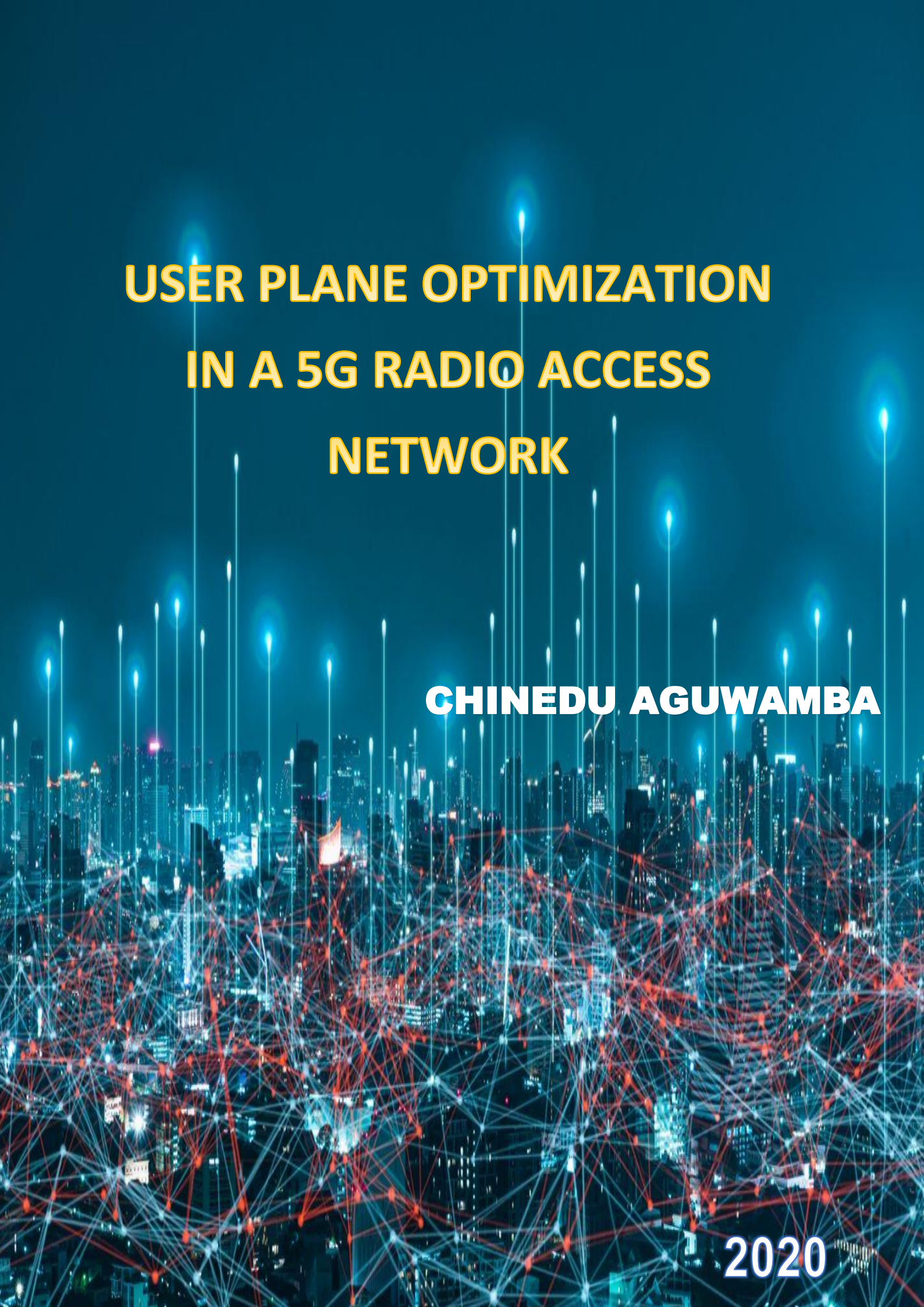
USER PLANE OPTIMIZATION IN A 5G RADIO ACCESS NETWORK

CHINEDU AGUWAMBA



2020

USER PLANE OPTIMIZATION IN A 5G RADIO ACCESS NETWORK

BY

CHINEDU AGUWAMBA

In partial fulfilment of the requirements for the degree of

Master of Science

In Embedded Systems
Specialization: Software and Networking

at the Delft University of Technology
to be publicly defended on Friday August 28, 2020 at 9:30 AM.

Thesis Committee:

1.	Dr. Remco Litjens, MSc	TU Delft, TNO
2.	Dr. Ir. Stephan Wong	TU Delft
3.	Ir. Rogier Noldus	TU Delft, Ericsson
4.	Maria Raftopoulou, MSc	TU Delft

TU Delft
Delft University of Technology

NAS

Preface

This thesis marks the finalization of my studies at Delft University of Technology, The Netherlands, for the Master of Science in Embedded Systems in the specialization track of Software and Networking. My experience while working on this thesis research has been challenging and academically intriguing. I would like to use this opportunity to thank everyone that contributed to the successful completion of the research for their immense support, mentorship and guidance.

Specifically, I would like to express my sincere gratitude to Maria Raftopoulou, my daily supervisor for her tremendous devotion of time out of her busy schedule in providing me with feedback on the progress of the research work and her guidance in writing the report. Though her feedback has been critical and challenging, working on the feedback has drastically improved the quality of the research work. Also, I would like to thank Rogier Noldus, my second daily supervisor, who inspired me on the research work, for his extensive guidance and despite his busy schedule accepted to become my daily supervisor. I also acknowledge all and sundry who helped me in one way or the other towards the course of the research work.

Finally, I would like to thank all my friends and family and most especially Paul Douglas and Ngozi Aguwamba who supported me financially and mentally during my study in The Netherlands.

Chinedu Aguwamba
Delft, August 2020

Abstract

5G as a future network is expected to be commercially deployed in 2020 and beyond. At the present time facilitated by industry need, the deployment option is to introduce 5G base stations alongside the existing 4G base stations in order to expedite 5G deployment. This deployment option is what presently is referred to as the Non-Standalone Architecture (NSA). To fully unlock the 5G potential such as enhanced end-user experience, service agility, Ultra Reliable Low Latency Communications (URLLC), improved network capabilities, critical Internet of Things (IoT) and industrial automation use cases, it becomes imperative to deploy a full 5G architecture with its own New Radio (NR) access and 5G Core Network (5GCN).

The goal of the thesis is to design a 5G standalone architecture that leverages on the principle of Control and User Plane Separation (CUPS) to be introduced in the 5G Radio Access Network (RAN). Such separation enables scaling of each plane's resources and also allows for a flexible deployment of the architecture as chosen by the Mobile Network Operator (MNO). To this effect the New Radio-New Radio (NR-NR) architecture is introduced which makes use of two 5G base stations such that a user can connect simultaneously to the two base stations in what is called Dual Connectivity (DC). One base station, which is referred as the Next Generation NodeB (gNB-CP), specifically handles all the Control Plane (CP) signalling in the RAN and the second base station, called gNB-UP, is dedicated specifically to handle User Plane (UP) traffic. To investigate how the new architecture handles control signalling and optimizes the UP as a result of decoupling the UP functions from CP signalling, IP Multimedia Subsystem (IMS)-based voice telephony, that is voice call made over a 5G network specifically called Voice over New Radio (VoNR), is chosen as an application and two distinct use cases are considered. The first use case is to investigate through signalling messages how the proposed architecture handles control signalling for setting up a VoNR call. The second use case is to investigate through signalling messages and data flow path how user mobility and handover procedures are handled during an ongoing VoNR call. Finally, a comparative study was conducted with the NSA.

From the results obtained and from the comparative study conducted, it is shown that the NR-NR architecture decouples the UP functions from CP signalling. For handover procedures in the NR-NR architecture involving a VoNR call, the gNB-CP initiates and handles all control signalling while maintaining the VoNR call, which allows for the direct forwarding of a voice call from the serving gNB-UP to the target gNB-UP. This handover procedure eliminates any interruption of the ongoing voice call. Finally, we foresee there is a possibility of increased signalling load in the NR-NR architecture proposed because proper co-ordination is needed between a gNB-CP and a gNB-UP to ensure optimal network functionality when compared to the NR architecture which uses a single 5G base station.

Table of Contents

List of Figures

List of Tables

CHAPTER 1

INTRODUCTION

This research study proposes an enhanced architecture that decouples the control plane and the data plane/user plane of Fifth Generation (5G) mobile networks. By decoupling the planes, allows the user plane functions which carry the user data or traffic to be optimized while enabling also both the control plane and the user plane resources to be scaled independently as desired by network operators. This chapter is made up of three main sections; the first section introduces the road to 5G and its global adoption as the future network, the second section introduces the background knowledge of enabling technologies that evolved 4G to 5G. Then, the final section discusses the general research motivation, the thesis contribution, and the research objectives.

1.1 The Road To 5G

From 2020 and further, the mobile broadband data consumption is fast accelerating with almost 75% of these data traffic video centric [1]. In the same vain, the Internet of Things (IoT) will have tremendous impact on the future mobile network with approximately18 billion out of 29 billion connected devices relating to IoT data traffic [2]. Thus, it can be shown that the four major factors that will expedite the adoption of 5G in the next coming years are data creation/consumption, more connected devices, the demand of high speed for real-time access applications and applications with low latency. Apart from building out the 5G network infrastructure, Mobile Network Operators (MNO) shall tremendously ensure return on their investment by embracing unique applications, and by adopting 5G best practices. Also, partnerships with other industrial experts are highly crucial.

Each generation of mobile technology unlocks new opportunities for telecommunication players. In 2020, worldwide 5G wireless network infrastructure revenues will amount to about 4.2 billion dollars, an 89% growth from 2019 revenue of 2.2 billion dollars according to [3]. 5G deployments introduced in 2019 and 2020 use the Non-Standalone Architecture (NSA) deployment technology and thus the already completed New Radio (NR) equipment can be rolled out alongside the existing 4G core network infrastructure. Therefore, it allows for network operators to introduce 5G services that will be deployed speedily. Figure 1.1 shows the global adoption of 5G for the next five years. It indicates that by 2025, the majority of the adoption will be linked to the Asia-Pacific region while the least adoption is associated to the region of Latin America. The 5G network launch has already begun in South Korea, US and other countries. Some European countries like Finland, Sweden

and Germany, which have anticipated plans to launch 5G network infrastructure in 2020 or later.

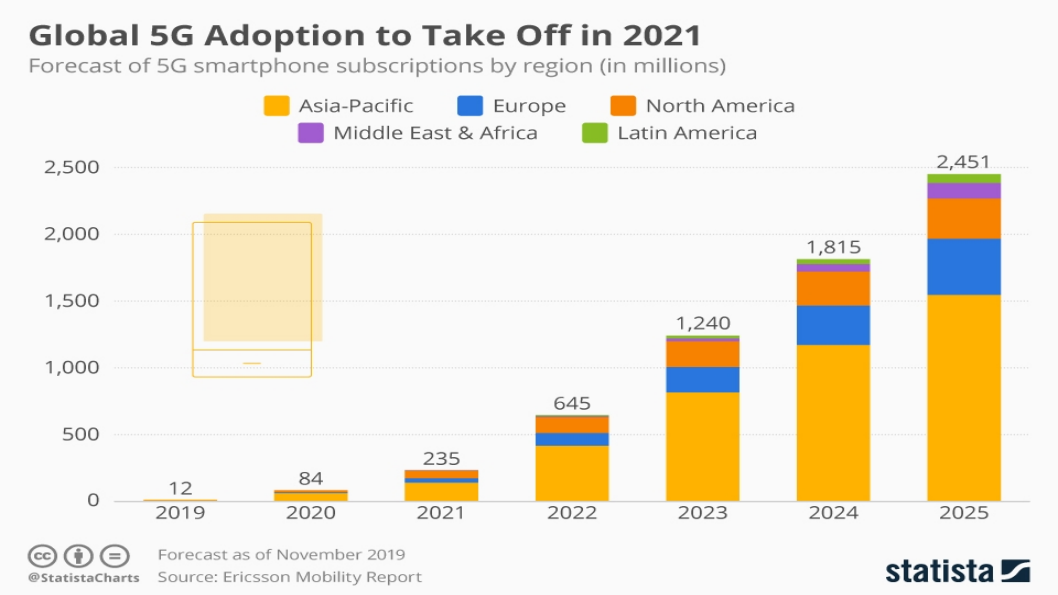


Figure 1. 1: Global 5G Adoption [4].

A graph communicating the global adoption of 5G. It shows that by 2020 84 million devices will be connected.

According to [4], the standalone deployment option of 5G will not really kick off until 2020/2021 and it is estimated to map the global 5G smartphone subscription of 12 million by 2019 and 84 million by the end of 2020. One of the main targets for the need of 5G deployment including industrial automation and mission-critical applications remains global mobile broadband. The global demand for "high-quality video content" continues to surge, resulting in increase in data usage and the need for a greater network capacity. Having said that, 5G remains hugely important for the overall development of IoT. However, until at least the mid-2020s, IoT developments will remain dependable on earlier generations of network technologies. Ericsson predicts that by the end of 2024 there will be about 1.9 billion 5G subscriptions, 35% of the mobile data traffic will be carried by 5G networks and up to 65% of the world population will be covered by the emerging technology [4], in other words, have access to 5G technology. This makes it the fastest generation to be rolled out on a large global rate.

1.2 5G as a Future Network

In particular, 5G will support numerous services such as enhanced Mobile Broadband (eMBB) with high data rates, massive Machine Type Communication (mMTC) for the enormous number of IoT devices available and Ultra Reliable Low Latency Communication (URLLC) for mission-critical services like remote surgery and self-driving cars. This is illustrated vividly in the 5G triangle as shown in Figure 1.2. The emergence of new applications and services including ultra-high definition video streaming, remote surgery, Augmented Reality (AR), Virtual

Reality (VR), self-driving cars, smart homes and IoT-related services like smart cities and smart agriculture have introduced new challenges for the existing mobile network infrastructure and capacity. These applications and services may need one or more requirements like high data rates, low latency, high reliability and high availability. To meet these service requirements, the design principles of the existing Long-Term Evolution (LTE) have to be recalibrated in order to provide adequate resources and lower latency for these new applications/services. The 5G standalone network is the solution to fulfil these diverse service requirements and thus enable the new applications and services.

