulator," University of Washington, Simulation Software, [Online]. Available: https://www.nsnam.org, Last accessed: 27-02-2021.

[29] J. Schulz-Zander, L. Suresh, N. Sarrar, A. Feldmann, T. Hühn, and R. Merz, "Programmatic Orchestration of WiFi Networks," in *USENIX Annual Technical Conference*, 2014, pp. 347–358.

[30] E. Coronado, R. Riggio, J. Villa1ón, and A. Garrido, "Lasagna: Programming Abstractions for End-to-End Slicing in Software-Defined WLANs," in *IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM)*, 2018, pp. 14–15.

[31] E. Coronado, S. N. Khan, and R. Riggio, "5G-EmPOWER: A Software-Defined Networking Platform for 5G Radio Access Networks," *IEEE Transactions on Network and Service Management*, vol. 16, no. 2, pp. 715–728, 2019.

[32] X. Jin, L. E. Li, L. Vanbever, and J. Rexford, "Softcell: scalable and flexible cellular core network architecture," in *ACM conference on Emerging networking experiments and technologies (CoNEXT)*, 2013, pp. 163–174.

[33] K. Pentikousis, Y. Wang, and W. Hu, "Mobileflow: Toward software-defined mobile networks," *IEEE Communications magazine*, vol. 51, no. 7, pp. 44–53, 2013.

[34] 3GPP TS 23.214 V15.0.0, "Architecture enhancements for control and user plane separation of EPC nodes," 3GPP, Technical Specification, 2017, [Online]. Available: https://www.3gpp.org/ftp//Specs/archive/23_series/23.214/23214-f00.zip.

[35] A. Gudipati, D. Perry, L. E. Li, and S. Katti, "SoftRAN: Software Defined Radio Access Network," in *ACM SIGCOMM workshop on Hot topics in Software Defined Networking (HotSDN)*, 2013, pp. 25–30.

[36] A. Gudipati, L. E. Li, and S. Katti, "Radiovisor: A Slicing Plane for Radio Access Networks," in *ACM SIGCOMM workshop on Hot topics in software defined networking (HotSDN)*, 2014, pp. 237–238.

[37] X. Foukas, N. Nikaein, M. M. Kassem, M. K. Marina, and K. Kontovasilis, "FlexRAN: A Flexible and Programmable Platform for Software Defined Radio Access Networks," in *ACM International Conference on emerging Networking EXperiments and Technologies (CoNEXT)*, 2016, pp. 427–441.

[38] S. Costanzo, I. Fajjari, N. Aitsaadi, and R. Langar, "A network slicing prototype for a flexible cloud radio access network," in *IEEE Annual Consumer Communications & Networking Conference (CCNC)*, 2018, pp. 1–4.

[39] IETF RFC 1098, "A Simple Network Management Protocol (SNMP)," IETF, Technical Specification, 1989, [Online]. Available: https://tools.ietf.org/html/rfc1098.

[40] R. Sherwood, M. Chan, A. Covington, G. Gibb, M. Flajslik, N. Handigol, T.-Y. Huang, P. Kazemian, M. Kobayashi, J. Naous *et al.*, "Carving research slices out of your production networks with OpenFlow," *ACM SIGCOMM Computer Communication Review*, vol. 40, no. 1, pp. 129–130, 2010.

[41] M. A. Habibi, M. Nasimi, B. Han, and H. D. Schotten, "A Comprehensive Survey of RAN Architectures Toward 5G Mobile Communication System," *IEEE Access*, vol. 7, pp. 70 371–70 421, 2019.

[42] V.-G. Nguyen, A. Brunstrom, K.-J. Grinnemo, and J. Taheri, "SDN/NFV-Based Mobile Packet Core Network Architectures: A Survey," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 3, pp. 1567–1602, 2017.

[43] H. Zhang, N. Liu, X. Chu, K. Long, A.-H. Aghvami, and V. C. Leung, "Network Slicing Based 5G and Future Mobile Networks: Mobility, Resource Management, and Challenges," *IEEE communications magazine*, vol. 55, no. 8, pp. 138–145, 2017.

[44] I. Afolabi, T. Taleb, K. Samdanis, A. Ksentini, and H. Flinck, "Network Slicing and Softwarization: A Survey on Principles, Enabling Technologies, and Solutions," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 3, pp. 2429–2453, 2018.