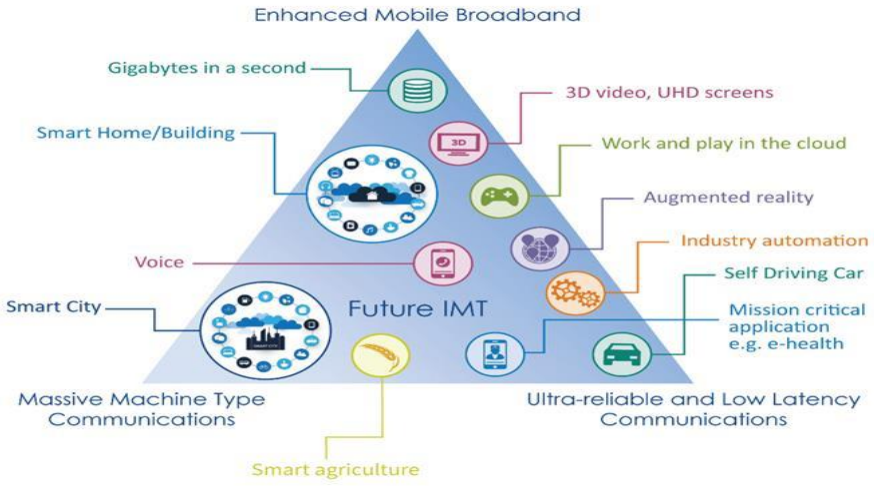


Figure 1. 2: 5G Triangle [5].

A chart showing different applications and services supported in the 5G Network. They include mostly URLLC, MTC, eMBB and others.

In order to support some of the 5G services already mentioned and meet the application requirement, drastic changes have been made in the architecture of the 5G network [6]. Network Function Virtualization (NFV), Mobile Edge Computing (MEC) and Software Defined Network (SDN) have all been introduced either in the 5G Core Network (CN) or the 5G Radio Access Network to enable the above-mentioned eMBB, URLLC and mMTC applications on the same physical infrastructure. These technologies form the fundamental pillars that enabled the paradigm shift to programmable networks which allows for flexibility in managing and controlling network traffic [7].

Researchers have proposed various new solutions to reduce the latency in 5G networks [6]. These solutions can be deployed in the RAN, the CN and as caching solutions. Proposed solutions in the RAN include flexible numerologies, and cloud RAN among others. Solutions in the CN include SDN, Network Functions (NF)s and MEC while for the caching solution can be in the form of content delivery, distributed caching or centralized caching. The 3rd Generation Partnership Project (3GPP) has finished the standardisation of the 5G RAN in

Release 15 and the first commercial deployment of the 5G CN network would be in 2020/2021. A widespread launch will take place in 2022 or later [8].

The future 5GCN is designed as a service-based architecture. In a service-based architecture, there are many Network Functions (NFs) (for example Session Management Function (SMF)) which allows any other NF to access its services. This service access can be done by having the NFs disclosing their service capabilities to other NFs in the network via a flexible and extendible Application Programable Interface (API). As a result, a flexible and scalable deployment of User Plane Functions (UPFs) is enabled. These concepts will be discussed further in detail in subsequent sections.

1.3 The 5G Network Enablers

We are going to briefly discuss some technologies that enabled the evolution to 5G networks. It includes Network Slicing, Service-based architecture and Software Defined Networking.

1.3.1 Network Slicing

According to [9], the sophistication and advancement of 5G revolves around *network slicing*. The concept of network slicing involves splitting of network resources to logical or virtual networks called "slices" where each slice addresses a specific application with distinct characteristics and service requirements. The idea is that each slice supports a given service application with various degrees of resource allocation [10]. For example, a slice that supports mission-critical services like remote surgery and self-driving cars, will differ in terms of throughput, latency and reliability requirements from a slice handling typical voice or video calls. The network has to ensure that each slice provides its users with a specified service quality and end-to-end network performance. Thus, an appropriate amount of resources is need to be allocated to each slice. Furthermore, the network can also dynamically turn down requests for new slices. Despite the potential gains from network slicing, it has its own challenges such as to dynamically co-ordinate the 5G slices so that they are optimally created and maintained based on resource availability and time-varying use case requirements [10]. Network slicing provides both optimal and domain-specific end-to-end performance within the underlying infrastructure. Hence, de-prioritization of less critical applications and services looks like a feasible solution [9].

1.3.2 Service-Based Architecture

Network Function Virtualization (NFV) and Software Defined Networks (SDN) are emerging technologies for managing and deploying network services [11]. They form the basis of the service-based architecture of the 5G network. As already mentioned, the CN is a service-based architecture. As per 3GPP standardization, all NFs within the control plane shall use the service-based approach to access network functionalities and maintain interactions. One benefit of *service-based*

architecture is that it grants service applications to be flexible, scalable and customizable for future network. Secondly, it facilitates the functional decoupling of the RAN and CN of the 5G architecture which greatly minimizes network dependencies between the RAN and the CN. This benefit will mostly allow the network to meet different performance targets for various services and applications. Thirdly, service-based architecture allows the combination between 3GPP and non-3GPP access network, such as Wi-Fi, the functionalities to interact compatibly without service deficiencies [11].

Another important benefit of the service-based architecture is that customized costly hardware platforms are now implemented as software applications running on low-cost hardware [12]. Virtualized Network Functions (VNF) are software that can run in an embedded network entity or in a network server hardware which can be moved in various locations as required by the network operator without the need of re-installing new network equipment. More concretely, it decouples the hardware from software. The concept of VNFs and SDN are related but independent, and they are both utilized to bring about a gain in network performance [7]. NFV delivers support to SDN by providing the enabling physical infrastructure upon which SDN software can run or operate for optimal network performance. The implication of this changes the architectural landscape of the 5G network in shifting from a vendor-specific model to an SDN-generic model that allows for control plane and user plane separation:

- **Control Plane (CP):** This is the part of the network architecture which carries control signals between the network and the user/device also internal in the network. It carries information such as Radio Resources Control (RRC) messages, Non-Access Stratum signalling, session control messages and control signals for routing/forwarding user data. Without a reliable and efficient control signal connection established, no radio connection or data transmission can be supported.
- **User Plane (UP**): This is part of the network architecture which carries or forwards actual user data. It is also referred as the data plane. It carries high speed data or voice/video traffic to/from an external data network via the UPF to the user terminal.

1.3.3 Software Defined Networking

SDN is one technique that allows the decoupling of CP and UP in the CN. The SDN consists of an SDN controller for decision-making and it generates sets of forwarding rules applied in the UP which directs user traffic as configured by the SDN controller. In other words, the CP in SDN is involved in decision-making that determines where and how the traffic should be forwarded, while the UP refers to the system that forwards the packets according to the decisions made by the controller running on the CP as shown in Figure 1.3. Furthermore, SDN simplifies network management procedures [1]. This is achieved by moving the CP to a software application residing in the SDN controller which results to a programmable network entity. In a nut-shell, intelligence is removed from most

network entities or nodes and to be centralized (intelligence) in fewer nodes. The advantage of this includes cost saving and a higher resource efficiency.

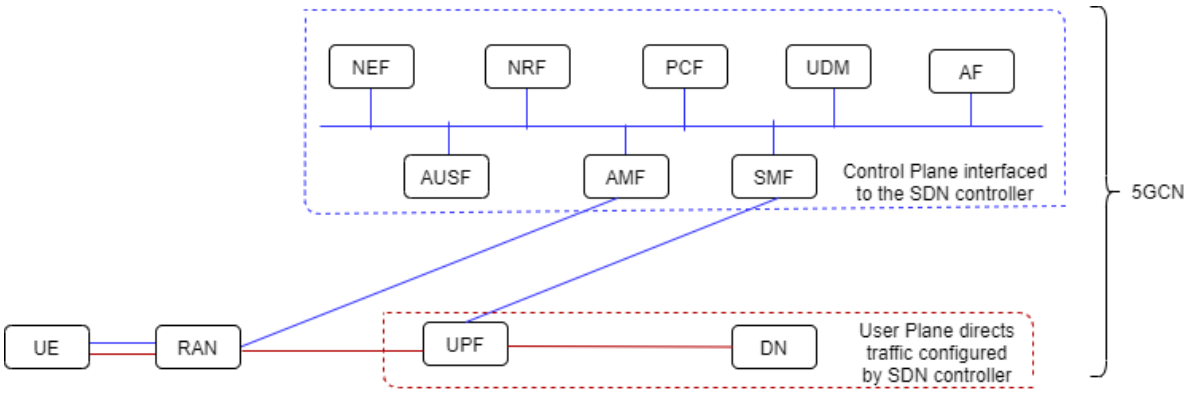


Figure 1. 3: SDN separation of CP and UP in 5GCN.

The SDN controller co-ordinates the activities of the CP and the UP for optimal network functionality.

In literature an integration has been proposed of SDN in the RAN with one controller controlling both RAN NFs and CN NFs to what is referred as Soft-RAN [13]. This entails that the base stations function as Virtual Machines (VM) running on network elements. For the SDN, the protocol used between the CP and UP is the OpenFlow protocol which is specified by 3GPP [5]. One known disadvantage of SDN is the initial delay when starting up the network, [14]. Though the initial delay as a disadvantage is not important since a mobile network is started up once and then runs for several years. All network functions shown in the figure is discussed extensively in Chapter 3.

1.4 Dual Connectivity

Dual Connectivity (DC) refers to when a mobile user or User Equipment (UE) is connected simultaneously to two distinct cells usually referred as master node and secondary node. The two nodes can be of various RATs as in NSA, or they can be from the same RAT such as the NR-NR architecture introduced in this thesis. DC improves user throughput, mobility robustness and also supports load balancing among various secondary nodes. DC allows relatively lower backhaul latency between two serving cells and usually requires packet reordering in the higher layer for bearer split. Besides to further improve user throughput, Multi-Connectivity (MC) technology is used to allow the user receive data from multiple cells in a heterogenous network [1].

1.4.1 Non-Standalone Deployment

The Non-Standalone Architecture (NSA) deployment option is a type of deployment architecture that is composed of two Radio Access Technologies (RAT) namely 4G Long Term Evolution (LTE) and 5G NR. The two base stations are connected to the Evolved Packet Core (EPC) which is the CN of the 4G network. The UE is configured to support both radio technologies and it is

connected to the two base stations in DC mode. The architecture ensures that mobile broadband that require 5G data rates are transmitted via a 5G base station. There are multiple options for NSA deployment [15] namely Option 3, Option 3a, Option 3x and so on. Option 3x is a combination of Option 3 and Option 3a. It is important to mention that Option 3a, also called the E-UTRAN New Radio Dual Connectivity (EN-DC), which utilizes the 4G EPC, a 4G base station and a 5G base station, is the most popular deployment option for mobile operators that want to quickly deploy 5G networks. Whenever NSA is mentioned in this thesis, we refer to Option 3a EN-DC.

The NSA deployment option utilizes the existing 4G infrastructure, but this deployment option comes with a disadvantage that it does not support some 5G applications such as URLLC, MTC, network slicing and so on. The benefit of the NSA is that it allows operators to launch 5G quickly for eMBB to gain throughput for mobile users while grabbing market leadership for 5G network. Figure 1.4 shows the NSA. The details of the architecture will be discussed in Chapter 4. One demerit of the NSA is that it is difficult to efficiently manage the signalling of different RATs and secondly, a need for higher UE computation in terms of resource computation as obtainable in the NSA deployment option is a challenge. This means that the mobile terminal needs to be configured for both LTE and NR all the time thereby increasing mobile terminal power consumption. The reason is that the UE can transmit/receive data using 4G base station or 5G base station. For example, consider a situation where the 5G NR is temporarily unavailable to transmit UE data as a result of cell outage or other reasons, the UE can receive/transmit via the 4G network or vice versa. Network architects are also looking for alternatives to save the UE's energy consumption and reduce configuration resources.

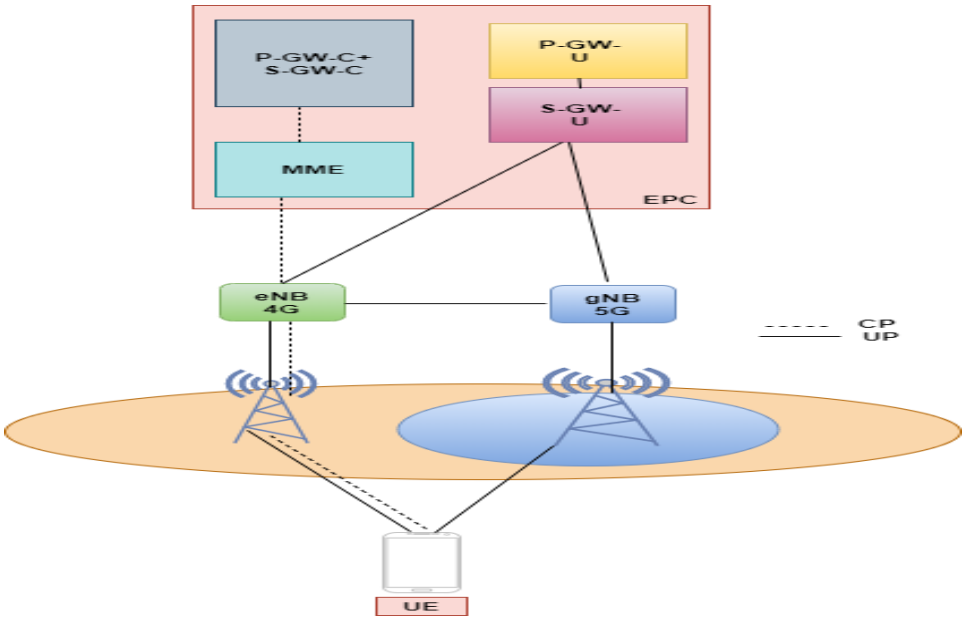


Figure 1. 4: Non-Standalone Architecture.

The user is connected to both 4G base station and 5G base station at the same time. Both 4G and 5G base stations transmit user data.

1.5 Heterogeneous Networks

Heterogeneous Networks (HetNets) refer to a kind of network deployment that uses various sizes of cells called femtocells, picocells and macrocells used to efficiently transmit user data especially in traffic hot spots as shown in Figure 1.5. Relays are mostly used to extend coverage. The advantage of a heterogenous network is the flexible deployment of base stations and provides an almost uniform broadband connection for mobile users anywhere in the network. It is a solution that is convenient to handle the ever-increasing high demand of mobile broadband connectivity for mobile users. To obtain an optimized HetNet, smarter resource allocation and co-ordination among base stations are required. It ensures efficient management of interference in such a network and maximize user throughput [16]. The proposed NR-NR architecture introduced is intended for a HetNet deployment.

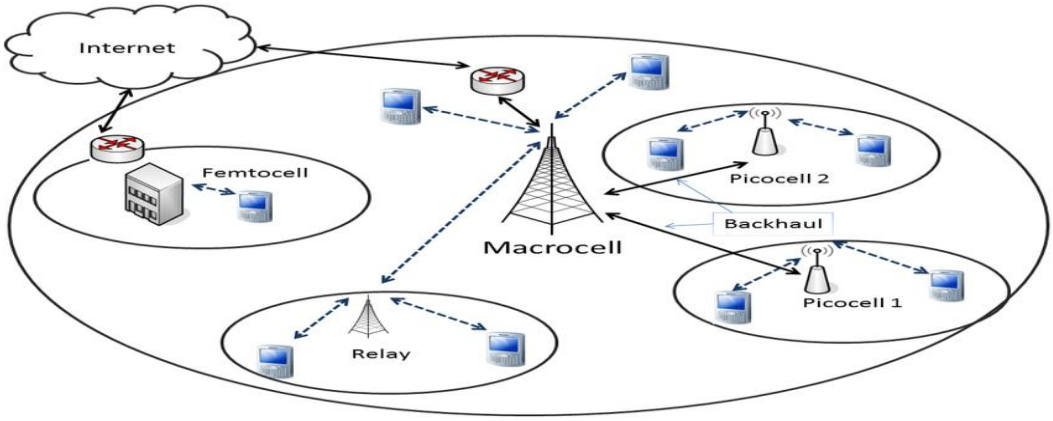


Figure 1. 5: Heterogenous Network [17].

A HetNet comprises different types of cells such as femtocells, picocells, relays and others are used to transmit user data in a non-uniform densely populated environment.

1.6 Research Motivation and Problems

The reason why it is imperative to decouple the 5G RAN into the CP and the UP is based on the fact that the tight coupling of CP and UP functions as it is in the recent architecture means that for a case of replacement or upgrade of the CP functions will inevitably require the replacement of UP functions. There could be remarkable savings in the cost of operation by preventing this replacement through a CP/UP split in the RAN. The advantage is that it will accelerate the roll out of new network functions which enables a flexible deployment of CP and UP functions. Also, in a multivendor network, it allows for CP and UP functions to be purchased from different network vendors, therefore preventing dependability from a specific network vendor.

Another important reason to decouple CP and UP is to enable independent scaling of each plane's resources. Take for instance a case where the CP and UP is not split, it means that any time the UP resources are scaled, it will call for the scaling of the CP resources and vice versa. The problem is that only a small

amount of control signalling is required for the enormous data traffic; hence, scaling each plane's resources should be strictly independent to allow for a flexible deployment. Besides, it also allows for the efficient CP signal delivery. By efficient CP signal delivery, it means that a dedicated network node is strictly transmitting only CP signals while small cells are transmitting only user data traffic. The ultimate reasoning is that decoupling the RAN allows for CP network functions and UP network functions not to be co-located in the same physical network node.

Splitting the 5G New Radio (NR) is challenging since the CP and UP functions are tightly coupled in the RAN [18]. It might be possible to fully decouple the CP and UP functions but standardization is often required in case the interface between the CP and UP has to be expanded to initiate new features such as inactivity detection, load management, bearer notification and so on which may also stagnate these features. Merging additional interfaces in an exclusive manner with combination of standardization is not a recommended solution [18]. This will tear down the benefits of CP/UP split. For example, a flexible change/upgrade of a UP function will no longer be possible if the CP functions have exclusive interface, that is, it supports only certain interfaces. Another challenge is that more testing is needed to guarantee the functional compatibility of CP and UP functions from different network vendors.

1.7 Research Objective

The 5G UP is poised to play a remarkable role in fulfilling the dynamic demand of data from mobile users. The concept of CUPS was first deployed in EPC and hitherto deployed in the CN of 5G Networks. The separation makes the 5G CN more agile, flexible, scalable and the efficient utilization of the available resources. In this same vain, CUPS can also be introduced in the 5G RAN to allow for the scaling of each plane's resources and optimize the data path for mobile users. The purpose of this research study is to propose an architecture using the principle of CUPS that can decouple the CP and UP in the 5G RAN all the way up to the UE. We envisage a non-uniform densely populated environment HetNet scenario where CP goes through separate, typically umbrella-like signalling-oriented macro cells and UP goes through data-oriented small cells. A further objective is to demonstrate the workings of the proposed architecture by analysing how the architecture handles control signalling for IMS call (VoNR); how voice traffic is transmitted in the UP during mobility, and using the NSA architecture as a baseline.

It is anticipated there will be other classes of communication services realised through the IMS. IMS itself is a generic communication framework where communication service comprises CP carrying the Session Initiation Protocol (SIP) messages and UP carrying Real Time Protocol (RTP) packets. The principle of the proposed architecture is hence, equally applicable for these IMS based communication services. Therefore, using an IMS voice call provides a general

insight on how the proposed architecture will function. It will further provide functional intuition despite that voice call contributes to a small proportion of overall network traffic.

In addition, we carry out a comparative study between the NSA and the proposed NR-NR architecture with respect to architectural differences because of the similar characteristics between the two architectures in terms of transmission between two base stations. Furthermore, both architectures are intended for UEs configured for DC and for HetNet deployment. Another justification for the comparison is also based on the fact that small cells are part of each architecture (NSA and NR-NR) and these small cells are used to improve UE throughput.

1.8 Contribution

We propose an architecture for the 5G RAN that decouples the UP and CP by introducing the principle of CUPS in the RAN. In this model, we look into the benefits of the separation of UP and CP. Furthermore, we find what is a good configuration for the suggested architecture, the design principles and what the interface between the separated planes would look like. In the course of doing this, we also investigate UE mobility and how the proposed architecture handles a VoNR call including handover procedures in the proposed architecture, and we compare them with the NSA deployment in terms of observable architectural differences.

CHAPTER 2

LITERATURE STUDY

There are several research studies in the domain of 5G New Radio which look at different methodologies to decouple the CP and the UP by applying new technologies like SDN, Cloud RAN, NFVs to split the New Radio (NR) in order to improve network performance like latency and bandwidth. Various papers indicate that decoupling of the CP and UP will allow the independent scaling of each network entities in the NR and the reduction in operational cost. Therefore, the attempt to decouple the 5G NR and to adopt a flexible CP and UP split in the RAN architecture has become an interesting topic in the research community.

Arnold et al. [18] proposed the use of SDN for the CP/UP split in the RAN and CN as shown in Figure 2.1.

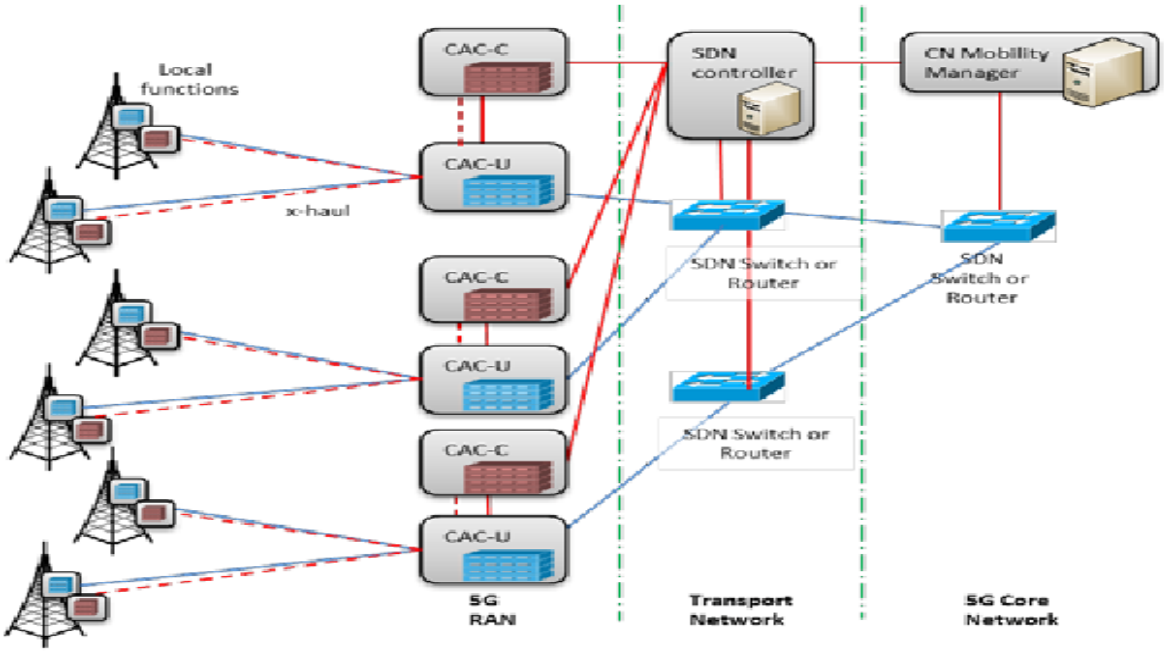


Figure 2. 1: An SDN Approach Supporting CP/UP in 5G Network [9].

SDN is an approach that can be used to enable a CP and UP split in 5G RAN network. The SDN controller coordinates the various CUs while SDN switches or routers form the transport network.

The authors presented Central Access Controllers (CACs) that centrally host both CP and UP functions in the RAN. The CP part (CAC-C) and UP part (CAC-U) perform CP and UP functions respectively while lower layers of the radio protocol stack are hosted close to the antenna sites in the 5G RAN. The transport network, which forwards the UP data from the UPF is implemented using SDN switches and the CN Mobility Manger is connected to an answerable SDN controller which imposes that data are forwarded to the correct antenna site. In their

implementation, SDN switches and routers are used to forward the UP data from the CN to the CU of the RAN. The main role of the SDN controller is to compel the data to be forwarded to the correct antenna site with a special case for mobile users. These SDN switches connect the UPF residing in the CN and the centralized unit of the RAN which forms the backhaul of the architecture. These switches and routers are controlled by the SDN controller, and the SDN controller is then connected to the AMF (CN Mobility Manager) in order to enforce decision about mobility management as shown in the figure. The authors of [18] also claim that the SDN-based approach offers additional improvements in terms of reduced control data overhead and an enhanced interoperation with fixed networks when compared to the evolved GPRS Tunnelling Protocol (eGTP) [19]. Furthermore, the authors implemented several CACs, each controlling the radio processing for a certain number of antenna sites, which are implemented as virtual functions on a server platform based on NFV principles.

They considered three mobility use cases

- Between sites within the domain of a CAC: In this case, mobility is controlled internally by the CAC within the domain and no signalling traffic is required between the RAN and the CN.
- Inter-CAC-U handover: In this case, the User Equipment (UE) moves from one CAC domain to another one. When this happens, the SDN controller generates a change of direction of the data flows.
- CN-based handover: In this case, it is the function of the AMF/SDN controller to send a path change command to the SDN switches. The SDN controller in collaboration with the AMF recalibrates a new route in the transport network set by the corresponding SDN controller and then forwards the new route to the SDN switches to implement these changes and consequently redirect the packets to the mobile user.

A similar study done by H. Gaoining et al. [20] also discusses how SDN can be used to separate the UP and the CP in the RAN. In this study, the structure of the SDN architecture is discussed.

The proposed SDN architecture consists of three parts:

- The SDN Controller
- The applications running on the North-Bound Interface (NBI)
- The South-Bound Interface (SBI)

The RAN architecture enhances a new minimum architecture by generating adequate extensions in the controller layer of the SDN controller. These extensions enable RAN programmability in terms of RAN control functions, as a specific application implementation. The applications run separately from the SDN controller and they are functionally connected to the SDN controller via the NBI over a cross-slice controller. A cross-slice controller is a specific controller that controls a given network slice and NBI is the interface of the SDN architecture. The data plane of the SDN architecture is the SBI which is the interface that connects SDN switches and routers. The NBI and the cross-slice controller communicate within the RAN and to the SBI to forward packets accordingly

depending on the rules enforced by the NBI. These NBI (SDN applications) provide control functionality and can support RAN control functions such as Radio Resource Management (RRM). To assure that the most demanding use cases such as URLLC which is most crucial for safety-critical vehicular applications, local end-to-end paths are introduced to reduce latency between vehicles and road users located in respective proximity.

Mohamed et al. [1] conducted a holistic survey of existing literature on the Control-Data Separation Architecture (CDSA) for cellular radio access since they believe that a logical separation between the control and data planes is a promising solution in a heterogenous environment deployment. The concept is to separate the signals required for full coverage from those needed to support high data transmission. A heterogeneous environment consists of a macro cell called Control Base Station (CBS) which transmits/provides control signals such as RRC and reference signals and of dedicated small cells which are called Traffic Base Station (TBS) that provide high data transmission within the CBS coverage. This will naturally lead to the separation at the signalling level. The data signal is transmitted by one of the TBSs and the corresponding control signal by the CBS. The idea is that for any UE that is inactive, which means that it is not transmitting any data, it will only receive RRC messages from the CBS for purposes like paging, and scheduling. Then when the UE is active, which means that the UE is transmitting data, its data will be dedicatedly received from the TBS. TBSs can be shut down to save energy and turned on whenever a UE needs data to transmit. Since one CBS controls multiple TBSs within its coverage, there is usually a high signalling overhead between the CBS and TBSs. This tight collaboration and excessive signalling between the CBS and TBS provide a reliable, robust and up to date information. However, lowering the CBS/TBS signalling minimizes the backhaul requirements at the cost of less reliable and out-of-date information.

A study similar to [1] was carried out by Bartelt et al. [21], where the authors investigate, characterise and model a 5G air interface while considering CDSA as a candidate. The authors exploit the global view of the CBS that controls power and the resource allocation of the various TBSs under its influence or control by using specialized interface scheduling mechanisms. Information from the UE such as position and movement history is used by the CBS to predict mobility and enforce handover optimization. It also uses a Signal-to-Noise (SNR) database to predict the channel quality. In addition to this, the authors also investigate the usage of carrier aggregation to enable a seamless implementation of the CDSA standard.

Another proposed method for decoupling the RAN into UP and CP functions as investigated by Ali-Ahmed et al. [22] is based on the concept of combining SDN and CDSA. It uses the classical SDN approach of using a centralized controller and it reduces signalling overhead by terminating some of the control information in the local controller resulting in a hybrid centralized controller. In other words, this model uses two controllers, the main centralized SDN controller and a local controller. The CP is built as software on the local controller which is then hosted in a RAN element and is generally used for fast grained control functionalities.

Several local controllers are connected to the main SDN controller hosted in a data centre, which is used for non-latency sensitive control functionalities. The CDSA is adopted by directing the control path of the UE to the local controllers and forwarding the data path to the UPF which supplies the data from the data network.