[45] K. Kozlowski, S. Kuklinski, and L. Tomaszewski, "Open issues in network slicing," in *International Conference on the Network of the Future (NOF)*, 2018, pp. 25–30.

[46] "O-RAN: Towards an Open and Smart RAN," O-RAN Alliance, White Paper, [Online]. Available: https://www.o-ran.org/s/O-RAN-WP-FInal-181017.pdf, 2018.

[47] 3GPP TS 38.401 V0.3.0, "NG-RAN; Architecture description," 3GPP, Technical Specification, 2017, [Online]. Available: https://ftp.3gpp.org//Specs/archive/38_series/38.401/38401-030.zip.

[48] 3GPP TS 28.530 V17.0.0, "Management and orchestration; Concepts, use cases and requirements," 3GPP, Technical Specification, 2020, [Online]. Available: https://www.3gpp.org/ftp//Specs/archive/28_series/28.530/28530-h00.zip.

[49] 3GPP TS 28.531 V16.8.0, "Management and orchestration; Provisioning," 3GPP, Technical Specification, 2020, [Online]. Available: https://www.3gpp.org/ftp//Specs/archive/28_series/28.531/28531-g80.zip.

[50] 3GPP TS 28.532 V16.6.0, "Management and orchestration; Generic management services," 3GPP, Technical Specification, 2020, [Online]. Available: https://www.3gpp.org/ftp//Specs/archive/28_series/28.532/28532-g60.zip.

[51] I. F. Akyildiz, P. Wang, and S.-C. Lin, "SoftAir: A Software Defined Networking Architecture for 5G Wireless Systems," *Elsevier Computer Networks*, vol. 85, pp. 1–18, 2015.

[52] A. Nayak M., P. Jha, and A. Karandikar, "A Centralized SDN Architecture for the 5G Cellular Network," in *IEEE 5G World Forum (5GWF)*, 2018, pp. 147–152.

[53] 3GPP TS 38.300 V16.4.0, "NR; NR and NG-RAN Overall description; Stage-2," 3GPP, Technical Specification, 2021, [Online]. Available: https://www.3gpp.org/ftp//Specs/archive/38_series/38.300/38300-g40.zip.

[54] 3GPP TS 29.500 V17.1.0, "5G System; Technical Realization of Service Based Architecture; Stage 3," 3GPP, Technical Specification, 2020, [Online]. Available: https://www.3gpp.org/ftp//Specs/archive/29_series/29.500/29500-h10.zip.

[55] 3GPP TR 38.801 V14.0.0, "Study on new radio access technology: Radio access architecture and interfaces," 3GPP, Technical Report, 2017, [Online]. Available: https://www.3gpp.org/ftp//Specs/archive/38_series/38.801/38801-e00.zip.

[56] F. H. Khan and M. Portmann, "A System Level Architecture for Software Defined LTE Networks," in *IEEE Signal Processing and Communication Systems (ICSPCS)*, 2016, pp. 1–10.

[57] J. Zhang, W. Xie, and F. Yang, "An Architecture for 5G Mobile Network based on SDN and NFV," in *IET International Conference on Wireless, Mobile and Multi-Media (ICWMMN)*, 2015.

[58] 3GPP TS 23.502 V15.0.0, "Procedures for the 5G System (5GS)," 3GPP, Technical Report, 2017, [Online]. Available: https://www.3gpp.org/ftp//Specs/archive/23_series/23.502/23502-f00.zip.

[59] I. Alawe, Y. Hadjadj-Aoul, A. Ksentini, P. Bertin, and D. Darche, "On the scalability of 5G Core network: the AMF case," in *IEEE Annual Consumer Communications and Networking Conference (CCNC)*, 2018, pp. 1–6.

[60] A. Basta, W. Kellerer, M. Hoffmann, K. Hoffmann, and E.-D. Schmidt, "A Virtual SDN-enabled LTE EPC Architecture: A case study for S-/P-gateways Functions," in *IEEE SDN for Future Networks and Services (SDN4FNS)*, 2013, pp. 1–7.

[61] L. Peterson, A. Al-Shabibi, T. Anshutz, S. Baker, A. Bavier, S. Das, J. Hart, G. Palukar, and W. Snow, "Central office re-architected as a data center," *IEEE Communications Magazine*, vol. 54, no. 10, pp. 96–101, 2016.