R. Guerzoni et al. [23] proposed a novel architecture for the advanced 5G network infrastructure by harvesting the latest advances in SDN, NFV and edge computing platforms. A unified controller is proposed which consists of three logical controllers: A Device Controller (DC), an Edge Controller (EC) and an Orchestration Controller (OC). These controllers run different applications to implement 5G control plane. This architecture allows 5G network operators to dynamically instantiate logical architectures and implement network functions and services in an optimal location with respect to network requirements. The authors show a reduction in latency even though it is far from the 1 ms latency requirement of some 5G applications. The DC is located in the device and it is responsible for the physical layer connectivity to the 5G network. The DC handles NAS functions such as access selection and network selection. The EC implements the 5G network CP, packet routing and transfer. The implementation of the EC is distributed over the cloud infrastructure via a set of control applications. The OC coordinates the utilization of cloud resources. In this proposed architecture, when a user performs network attachment, an address is allocated at the UPF and a forwarding path to route the data from the UPF to the device is established. One benefit of this architecture is reconfigurability which is a key feature of the plastic architecture. Furthermore, the proposed model does not make use of tunnelling protocols such as GTP. The authors claim that this performance improvement is valid for device trigged requests as well as handovers with mobility reallocation. Also, the 'always on' concept which consists of establishing data bearers after the device is powered on as discussed in the high-level functional description of [23] indicates that latency for creating the forwarding paths can be possibly reduced to zero by dynamically configuring the SDN based infrastructure for devices performing mission critical devices. Finally, for delay critical services the UP latency can be minimized by building ad-hoc virtual link embedding algorithms in the controllers to guarantee that the cloud infrastructure is optimally employed while fulfilling service latency requirements.

Cloud Radio Access Network (Cloud-RAN) is a new mobile network architecture. It uses the cloud computing capabilities to enhance the Quality of Service (QoS) in 5G networks. The basic idea of Cloud-RAN is to separate the BBU from the RRH and move the BBU to the cloud for processing and management. Chabbouh and Rejeb [3] proposed a cloud-RAN architecture to provide flexibility in the functional split of the radio stack in the deployment of RAN functions by introducing a cloud RRH. The cloud RRH will provide a flexible access network function split between the edge and the central cloud depending on actual network needs and characteristics. The functional split introduces more degrees of freedom in network processing and function executions. Moreover, additional computation and storage resources are included in the cloud RRH for computation offloading

and these resources are represented by what is called “cloud containers” which are storage resources. When it receives a task (task from the UE to be processed), it will be executed in the cloud RRH. The proposed model aims to make better usage of resources while improving network performance. Furthermore, additional resources can be deployed closer to the user to allow total or partial offloading of computation in cloud RRH and improve energy efficiency. Thus, the unique model entails redefining the RRH and BBU functions and change the interface between BBU and RRH from CPRI to Ethernet. The central element in the presented model is the adjustable and flexible functional split of the radio protocol stack between the cloud platform and the edge cloud RRH. If the received task is an offload request, it will be processed and executed in the cloud-RRH which is usually close (suited at the cell level) to the user and thus reduction in user latency may be achieved. The model is characterised by on-demand resources provisioning and elasticity which ensures mobile communication service delivery to be closely adapted to meet the needs of the operators. The authors also propose a new functional entity called Cloudlet Manger (CM) with functionalities such as container placement, container management and application scheduling. Mobile users can access their services directly in the edge cloud by requesting the CM to instantiate containers in the edge and offload the service.

Du et al. [24] propose to manage the UP and CP transmissions for Multi-Connectivity (MC) sessions separated from mono-connectivity which is called CP/UP (C/U) split MC to boost user throughput. The background information is that in 3GPP LTE standardization for Dual Connectivity (DC), Carrier Aggregation (CA) has been specified to boost capacity. The results presented are based on simulations conducted in a Hetnet scenario with two frequency layers used for macro and small cells respectively. The authors investigated the performance evaluation of the C/U split solution for an eNB+gNB scenario. In this model, the eNB forms the macro site while the gNB forms the small cell. The results obtained showed that the average throughput of the UE increased significantly. For all simulation results shown, the C/U MC split showed the best throughput performance when the transmission capability is unbalanced between the macro cells and the small cells. For their research study, they concluded it will be preferred to move all the UP transmissions to the small cells while keeping only the CP transmissions in the macro cells.

To summarize, the literature study considered in this thesis shows that the biggest challenge of using SDN in the RAN is the inherent delay between the centralized unit and the individual radio elements. This delay is crucial since some RAN operations such as scheduling require real time control. The controller operations for such a network function should be extremely fast, supporting hard real-time applications. Another challenge of SDN solution is the difficulty in managing both RAN functions and CN functions simultaneously using a single SDN controller. It is suggested that using two SDN controllers could be a way to decouple the RAN though it could introduce additional complexity to the already complex 5G network as proposed by H. Gaoining et al [20]. The disadvantage of using SDN to split the RAN increases the bandwidth requirement of the transport network which strongly

increases the limitation of the proposed approach. The introduction of a cross-slice controller would introduce more flexibility in the RAN specifically for a given slice, but it will come with a price of a more complicated SDN which might be difficult to manage. From the review of the SDN approaches, it can be argued that SDN approach is not expected to be a good solution for a CP/UP split in the RAN as a result of the complexity in manging the SDN controller. The Cloud RAN proposed by Chabbouh [3] is not an ultimate solution that splits the RAN into CP and UP both; rather their solution provides a functional split between the edge cloud and central cloud for the UP data. The model aims to make better usage of network resources to improve network performance by introducing Cloudlet Manager to allow users access services directly. From the works of Du et al [24]. it is suggested that to improve throughput of mobile users and allow for scalability of resources, it is preferred to move all UP transmissions to small cells while keeping only CP signalling in macro cells.

Finally, we present the NR-NR architecture using the principle of CUPS to separate the CP and UP of the 5G RAN by introducing two NR base stations connected in DC. By decoupling the CP and UP of the RAN using two NRs has allowed for independent scaling of each plane's resources and for a flexible deployment. Besides, the NR-NR architecture presents a framework of the future 5G network which ensures an uninterrupted VoNR call during handover procedures, and also allows for a multi-vendor NFs to be deployed within the same physical infrastructure (gNB).

CHAPTER 3

5G NETWORK ARCHITECTURE

This chapter discusses four major aspects of the 5G network analysis. The first section discusses the architecture of the 5G RAN and the deployment options available in the RAN. It also discusses the key interfaces and network protocols of the RAN. The second section presents the architecture of the 5GCN and the major NFs are described. The third section discusses in detail the principle of CUPS, the advantages and disadvantages of CUPS. Finally, the last section discusses the IMS, the IMS network architecture and the enabling protocols such as SIP and RTP. In addition, the connection between IMS and the 5G CN is extensively discussed.

3.1 The Architecture of the 5G Radio Access Network

The 5G RAN comprises an antenna with a Remote Radio Head (RRH) also called Remote Radio Unit (RRU) which is located just below the antenna and a Baseband Unit (BBU) which is traditionally located at the bottom of the cell tower. The RRH is comprised of radios, Radio Frequency (RF) amplifiers and mixers that perform some radio signal processing. The BBU performs both real time and non-real time control functions such as scheduling, error correction, multiplexing and so on. The 5G radio node is called the next-generation NodeB (gNB) which is equivalent to what is called evolved NodeB (eNB) in the 4G network. The new air interface of the 5G network is called New Radio (NR). The functions of the gNB entity include performing dynamic allocation of resources to UEs in both uplink and downlink, radio admission control, radio bearer configuration and so on.

The BBU is divided into the Distributed Unit (DU) and the Centralized Unit (CU). The CU is a logical node or entity that manages non-real time control functions of the gNB while the DU is a logical node or entity that manages physical layer processing and layer-2 real time control functions of the gNB. This description is based on a certain split option; multiple split options are possible which shall be discussed later in the chapter. The activities of the DUs are controlled by the CU. Figure 3.1 shows the architecture of the 5G RAN which comprises the RRH connected to the DUs and from the DUs connected to the CU and then finally to the 5GC or 5GCN. The digitalized RF signals coming to/from the antenna are transported to the DU via the fronthaul interface using the evolved Common Public Radio Interface (eCPRI) [25] standard. The fronthaul is the transport network linking the RRH to the DU and the connection usually runs over fiber. The eCPRI standard is an outcome of an industry co-operation that defines specifications for RRU and the DU via the fronthaul. Normally, the RRH is close to the antenna to reduce cable loss unlike the legacy base stations (2G and 3G) where they are significantly separated by some few metres located within the same site.

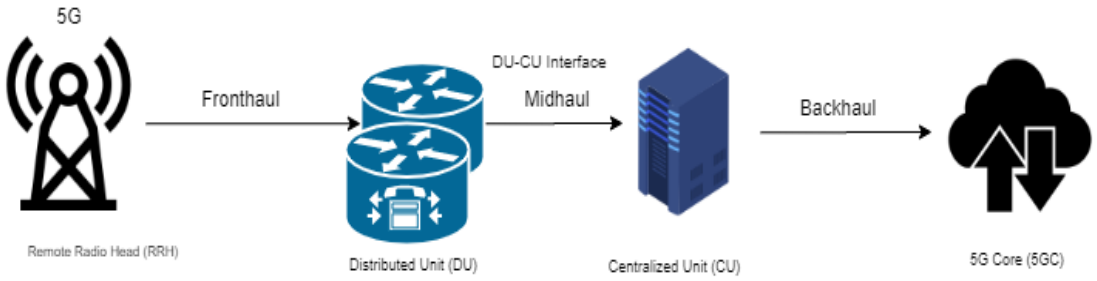


Figure 3. 1: The Architecture of 5G RAN.

The mid haul is the transport network infrastructure that connects the DU to the CU also called DU-CU interface. The backhaul is the transport network infrastructure that connects the CU to the 5GCN which is further connected to internet. As a result of high video-centric traffic and the need to reduce latency for URLLC, the backhaul connection is often carried out via a high-speed fiber. The backhaul connection can also be ethernet, copper wire, microwaves or satellite systems [14].

The reference point that connects the 5G RAN to the 5GCN is the Next Generation (NG) interface which is divided into the NG-Control plane (NG-C) and the NG-User plane (NG-U) with connections to the Access and Mobility Function (AMF) and the User Plane Function (UPF) of the 5GCN respectively which is visualised later under Section 3.3. The AMF is in charge of user mobility/granting of access to the network as well as other functions and the UPF forms the anchor point for the transmission of user IP address to/from the internet to the user. The NG interface supports procedures to establish, maintain and the release of bearers (tunnels) in order to perform inter-RAT and intra-RAT handover. Inter-RAT handover involves releasing or transferring of radio connection of a given access technology to another access technology; for example, from 4G technology to 5G or 3G technology.

There are various types of deployments of the RAN; such as the Distributed RAN (D-RAN) and the Centralized RAN (C-RAN). The D-RAN deployment entails that the RRH and the BBUs are decentralized. A distributed architecture involves adding more functionalities of the network protocol stack to the distributed entities (DU). This option is chosen when there is less demand in the requirements of the transport network such as bandwidth, latency and so on [18]. In the C-RAN deployment, a centralized BBU usually located at a different site co-ordinate various RRHs. A centralized architecture involves adding more functionalities of the network protocol stack to the centralized entities (CU).

Figure 3.2 (a) shows the D-RAN architecture, it consists of decentralized BBUs connected via fiber or copper to the 5GCN. The advantages of D-RAN are; there is less complex co-ordination of radio capabilities across a set of RRHs that is, all Radio Resource Management (RRM) functions are close to the radio interface which makes it fast to respond to variations and secondly, D-RAN has less stringent demand on the fronthaul requirements which reduces the cost of deployment. On the other hand, the D-RAN deployment has some disadvantages

that accelerated its evolution to the C-RAN. In D-RAN, the BBU is limited for resource pooling and it experiences a low energy efficiency of the RRH [26]. Secondly, the RRH is usually more expensive.

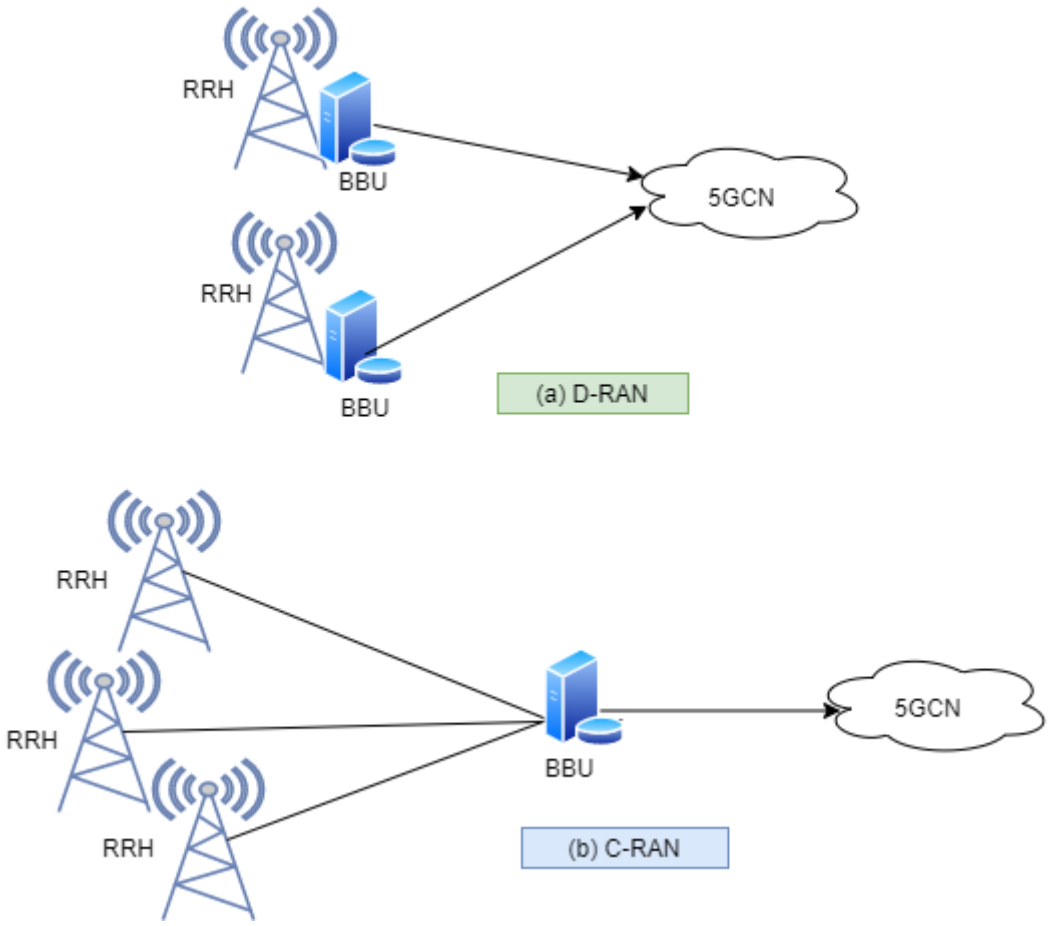


Figure 3. 2: Deployment Scenarios for the 5G RAN.

The C-RAN architecture is shown in Figure 3.2 (b). In C-RAN deployment, the baseband processing for a large number of RRHs is centralized and as a result of the BBU having efficient ability to co-ordinate among multiple RRHs. The advantage of C-RAN is the pooling of high level BBU functions resulting in trunking gains and also for a possibility of Distributed Multiple Input Multiple Output (D-MIMO). D-MIMO is a technique that improves the spectral efficiency of cellular networks [27]. One disadvantage of C-RAN is that it leads to stringent fronthaul requirements in terms of latency and bandwidth which makes fiber the only possible solution to meet the demanding fronthaul requirement. Secondly, a complete C-RAN makes it difficult to handle the uncertainty in user mobility as density of base stations increases [28].

The protocol stack in the RAN consists of Layer 1 (low/high-Phy sub layers), Layer 2 (MAC and RLC sub layers) and Layer 3 (RRC, SDAP and PDCP sub layers). Figure 3.3 shows 8 possible split options in the RAN for either downlink or uplink data.

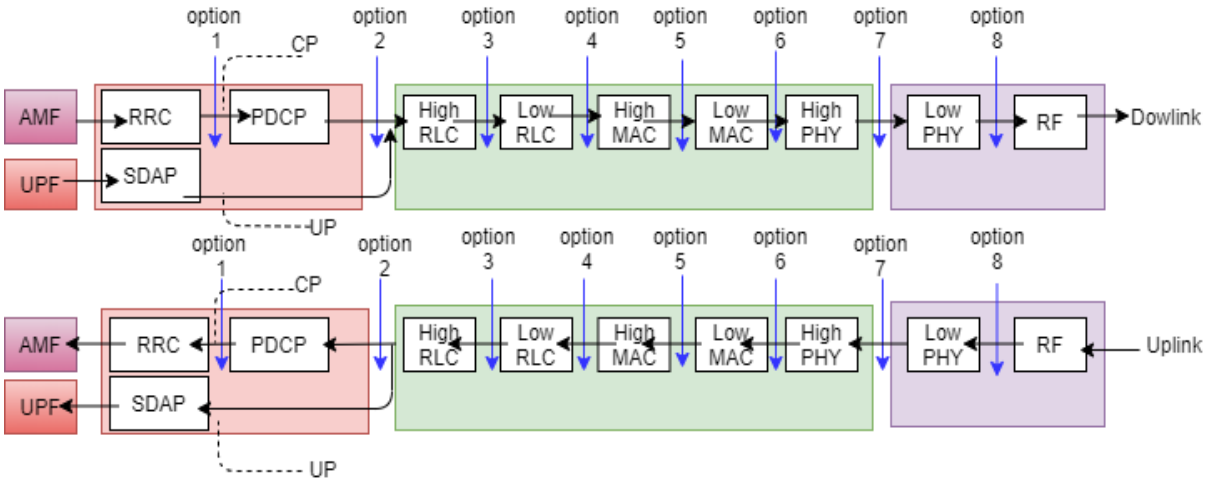


Figure 3. 3: Possible Split Options for 5G-NR.

Split options showing the possible deployment of the various radio protocol stacks either in the CU or DU. Split Option 2 means that the RRC, PDCP and SDAP layers are residing in the CU.

The CU is represented by the pink block where the RRC layer and PDCP layer belong to the CP and SDAP layer belonging to the UP. The green block and purple block represent the DU and RRH respectively. Higher layer split options refer from split option 1 to split option 5 while the lower layer split options refer from split option 6 to split option 8. The split option chosen depends on the network requirements and services with each option having its own advantages and disadvantages. The choice of the split option aims to optimize either for a more centralized or distributed system as required by the network operator [14].

In this study, split Option 2 is chosen. The reason for choosing option 2 is to allow for an implementation separating the Radio Resource Control (RRC) layer and the Packet Data Convergence Protocol (PDCP) layer into two entities. One PDCP entity for the CP stack and the other PDCP entity for the UP stack allowing for a flexible deployment in any CU. Split Option 2 involves moving the PDCP and the RRC in the CU and the Radio Link Control (RLC), Medium Access Control (MAC), and the physical layer (PHY) in the DU.

3.1.1 NG-RAN Key Interface and Protocols

The NG-RAN nodes are connected by means of horizontal Xn interface which are fundamentally used for mobility, Multi-Connectivity and Self Optimized Network (SON) [30]. The Xn interface [31] is used to interconnect two gNBs or a gNB and a ng-eNB (LTE base station that can connect to the 5GCN) as shown in Figure 3.4.

In other words, the Xn interface is a logical point-to-point interface between two NG-RAN nodes as shown in the figure. The Xn interface transmits CP signals and UP data via the Xn-C and Xn-U interfaces, respectively. The UP-protocol stack of the Xn-U interface, relies on the GPRS Tunnelling Protocol-User (GTP-U) [32] running on top of the User Datagram Protocol (UDP) and the Internet Protocol (IP). A GTP-U is used to create, modify, encapsulate and delete tunnels/bearers that transmit IP payload between the UE, the gNB and the internet.

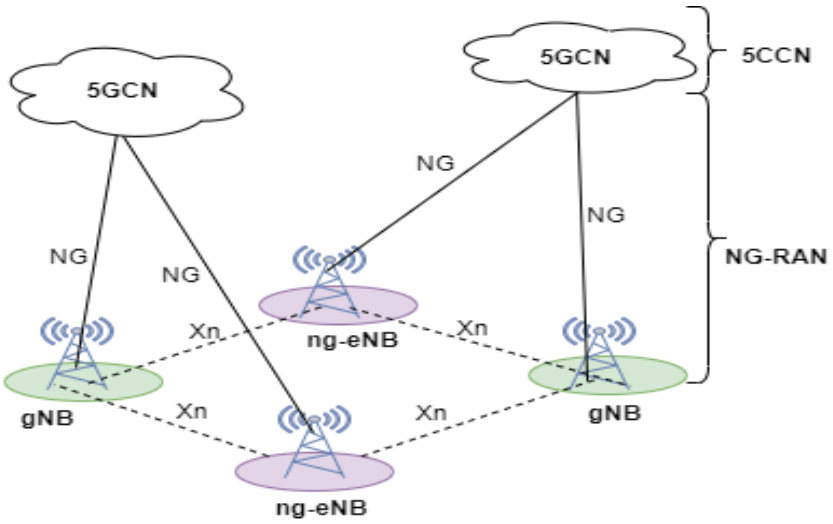


Figure 3. 4: NG-RAN Architecture showing Xn Interface.

The ng-eNB is an enhanced eNB that can connect to the 5GCN.

A Tunnel Endpoint Identifier (TEID) [33] is a value assigned to each tunnel to identify which particular tunnel is carrying user data for either uplink or downlink. For example, a tunnel carrying voice traffic is different from a tunnel carrying data traffic and the identification of these tunnels is done using TEID. For the CP architecture, Scream Control Transmission Protocol (SCTP) is used. SCTP is a transport layer protocol that is used for transmitting multiple streams of control signals between two end points when a connection has been established between these two points.

The protocol stack of Xn-U and Xn-C are shown in Figure 3.5. The left-hand side of the Figure shows the Xn-U protocol stack which transmits UP PDU payload running on top of UDP/IP.

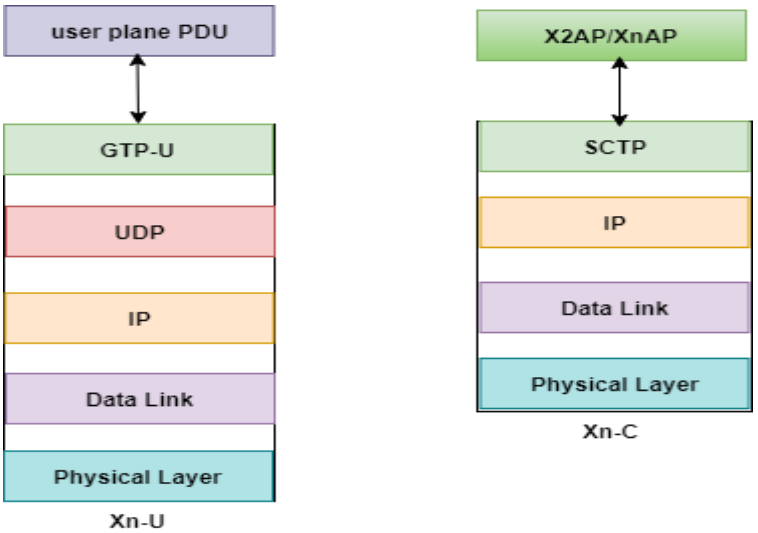


Figure 3. 5: Xn-U and Xn-C control stack.

GTP-U is the protocol handling the user data running on UDP/IP while SCTP is the protocol handling the control signals running on IP. SCTP is used in the control stack because it ensures a guaranteed delivery of control signals.

The Xn-U interface provides a non-guaranteed delivery of UP data between any two NR-RAN nodes and also supports mobility operation. The right-hand side of Figure 3.5 shows the Xn-C interface protocol stack which uses Xn-Application Protocol (Xn-AP) running on top of SCTP/IP to provide a guaranteed delivery of CP signals between any two NR-RAN nodes. The protocol allows for interface maintenance, mobility and multi-connectivity operation [30]. The protocols layers in the RAN consist of the UP stack which is used for the transfer of user data (IP packets) between the gNB and the UE and the CP stack used for the transfer of control signalling (RRC messages) between the gNB and the UE. The UP-protocol stack is shown in Figure 3.6 for the UE and gNB.

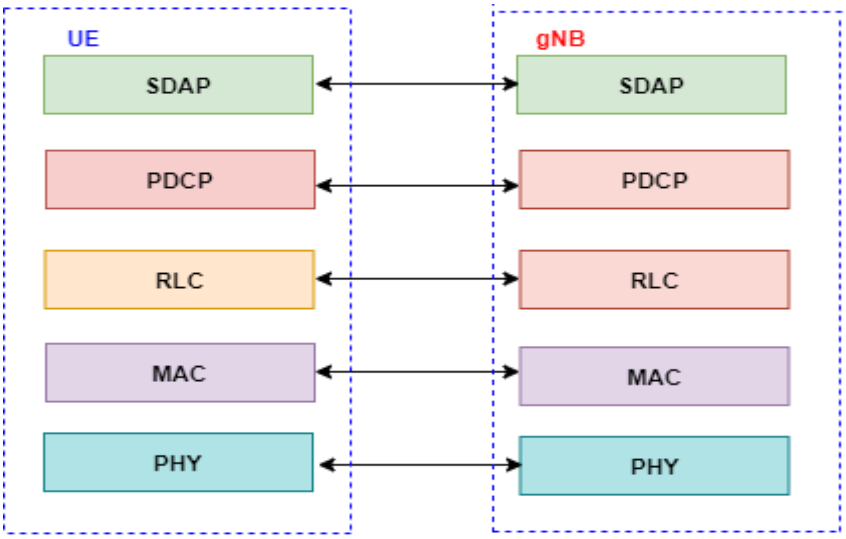


Figure 3. 6: User plane protocol stack.

UE 'talks' to gNB via each layer in the protocol stack

The UP-protocol stack is comprised of the following layers:

- **Service Data Adaptation Protocol (SDAP):** This is introduced in 5G to support the new QoS model of the 5G packet data session. In LTE there is no packet differentiation between different flows of packets that have the same QoS index. To solve this problem (packet differentiation between different flows), SDAP layer was introduced as a new layer allowing the CN to configure different QoS requirements for different QoS flows for a packet data session. Also, the SDAP layer provides mapping of QoS flows with different QoS requirements to radio bearers.
- **Packet Data Convergence Protocol (PDCP):** The main function is to provide header compression and decompression and security functions like ciphering and integrity protection. It also reorders and duplicates messages.
- **Radio Link Control (RLC):** The main function is to provide segmentation in order to match the transmitted IP packets size to the available radio resources. The RLC layers also performs error correction through AQR. Note that this layer no longer provides concatenation compared to LTE.

- **Medium Access Control (MAC):** The main function is scheduling of radio resources and also to provide multiplexing and de-multiplexing of data from different radio bearers to the transport blocks that are carried by the physical layer. It also performs error correction by Hybrid Automatic Request Repeat (HARQ). The MAC layer also carries the control signal for the purpose of beam management within the physical layer.
- **Physical Layer (PHY):** The main function of the physical layer is in the transmission of electromagnetic waves or radio signals. It also specifies the bandwidth and maximum capacity of the transport medium.

Figure 3.7 shows the CP protocol stack.

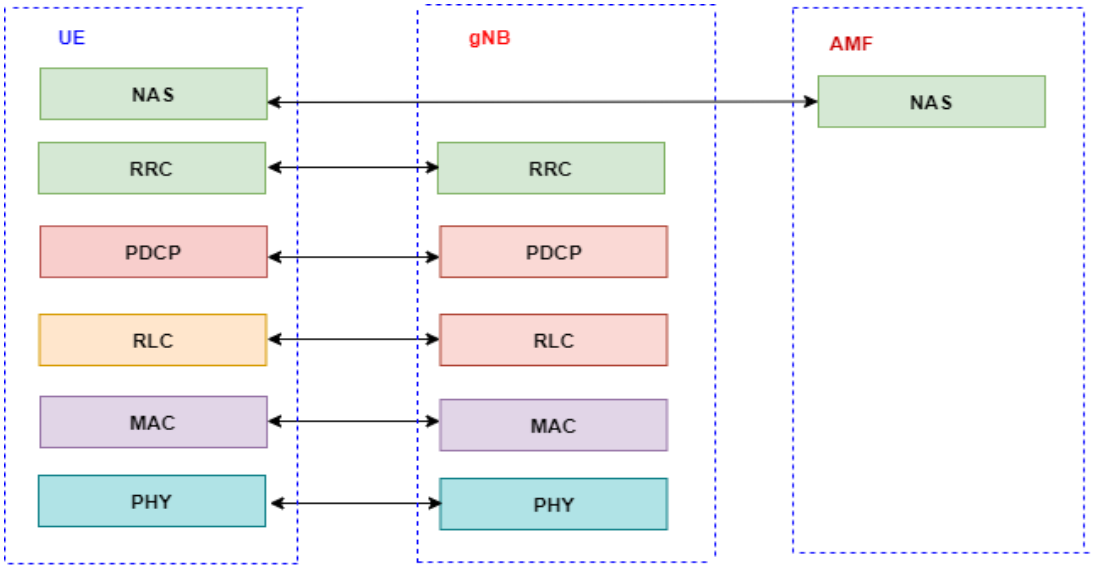


Figure 3. 7: Control Plane Protocol Stack.

UE 'talks' directly to the AMF via the NAS signalling. NAS signalling is a non-radio control signalling.

The CP stack is made of the following layers:

- **Non-Access Stratum (NAS):** This responsible for non-radio control signalling between the UE and the AMF. Such control signalling is related to mobility, registration, authentication and session management.
- **Radio Resource Controller (RRC):** The RRC layer is used for control and configuration of the radio related functions of the UE. In addition, the RRC layer supports the mechanism that enables the UE to request specific system information for the consumption of radio resources such as reporting measurements for the support of handover [17]. The other layers in the control stack perform the same functions as in the UP-protocol stack.