[62] 3GPP TR 25.912 V14.0.0, "Feasibility Study for Evolved Universal Terrestrial Radio Access (UTRA) and Universal Terrestrial Radio Access Network (UTRAN)," 3GPP, Technical Report, 2017, [Online]. Available: https://www.3gpp.org/ftp//Specs/archive/25_series/25.912/25912-e00.zip.

[63] 3GPP TR 22.891 V14.2.0, "Study on new services and markets technology enablers," 3GPP, Technical Report, 2016. [Online]. Available: https://www.3gpp.org/ftp/Specs/archive/22_series/22.891/22891-e20.zip

[64] A. Roy, P. Chaporkar, and A. Karandikar, "Optimal Radio Access Technology Selection Algorithm for LTE-WiFi Network," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 7, pp. 6446–6460, 2018.

[65] A. Nayak M., A. Roy, P. Jha, and A. Karandikar, "Control and Management of Multiple RATs in Wireless Networks: An SDN Approach," in *IEEE 2nd 5G World Forum (5GWF)*, 2019, pp. 596–601.

[66] A. Karandikar, P. K. Jha, A. Nayak M., N. Shah, A. Roy, O. A. Kanhere, P. Priyanka, A. Dandekar, and R. Kharade, "Methods and Systems for Controlling an SDN-based Multi-RAT Communication Network," 2019, US Patent Grant No. 10,187,928.

[67] K. S. Munasinghe, I. Elgendi, A. Jamalipour, and D. Sharma, "Traffic Offloading 3-tiered SDN Architecture for DenseNets," *IEEE Network*, vol. 31, no. 3, pp. 56–62, 2017.

[68] X. Mi, Z. Tian, X. Xu, M. Zhao, and J. Wang, "NO Stack: A SDN-based Framework for Future Cellular Networks," in *IEEE Wireless Personal Multimedia Communications (WPMC)*, 2014, pp. 497–502.

[69] 3GPP TS 29.244 V15.2.0, "Interface between the Control Plane and the User Plane nodes," 3GPP, Technical Specification, 2018, [Online]. Available: https://www.3gpp.org/ftp//Specs/archive/29_series/29.244/29244-f20.zip.

[70] 3GPP TS 38.463 V15.3.0, "NG-RAN; E1 Application Protocol (E1AP)," 3GPP, Technical Specification, 2019. [Online]. Available: https://www.3gpp.org/ftp//Specs/archive/38_series/38.463/38463-f30.zip

[71] 3GPP TS 38.463 V15.5.0, "NG-RAN; F1 Application Protocol (F1AP)," 3GPP, Technical Specification, 2019. [Online]. Available: https://www.3gpp.org/ftp//Specs/archive/38_series/38.473/38473-f50.zip

[72] 3GPP TS 32.425 V17.0.0, "Performance Management (PM); Performance measurements Evolved Universal Terrestrial Radio Access Network (E-UTRAN)," 3GPP, Technical Specification, 2020. [Online]. Available: https://www.3gpp.org/ftp//Specs/archive/32_series/32.425/32425-h00.zip

[73] 3GPP TS 28.552 V17.2.1, " Management and orchestration; 5G performance measurements," 3GPP, Technical Specification, 2021. [Online]. Available: https://www.3gpp.org/ftp/Specs/archive/28_series/28.552/28552-h21.zip

[74] G. Bianchi, "Performance Analysis of the IEEE 802.11 Distributed Coordination Function," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 3, pp. 535–547, 2000.

[75] 3GPP TS 36.839 V11.0.0, "Evolved Universal Terrestrial Radio Access (E-UTRA); Mobility enhancements in heterogeneous networks," 3GPP, Technical Specification, 2012. [Online]. Available: https://www.3gpp.org/ftp/Specs/archive/36_series/36.839/36839-b00.zip

[76] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 2014.

[77] IEEE 802.11-2016, "IEEE Standard for Information technology Telecommunications and information exchange between systems-Local and metropolitan area networks-Specific requirements Part 11: Wireless LAN Medium-Access Control (MAC) and Physical Layer (PHY) Specifications," IEEE, Technical Specification, 2016, [Online]. Available: https://standards.ieee.org/standard/802_11-2016.html.

[78] 3GPP TS 28.550 V16.7.0, "Management and orchestration; Performance assurance," 3GPP, Technical Specification, 2020, [Online]. Available: https://www.3gpp.org/ftp//Specs/archive/28_series/28.550/28550-g70.zip.