3.1.2 Common Public Radio Interface CPRI

The CPRI is an industry forum that defines the publicly available specification for the interface between a Radio Equipment Controller (REC) and a Radio Equipment (RE) in wireless networks [34]. In other words, CPRI specifies the

digital interface between the part of DU called REC and the RRH called RE. User data is transported via the CPRI data frame as baseband I/Q stream [1]. I/Q, which stands for In-Phase and Quadrature signals, refers to two sinusoids that are of same frequency but out of phase by 90 degrees. In the 4G RAN, the CPRI standard is used to transmit digital signals over the fronthaul. CPRI will not work effectively in the 5G RAN because of the stringent fronthaul requirements in terms of both throughput and latency. The high bandwidth requirement of the fronthaul is a result of centralization which imposes stringent requirements on the transport network because the higher layer processing of the radio stack is moved to the CU. In addition, CPRI requires data rate of about 24Gbps per cell and a round-trip fronthaul latency below 200μs, with low jitter and high reliability [21]. These requirements can only be achieved by using high capacity fiber or point to point wireless link [21]. This makes the deployment of fronthaul network expensive thereby reducing the gains from centralization. As a result of this high bandwidth requirement, enhanced CPRI (eCPRI) was introduced in 2017 to improve efficiency, link capacity and reduce the burden on the fiber. The eCPRI also enables functional decomposition and supports service points for UP traffic, synchronization and control and management [35]. These service points are generally operated by eCPRI protocol stack over IP/Ethernet.

Figure 3.8 shows eCPRI which is comprised of enhanced Radio Equipment Controller (eREC) and enhanced Radio Equipment (eRE) which are physically separated and connected via a transport network.

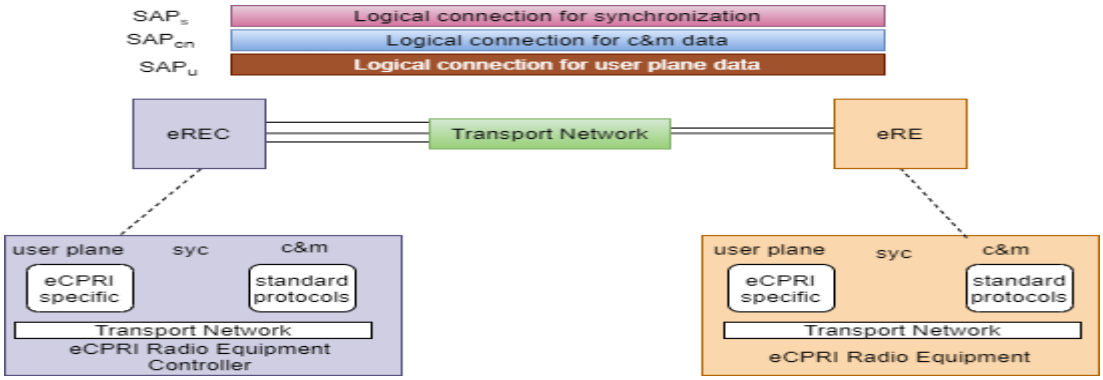


Figure 3. 8: eCPRI Interface [11].

The eCPRI is a standard used to transmit digital signal over the fronthaul. The fronthaul is the transport network linking the RRH and the DU. The eCPRI protocol stack runs over IP/Ethernet.

There are three Service Application Protocols (SAP) running on the transport network namely the SAP_s, SAP_{cn} and SAP_u that transmit different kinds of packets. SAP_s is the logical connection transmitting packets for synchronization, the SAP_{cn} is used to transmit control and management packets and the SAP_u is used to transmit user payload data. The eREC implements the higher layer functions and some parts of the physical layer functions while eRE implements the rest of the physical layer functions including the RF functions. In general, eCPRI packetizes data and also synchronizes control signals as mentioned previously. It should be noted that the eCPRI is not specific on the type of fronthaul transport network in

use, thus, any type of network can be used for eCPRI, provided that its fronthaul requirements (such as bandwidth) are met. Key features of eCPRI are:

- Reduction in the required bandwidth, which enables the required bandwidth to scale flexibly according to the UP traffic.
- It encourages the utilization of Ethernet and IP which guarantees future evolution.

The information carried on the eCPRI interface is the following:

- The UP, which contains the user data, real-time control data, and other eCPRI services.
- The control and management plane, which is exchanged between the eREC and the eRE for control signalling.
- The synchronization plane.

Advantages of eCPRI Compared to CPRI:

- A single ethernet network can simultaneously carry eCPRI traffic from different vendors.
- The interface is future proof, allowing new features to be installed by software updates in the radio network.
- Jitter and latency can be reduced for high priority traffic using Time Sensitive Networking (TSN) [36] standard such as IEEE 802.1CM which supports pre-emption of low priority packets.
- The eCPRI interface is a real-time traffic interface that enables the use of sophisticated coordination algorithms to ensure the best radio performance.

Disadvantages of eCPRI Compared to CPRI:

- The eCPRI standard is vendor specific which means that network operators cannot mix RRU and BBU components from different vendors.
- The eCPRI has no backward compatibility.

3.1.3 Fronthaul, Mid-haul and Backhaul

Fronthaul, as already mentioned in the previous sections, is the transport network infrastructure between the RRH and the DU. Figure 3.9 shows the various x-hauls in the 5G architecture namely the fronthaul, mid-haul and backhaul.

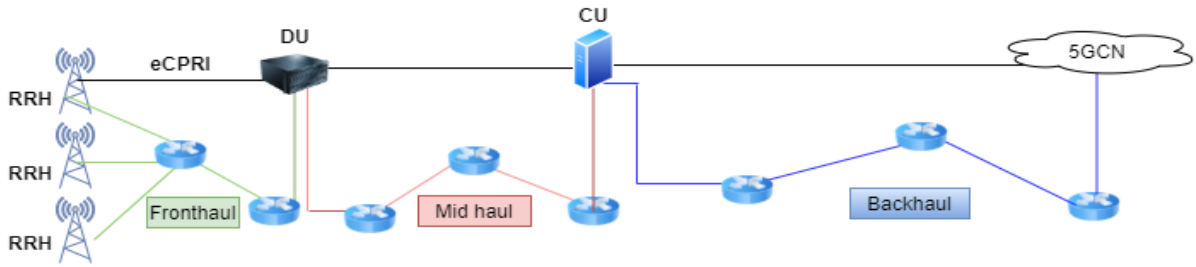


Figure 3. 9: 5G Architecture showing the x-haul.

The gNB node is made of RRU, DU, CU and CN. The RRU is connected to DU via the fronthaul while the DU is connected to CU via the midhaul and finally the CU is connected to the core network via the backhaul.

Since 5G supports high bandwidth and low latency services, it becomes imperative to use a high-speed link at the fronthaul which is the most demanding interface in the 5G RAN [37]. The challenge with CPRI over long distance is the very demanding bandwidth, which requires fiber pair in most deployments. In addition, it also does not scale well to Massive Multiple Input Multiple Output (MIMO) and other new RAN technologies [38]. For the C-RAN concept to be widely deployed, the cost of optical fiber needed for the fronthaul needs to be reduced and by adapting to eCPRI, the protocol running on the fronthaul becomes ethernet which is a solution that reduces the cost [39]. The mid-haul is the transport network infrastructure between the DU and CU. The latency on this link should be around 1 ms for some 5G applications and a CU may be controlling several DUs in an 80 km radius [37]. The backhaul is the transport network infrastructure connecting the CU to the CN. The backhaul is usually an optical fiber, ethernet or microwaves depending on the deployment scenario and area of deployment. For highly dense population areas, high speed optic fibers are usually the preferred backhaul to meet the enormous data consumption.

3.2 The Architecture of the 5G Core Network

3.2.1 The Core Network of 5G

The EPC, which is the CN in LTE, consists of the Mobility and Management Entity (MME) plus Home Subscribers Server (HSS), the Serving Gateway (S-Gw), the Packet Data Network Gateway (P-Gw) and the Policy and Charging Rules Function (PCRF). The MME performs mobility management such as registration, paging, handover etc and also performs authentication and verification of the UE. The PCRF provides the charging policy and determines the QoS to be applied to the packets and then passes this information to the P-Gw which acts as an anchor point for the allocation of the UE IP address. The P-Gw connects the UE to the PDN or the internet and meanwhile, the S-Gw relays the data packets from the P-Gw to the UE via the eNB. By the application of Control Plane User Plane separation (CUPS) in the EPC, the S-Gw and P-Gw are functionally decoupled into the UP and CP respectively. Hence, the S-Gw is decoupled into S-Gw-CP and S-Gw-UP. This is also applicable to the P-Gw; that is P-Gw decoupled into P-Gw-CP and P-Gw-UP. The main reason for this de-couple is to allow for independent scaling of each network entity and to also allow for flexibility of network entities.

The network entities identified in LTE perform the same functions and additional functions with respect to the 5GCN. In the 5GCN, the Access and Mobility Function (AMF) performs functions as the MME in LTE. AMF's main function is mobility management and registration of the UE. The S-Gw-CP and P-Gw-CP are now grouped together in the 5GCN to form the Session Management Function (SMF). The function of the SMF is to establish sessions to be used to forward packets from the PDN to the UE. Hence, the SMF is specifically a CP entity in the 5GCN. The S-Gw-UP and P-Gw-UP are grouped together to form the UPF which allocates an IP address to the UE in order to forward the UE data packets from the PDN to the UE via the gNB. The UPF is purely a UP network function. The Policy Control Function (PCF) performs the same function as the PCRF entity in

LTE. In general, it can be deduced that the 5GCN is an evolution of the EPS with added functionalities to meet the requirements of the 5G services and applications. Other entities that were added in 5GCN include the Authentication Server Function (AUSF), the Network Repository Function (NRF), the Network Exposure Function (NEF) and the Network Slice Selection Function (NSSF). Figure 3.10 shows the reference point architecture of the 5GCN. The reference points in the figure show the interaction between the NFs. Recall that NFs are software running on dedicated hardware to perform a specific/dedicated function in the network without the need of buying new equipment. Whenever there is an added feature or service, the software is updated thereby reducing operational cost.

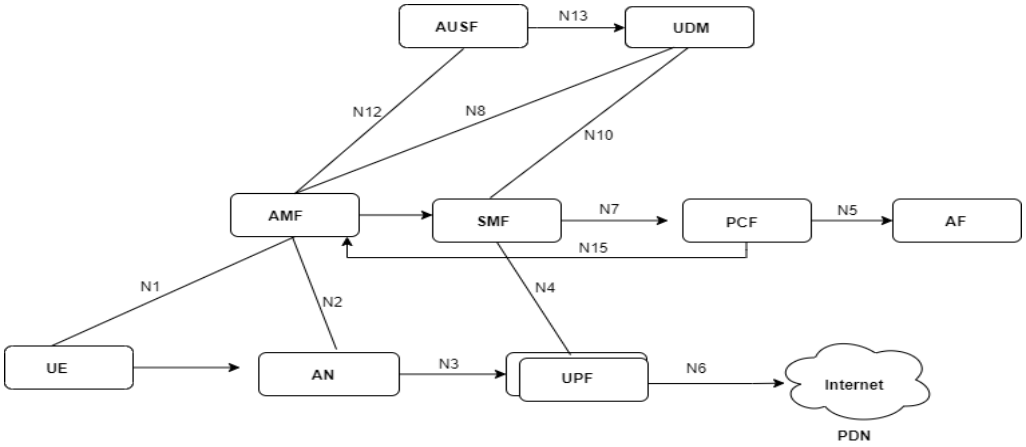


Figure 3. 10: Architecture of the 5G CN.

The reference points represented as Nx is a point-to-point interface that interconnects network elements. Usually the signalling procedure is specified for each reference point.

The functions of each component in 5GCN are briefly summarized below

- **Session Management Function (SMF**): This entity acts as an anchor for the UE IP address for the purpose of IP addressing, allocation and management. Furthermore, this entity performs session establishment, session management, and traffic steering to the UPF for data tunnelling.
- **Access Management Function (AMF):** This entity is analogous to Mobility Management Entity (MME) in LTE. It supports registration, mobility, access authentication and authorization, lawful intercept and transporting session management messages between the UE and SMF.
- **Policy Control Function (PCF):** This entity is responsible for enforcing all forms of charging polices to be used in billing the UE. Further, the PCF also provides mobility management policies and granting UE access to the AMF.
- **Application Function (AF):** This entity supports application functions that can influence traffic or data that are being delivered to the UE. The AF has interaction with the PCF to influence policy control on the UE's data.
- **Authentication Server Function (AUSF):** The AUSF acts as an authentication sever for the UEs in the network.
- **Unified Data Management (UDM):** This entity acts as a policing agent and supports Authentication and Key Agreement (AKA) credentials between the

UE and the network. It performs the 'challenge' and 'response' command to the UEs to authenticate and validate the UEs based on their subscription data.

- **User Plane Function (UPF):** The UPF is responsible for packet forwarding, packet inspection, QoS handling and routing to support traffic flow in the network.
- **Data Network (DN):** This is the internet or other operator's application service.

The UE attaches to the network by establishing a connection to the AMF via Non-Access Stratum (NAS) signalling through the gNB. The AMF then registers the UE to the network by Importing the subscriber's profile from the UDM after performing authentication and verification of the UE through the AUSF. The SMF allocates an IP address to the UE and also acts as an anchor point for the UE data to the AF in the DN. The SMF configures the UE data and selects which UPF to be used in tunnelling the data packets from the DN towards the gNB and thus to the UE. The PCF enforces all forms of policy and charging related controls and applies the appropriate QoS to the relevant data packets. Then, the Data Radio Bearers (DRB) or tunnels steer the traffic through the UPF and to the UE. Note that this connection establishment description is similar but different in the sequence of actions for the establishment of the PDU session to be discussed next.

3.2.2 The Protocol Data Unit Sessions

In LTE, the IP connectivity between the UE and the PDN is established via the PDN connection. In the same vain, the 5G PDN is referred as the Protocol Data Unit (PDU) session. A PDU session delivers IP packets that are labelled with the UE IP address through a logical path between the UE and the DN and can support IPv4, IPv6 and ethernet. PDU sessions are made of bearers that carry signals/data between the gNB and the UE. These bearers are encapsulated via tunnels by the GTP. There are two types of radio bearers namely the Signal Radio Bearer (SRB) and the Data Radio Bearer (DRB). The SRB is used to transmit RRC messages and NAS messages from the UE to the gNB. NAS is a non-radio signalling message that is transmitted from the UE to the AMF in the CN. A DRB is a logical tunnel through which IP packets are delivered from the DN to the UE. There can also be multiple DRBs per UE, each dedicated to a specific user application. A DRB is established through the UP upon request from the UE to the gNB. This is different from the 4G network whereby default radio bearers are established without request as long as the UE successfully attaches itself to the MME. The default bearer for the 4G network has no guaranteed bit rate and has best effort QoS which is also same as the 5G.

In contrast to the LTE bearers, where the bearers provide insufficient support to guarantee a specific QoS, a modified QoS model was introduced in the 5G architecture [40]. In this model, all traffic that is mapped to the same QoS flow receives the same mapping or forwarding rule. By mapping it simply means that they receive the same kind of rules or treatment in terms of QoS when transmitting the packets. In the QoS flow model, numerous QoS flows moving through a PDU

session tunnel set up between the CN and the gNB are mapped to separate radio bearers. Specifically, when there is a need for remapping a QoS flows and a radio bearer, the CN assigns a new QoS Flow Identifier (QFI) [41] labelled in the packets and sends it to the gNB. The SDAP entity is responsible for the mapping of QoS flows to a data radio bearer as illustrated in Figure 3.11. The different colours show different QoS flows that are mapped to a data radio bearer. The data coming from the UPF are carried as different QoS flows inside a GTP-U tunnel to the gNB and then to the UE via data radio bearers. The opposite sequence takes place for uplink data.

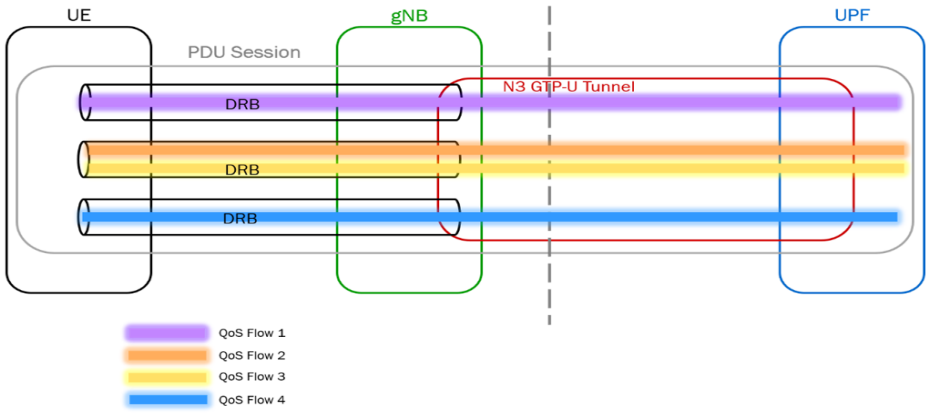


Figure 3. 11: QoS Differentiation within a PDU session [42].

A PDU session is made of a tunnel and data radio bearers. Inside the tunnel are various QoS flows having specific QFI values that are mapped to data radio bearers performed by the SDAP layer.

To establish a PDU session in the 5G network, a PDU session establishment message is sent from the 5G CN to the gNB that is serving that particular UE after the UE has made a request to the AMF via the NAS signalling message containing the QoS information to be set up. The gNB then sends a DRB setup request to the UE including the SRB and the NAS messages it received in response from the CN. The CN then creates a QFI which is used for mapping the QoS flows to the DRB and sends an RRC complete message containing DRB complete setup response to the gNB. The gNB then sends a PDU establishment acknowledgment to the 5G CN which indicates a successful establishment of the PDU session. Data is then sent from the PDN to the gNB via a tunnel with a particular TEID for both Uplink (UL) and Downlink (DL) data over the DRB encapsulated as tunnels to the UE. The tunnel runs from the UPF to the gNB while DRBs runs from gNB to the UE. Also, each radio bearer has an identification tag (DRB ID) which is used for mapping the DRBs to the QoS flows for UL transmission.

3.2.3 UP Data Flow in Layer 2

In this section, the UP data flow transmission from the gNB-UP to the UE is explained. Specifically, it is explained how IP packets are treated by the various sublayers of Layer 2. Recall that Layer 2 of the OSI model refers to the SDAP, PDCP, RLC and MAC sublayers. A Service Data Unit (SDU) refers to packets

received by a sublayer while Protocol Data Unit (PDU) refers to packets sent to sublayers or an output of a sublayer. The SDU given as input to Layer 2 is an IP packet consisting of an IP header and an IP payload. Then, the IP packet will be encapsulated in frames by the Layer 2 sublayers. Finally, the PDU/frames are transported as bits in the PHY sublayer to the UE via DRBs. Figure 3.12 shows an IP packet.

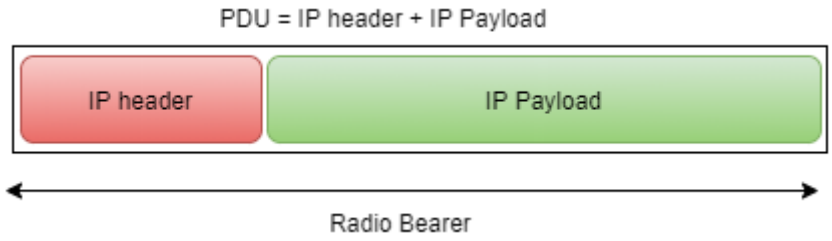


Figure 3. 12: Radio Bearer encapsulating PDU.

The IP header contains information about the processing of the packet while the IP Payload is the data being transmitted.

The UP data flow is shown in Figure 3.13 which shows three sections, labelled as (a), (b), (c) and (d) addressing the processing performed by each sublayer.

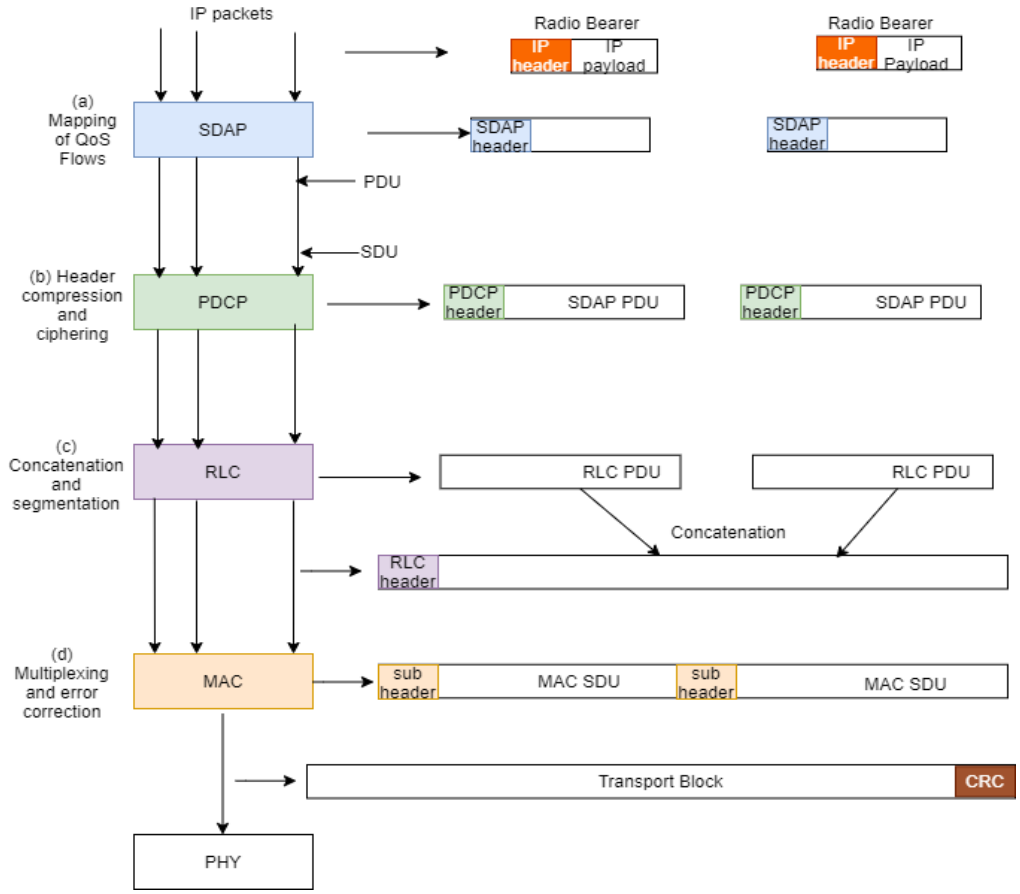


Figure 3. 13: Data Flow in RAN Protocol Layers.

IP packets coming from the UPF must pass through each layer. Each layer performs its own functions, adds a header to the packet and transmits it to the next layer and so on.

(a) This section shows IP packets coming from the UPF into the SDAP layer which performs mapping of QoS flows Recall the reason the SDAP layer was introduced in the 5G network is to map the User plane tunnels to radio bearers to improve the QoS of IP packets and also adds its own header to the IP packets.
(b) PDCP layer which major function is header compression and ciphering. It is noted that these packets come one by one in sequence and not simultaneously. The PDCP layer adds a PDCP header to each incoming IP packet (also referred as PDCP SDUs) to from the RLC PDU. The RLC PDUs are sent to the RLC sublayer one after the other in sequence.
(c) The RLC sublayer is responsible to concatenate multiple RLC SDUs into a frame and then add the RLC header which contains information, such as the frame length, that is required to split the frame back to the original frames. The frame after concatenation together with the RLC header, form the MAC SDU as the data are forwarder to the MAC sublayer. On the uplink side, RLC sublayer performs data segmentation.
(d) The main functions of the MAC sublayer are multiplexing and error correction. The MAC layer adds the MAC header, Cyclic Redundant Check (CRC) for error correction and padding bits to ensure that the frame fits into a subframe. It then transmits the frame to the PHY layer.

The PHY layer is responsible for the maximum data size that can be transported on the MAC sublayer and is defined by the Transport Block Size (TBS). The PHY layer transmits the data into slots of a subframe as bits.

For the uplink data transmission, the reverse process occurs in the gNB. The MAC layer receives a frame containing smaller IP packets and does error correction based on the CRC received, stripes off the MAC header and then transports them to the RLC sublayer. The RLC sublayer segments the huge frame into multiple SDUs based on the header information. The segmentation process is followed by the removal of RLC header. The SDUs is transported to the PDCP layer one at a time in sequence and each IP packet is stripped of the PDCP header exposing each IP payload.

3.3 Control and User Plane Separation

The concept of CUPS was first introduced in the EPC to decouple the CP from the UP [43]. Figure 3.14 shows an overview of the CUPS for the 5G RAN. The left-hand side of the Figure shows the CP while the right-hand side shows the UP. The Application protocol (AP) generates control signal and the CP encapsulates the signal bearers used to transmit control signals from the gNB to the UE. The UP provides the data bearers used to carry UE data from the DN to the UE using the GTP-U. The principle of CUPS supports the SDN technology which is already adopted in the 5GCN utilizing a service-based architecture.

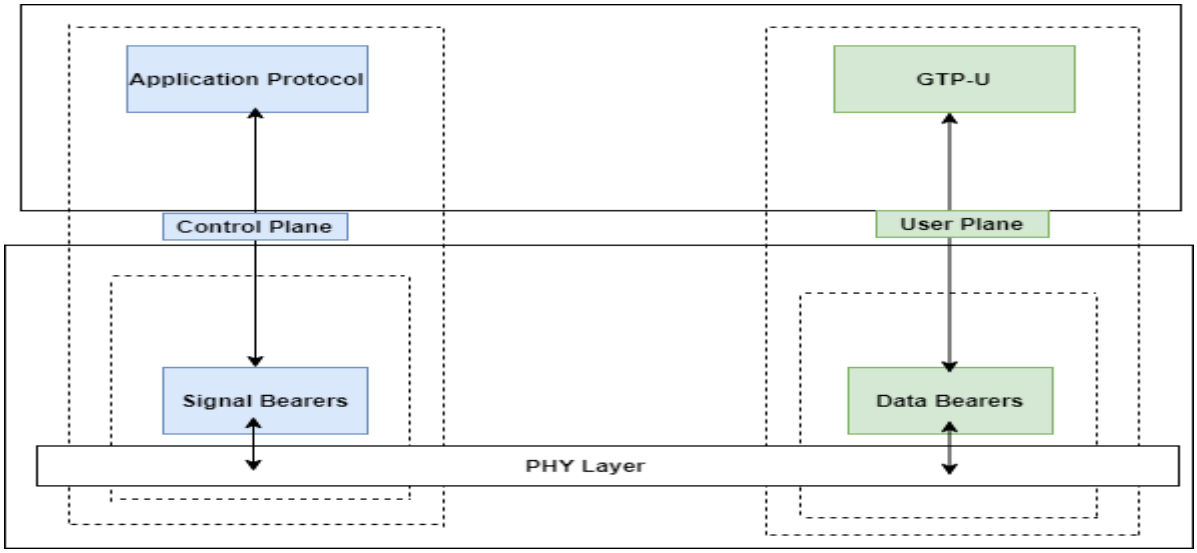


Figure 3. 14: Overview of CUPS [44].

The Application protocol is a CP protocol that transmit control signals via signal bearers while GTP-U is a UP protocol that transmits UE data via data bearers.

The advantages of CUPS can be summarized as follows:

- Resource Scaling: CUPS allows the scaling of UP and CP resources independently; separating the CP and UP allows for scaling of each planes resources since both CP and UP are no longer tightly coupled which results in scaling of both resources whenever a given plane is scaled.
- Enables efficient delivery of the UP data and CP signals.
- Allows for independent evolution of each plane.
- CUPS ensures that individual services are tailored or adapted to suit the scenario which makes data transmission more efficient. In other words, the principle of CUPS enables the concept of network slicing. For instance, consider MTC applications such as smart meters; MTC traffic has typically no mobility and in addition, it has low payload which have different traffic characteristics than traffic from a smart phone or other eMBB related services [12]. These various services (MTC and eMBB) can be running via different slices in the network made possible by CUPS.

The major disadvantages of CUPS is that it introduces complexity, standardization of interfaces between CP and UP and high cost of operation which may result in deployment delays [2].

Figure 3.15 shows the EPS without CUPS where the C/U indicates that both the CP and UP are tightly coupled in the S-Gw and in the P-Gw.