[79] A. Nayak M., P. Jha, A. Karandikar, and P. Chaporkar, "VirtRAN: An SDN/NFV-Based Framework for 5G RAN Slicing," *Springer Journal of the Indian Institute of Science (Invited Paper)*, vol. 100, no. 2, pp. 409–434, 2020.

[80] S. Kumar, D. Cifuentes, S. Gollakota, and D. Katabi, "Bringing Cross-layer MIMO to Today's Wireless LANs," *ACM SIGCOMM Computer Communication Review*, vol. 43, no. 4, pp. 387–398, 2013.

[81] ONF TR-502 V1.0, "SDN Architecture," ONF, Technical Report, 2014, [Online]. Available: https://www.opennetworking.org/wp-content/uploads/2013/02/TR_SDN_ARCH_1.0_06062014.pdf.

[82] A. Blenk, A. Basta, M. Reisslein, and W. Kellerer, "Survey on Network Virtualization Hypervisors for Software Defined Networking," *IEEE Communications Surveys and Tutorials*, vol. 18, no. 1, pp. 655–685, 2015.

[83] A. Nayak M., P. Jha, A. Karandikar, and P. Chaporkar, "Enhanced UE Slice Mobility for 5G Multi-RAT Networks," in *IEEE Conference on Network Function Virtualization and Software Defined Networks (NFV-SDN) MOBISLICE Workshop*, 2019, pp. 1–6.

[84] A. Nayak M, P. Jha, A. Karandikar, and P. Chaporkar, "Enhanced UE Slice Availability and Mobility through Multi-connectivity in 5G Multi-RAT Networks," *Wiley Internet Technology Letters (Invited Paper)*, vol. 3, no. 6, p. e184, 2020.

[85] A. Karandikar, P. Chaporkar, P. Jha, and A. Nayak M., "Methods and systems for Radio Access Network aggregation and uniform control of multi-RAT networks," 2021, uS Patent Application No. 16/842,095.

[86] ETSI GR NGP 001 V1.1.1, "Next Generation Protocols (NGP); E2E Network Slicing Reference Framework and Information Model," ETSI, Technical Specification, 2018, [Online]. Available: https://www.etsi.org/deliver/etsi_gr/NGP/001_099/011/01.01.01_60/gr_NGP011v010101p.pdf.

[87] "Unleashing the economic potential of network slicing," Nokia, White Paper, [Online]. Available: https://onestore.nokia.com/asset/202089, 2018, Last accessed: 27-02-2021.

[88] 3GPP TR 37.340 V15.5.0, "NR; Multi-connectivity; Overall description; Stage-2," 3GPP, Technical Specification, 2019. [Online]. Available: https://www.3gpp.org/ftp//Specs/archive/37_series/37.340/37340-f50.zip

[89] M. Polese, M. Giordani, M. Mezzavilla, S. Rangan, and M. Zorzi, "Improved Handover Through Dual Connectivity in 5G mmWave Mobile Networks," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 9, pp. 2069–2084, 2017.

[90] R. Wen, G. Feng, J. Zhou, and S. Qin, "Mobility Management for Network Slicing Based 5G Networks," in *IEEE International Conference on Communication Technology (ICCT)*, 2018, pp. 291–296.

[91] F. Meneses, R. Silva, D. Santos, D. Corujo, and R. L. Aguiar, "Using SDN and Slicing for Data Offloading over Heterogeneous Networks Supporting non-3GPP Access," in *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, 2018, pp. 1–6.

[92] H. Zhang and W. Huang, "Tractable Mobility Model for Multi-Connectivity in 5G User-Centric Ultra-Dense Networks," *IEEE Access*, vol. 6, pp. 43 100–43 112, 2018.

[93] N. H. Mahmood, M. Lopez, D. Laselva, K. Pedersen, and G. Berardinelli, "Reliability oriented Dual connectivity for URLLC services in 5G New Radio," in *IEEE International Symposium on Wireless Communication Systems*, 2018.

[94] P.-J. Hsieh, W.-S. Lin, K.-H. Lin, and H.-Y. Wei, "Dual-connectivity Prevenient Handover Scheme in Control/user-plane Split Networks," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 4, pp. 3545–3560, 2017.

[95] A. Alfoudi, M. Dighriri, A. Otebolaku, R. Pereira, and G. Lee, "Mobility Management Architecture in Different RATs Based Network Slicing," in *IEEE International Conference on Advanced Information Networking and Applications Workshops (WAINA)*, 2018, pp. 270–274.

[96] A. H. Celdrán, M. G. Pérez, F. J. G. Clemente, F. Ippoliti, and G. M. Pérez, "Policy-based Network Slicing Management for Future Mobile Communications," in *IEEE International Conference on Software Defined Systems (SDS)*, 2018, pp. 153–159.

[97] F. Z. Yousaf, M. Gramaglia, V. Friderikos, B. Gajic, D. von Hugo, B. Sayadi, V. Sciancalepore, and M. R. Crippa, "Network Slicing with Flexible Mobility and QoS/QoE support for 5G Networks," in *IEEE International Conference on Communications Workshops (ICC Workshops)*, 2017, pp. 1195–1201.

[98] 3GPP TS 36.360 V15.0.0, "Evolved Universal Terrestrial Radio Access (E-UTRA); LTE-WLAN Aggregation Adaptation Protocol (LWAAP) specification," 3GPP, Technical Specification, 2018. [Online]. Available: https://www.3gpp.org/ftp//Specs/archive/36_series/36.360/36360-f00.zip

[99] A. Karandikar, P. Jha, A. Nayak M., and P. N. Elango, "Sdn controlled overlay network," 2019, uS Patent Grant No. 10,433,234.

List of Publications

Part of Ph.D Thesis

Journal Publications

J1 **A. Nayak M.**, P. Jha, A. Karandikar and P. Chaporkar, "VirtRAN: An SDN/NFV-Based Framework for 5G RAN Slicing," ***(Invited Paper)*** *Springer Journal of the Indian Institute of Science, Special issue on 5G and Beyond*, vol. 100, no. 2, pp. 409–434, 2020.

J2 **A. Nayak M.**, P. Jha, A. Karandikar and P. Chaporkar, "Enhanced UE Slice Availability and Mobility through Multi-Connectivity in 5G Multi-RAT Networks," ***(Invited Paper)*** *Wiley Internet Technology Letters, Special Issue on MOBISLICE-SI-2019*, vol. 2, no. 6, pp. e184, 2020.

International Conference Publications

C1 **A. Nayak M.**, P. Jha, and A. Karandikar, "A Centralized SDN Architecture for the 5G Cellular Network," *in IEEE 5G World Forum (5GWF)*, 2018, pp. 147-152.

C2 **A. Nayak M.**, A. Roy, P. Jha, and A. Karandikar, "Control and Management of Multiple RATs in Wireless Networks: An SDN Approach," *in IEEE 5G World Forum (5GWF)*, 2019, pp. 596-601.

C3 **A. Nayak M.**, P. Jha, A. Karandikar and P. Chaporkar, "Enhanced UE Slice Mobility for 5G Multi-RAT Networks ," *in IEEE Conference on Network Function Virtualization and Software Defined Networks (NFV-SDN) MOBISLICE Workshop*, 2019, pp. 1-6.

Patents Granted/Filed

P1 A. Karandikar, P. Jha, **A. Nayak M.**, N. Shah, A., O. Kanhere, P. Pola, A. Dandekar and R. Kharade, "Methods and Systems for Controlling a SDN based MultiRAT Communication Network," US Patent Grant no. 10,187,928, 2019.

P2 P. Jha, P. Chaporkar, A. Karandikar, **A. Nayak M.**, "Methods and Systems for Radio Access Network Aggregation and Uniform Control of Multi-RAT Networks", US Patent Application no. 16/842,09, 2020.

Other Publications

International Conference Publications

C4 M. Khaturia, **A. Nayak M.**, P. Jha and A. Karandikar, "5G-Serv: Decoupling User Control and Network Control in the 3GPP 5G Network", *Accepted in IEEE Conference on Innovation in Clouds, Internet and Networks (ICIN)*, 2021.

Patents Granted

P3 A. Karandikar, P. Jha, **A. Nayak M.** and P. Nidhya E., "SDN controlled Overlay Network," US Patent Grant no. 10,433,234, 2019.

P4 A. Karandikar, P. Jha, K. Kumar, P. Nidhya E., P. Magar, S. A. Nadaf, M. Tripathi and **A. Nayak M.**,"Methods and Systems for Providing Standalone LTE based Communication Networks," US Patent Grant no. 10,182,354, 2019.

P5 A. Karandikar, P. Jha, **A. Nayak M.**, P. Nidhya E. and P. Magar, "Highly Available Network Architecture for a LTE based Communication Network", US Patent Grant no. 10,506,508, 2019.