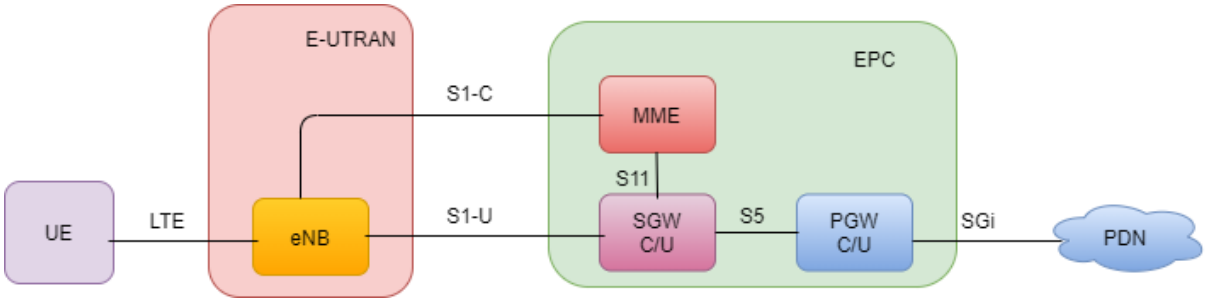


Figure 3. 15: EPS without CUPS.

The CP and UP entities are residing in the SGw and PGw respectively before the introduction of CUPS principle in the EPC.

The function of the P-Gw is to connect to the PDN or server via the SGi interface. The P-Gw entity also serves as the anchor point for the issuance of IP address for a given data session. The S-Gw acts as a router and forwards data between the PDN gateway and the eNB. The MME controls the high-level control and signalling between the UE and the eNB such as mobility management, session management, registration and authentication of UE etc. It should be noted that the PDN is the operator's server or data network.

CUPS involves the categorization and splitting of network entities such as S-Gw and P-Gw as being either part of CP or UP based on functional decomposition [13]. In 3GPP Release 14, EPS with CUPS was introduced which enables flexible deployment of Gateways (GWs) closer to the RAN in order to reduce latency. Figure 3.16 shows EPS with CUPS where a dedicated bearer travel through the GW-UP (more bearers are also possible) to convey user data with SGW-CP and PGW-CP supporting all the control signalling for the data transfer session.

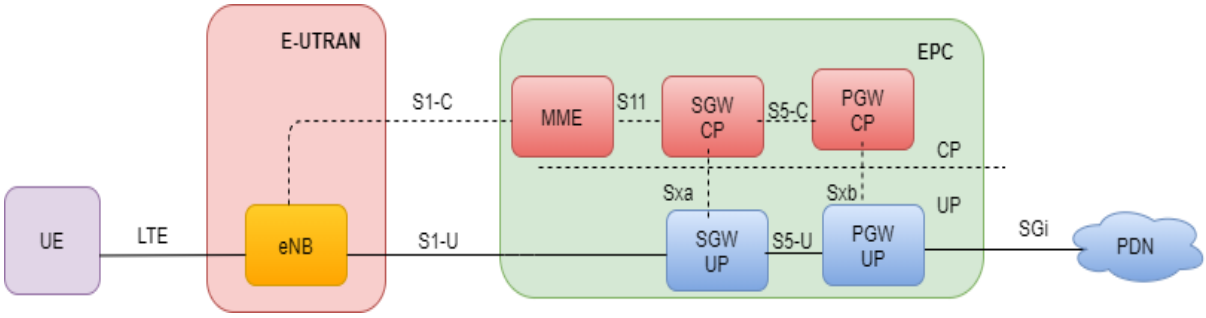


Figure 3. 16: EPS with CUPS.

After the introduction of CUPS principle in the EPC. The GWs (SGw and PGw) are now decoupled into GW-C and GW-U for CP and UP functionalities.

As user traffic increases, the independent evolution of UP and CP saves cost since only one CP is maintained while multiple UPs are established based on the user request. Figure 3.17 shows the high-level split of the gNB to allow CUPS in the 5G RAN.

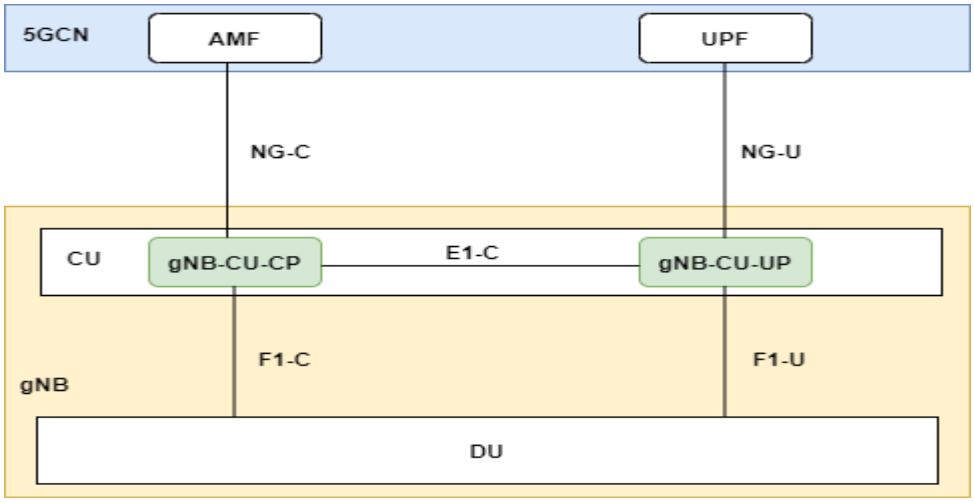


Figure 3. 17: High level Split of gNB showing Interfaces.

The E1-C interface interconnects the gNB-CU-CP and gNB-CU-UP and both entities are connected to the AMF and the UPF respectively.

A gNB may consist of a gNB-CU and multiple gNB-DUs (in C-RAN deployment). It is also possible to deploy multiple gNB-CU to increase the network resilience [30]. A gNB-CU is further split into gNB-CU-CP and gNB-CU-UP. The gNB-CU is connected to gNB-DU through an F1 interface. The F1 interface allows signalling exchange and data transmission between two end points specifically the CU and the DU. The F1 interface [45] is divided into F1-C and F1-U interfaces. The connection that exist between the gNB-DU and the gNB-CU-UP is established using UE context management function. One DU can be connected to multiple gNB-CU-UPs under the control of one gNB-CU-CP and also one gNB-CU-UP can be connected to multiple DUs under the control of one gNB-CU-CP.

The general functions of the F1-C interface are:

- System information management: The responsibility of the gNB-DU is for scheduling and broadcasting of system information [57].
- F1 interface management functions: The main functions consist of the F1 setup, gNB configuration update, reset function and error indication function [57].
- RRC message transfer function: This function is responsible mostly for the transfer of RRC messages from the gNB-CU to gNB-DU and vice versa [57].
- F1 UE context management functions: These functions are mostly performed for the establishment and modification of UE context. The gNB-CU initiates the establishment of F1 UE context which the gNB-DU can accept or reject depending on the admission control criteria. For example, the gNB-DU can reject a context setup when there are not available resources. The F1 context management function may also be used to create, modify and release the SRBs and DRBs.

The general Functions of the F1-U interface:

- Transfer of user data: This function enables the transfer of user data between gNB-CU and gNB-DU.

- Flow control function: This function enables the control of the downlink user data transmission towards the gNB-DU. Functionalities like PDCP PDU retransmission are introduced to improve data performance in case of loss due to radio link outage.

The Interface between the gNB-CU-UP and the gNB-CU-CP is called the E1-C interface and it is strictly a control plane interface.

The general Functions of the E1-C [45] Interface includes:

- It supports the exchange of signalling between the two end points.
- It is a point-to point interface between a gNB-CU-CP and gNB-CU-UP form a logical stand point.
- E1 interface is a control interface and it is not used for data forwarding

3.4 IP Multimedia Subsystem

IP Multimedia Subsystem (IMS) is a telecommunication system which controls multimedia services accessing different networks [46]. There are three functional domains identified in the IMS network: The Access Networks (AN), the Packet CN and the multimedia communication framework. The multimedia domain is responsible for multi identity, name calling, multi-user and a flexible UP. In the IMS network, there is strict separation between the functional domains. Figure 3.18 shows the overview of the IMS network which involves connecting different user devices via different ANs to the Packet CN and then to the IMS network. The interconnect indicates that the IMS network can be connected to other IMS network especially when two subscribers are connected to different IMS network.

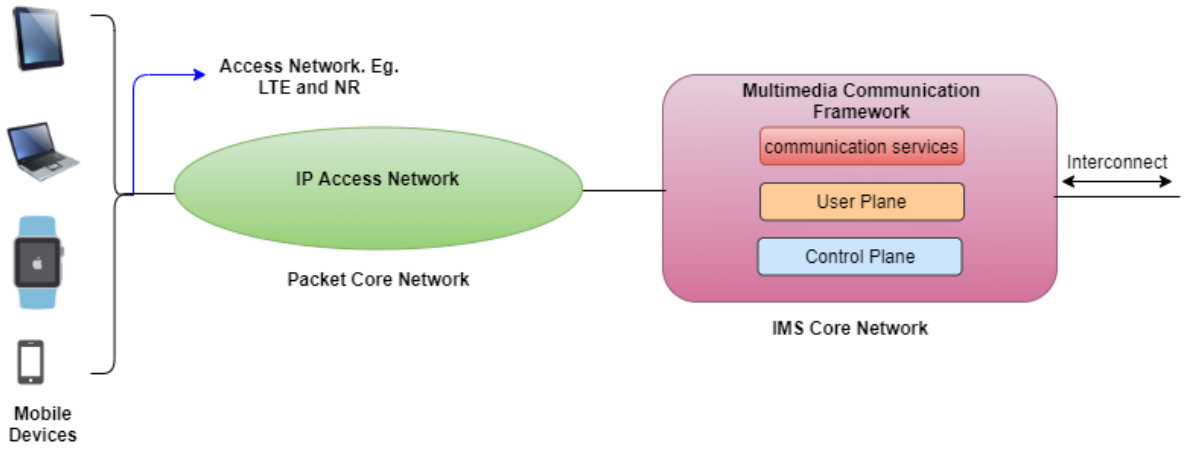


Figure 3. 18: Overview of IMS Network.

Various devices need access to a packet core network in order to connect to the IMS network. Without the IMS network, VoLTE and VoNR cannot be performed since these calls are performed as packet switching unlike in 2G or 3G where call establishment is done via circuit switching.

The IP AN can be UMTS, LTE or NR while the IP packet core network can be EPC for 4G networks or 5GCN for 5G networks. The multimedia communication framework is made of IMS communication services which are functionally decoupled into the CP and UP respectively. The whole multimedia

communication framework represents the IMS network. The IMS network is strictly based on IP infrastructure, hence, it can be mapped to both mobile and fixed-line networks [46]. IMS network allows for IP-based real-time services such as voice, video telephony, machine-to-machine communication and gaming. By principle, IMS is independent of the AN and so IP multimedia communication becomes a network on itself. Generally, a subscriber will register and attach to the CN of a given AN and then also to the IMS communication network. The key fundamental protocols used in the IMS architecture include the SIP and the RTP, which will be discussed in subsequent subsections.

3.4.1 Voice Over NR

Similar to LTE where voice over the LTE is called VoLTE [47], voice/video over 5G is called VoNR. VoNR provides a better user experience than the VoLTE because the latter uses the H.265 codec [48]. A codec is an encoding and decoding scheme for digital signals to be used by both the UE and the IMS network. Furthermore, compared to VoLTE, the bandwidth of VoNR is increased to improve the bit rate. In addition to this, the service robustness is improved and thus a better user experience. Additionally, in VoNR the voice/video call is connected within a period of one to two seconds from dialling to hearing the ringtone back [48]. In other words, it takes a period of 1-2 seconds to hear the ringtone from the moment of dialling which is almost twice as low as for VoLTE. As the call is in progress, subscribers can still enjoy high-speed internet access via the 5G network.

3.4.2 The Real-Time Transport Protocol

Real-time Transport Protocol (RTP) [49] is the protocol running on the UP for the transmission of media services in the IMS. The RTP encapsulates chunks of Pulse Code Modulated (PCM) data streams into RTP messages and transports them over a UDP/IP connection towards the subscribed user. RTP is an essential component of real-time communication in IMS because it provides the structure for reliable real-time transmissions over an IP infrastructure. In Figure 3.19 the RTP encapsulation and decapsulation is shown.

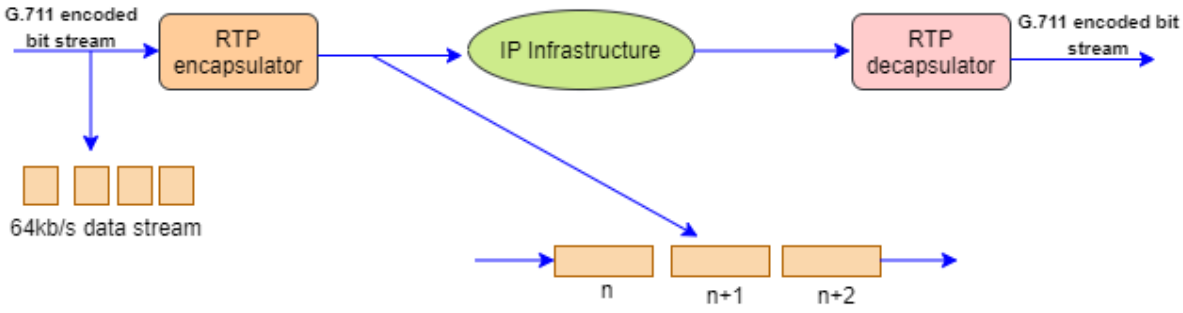


Figure 3. 19: RTP encapsulation and decapsulation [49].

The voice packets are transmitted in form of RTP packets. RTP runs on top of UDP which does not guarantee the delivery of packet. The reason is to avoid retransmission which distorts the quality of voice signals transmitted.

RTP encapsulation refers to the process of stacking a defined number of PCM words into RTP messages shown as. The G.711 encoded bit stream is the encoded voice input carrying a 64kb/s data stream to the RTP encapsulator. Furthermore, all RTP messages are numbered where n refers to the number given to the RTP message; since RTP runs on UDP, which does not provide any delivery guarantees, the numbering of packets becomes imperative to monitor the received RTP messages as they may arrive at the destination out of order or some of them might got lost. Therefore, numbering the RTP messages helps the receiver to re-arrange them and identify lost RTP messages. It is worth mentioning that the main reason why RTP runs over UDP and not TCP is because TCP can introduce high transmission packet delays due to its nature to deliver packets in a dependable way (flow control mechanism). A high transmission packet delay can result, in this case, to a skewed speech quality.

The Real-time Transport Control Protocol (RTCP) [49] is the control protocol for the RTP. Whenever an RTP media session is established between two hosts, an RTCP session will also be initiated as well. The RTCP control session is carried in the UP since RTP is the accompanied media session. The RTCP is used between two hosts, that usually want to begin a communication session, to negotiate the quality parameters of the media transfer such as jitter, packet count, percentage packet loss and transmission latency. Once these messages are successfully exchanged, the media can be transported via the UP under the control of RTCP.

3.4.3 Session Initiation Protocol

The Session Initiation Protocol (SIP) [49] is a CP protocol used in the IMS network for registration, session management/establishment, message routing and so on. SIP is a transaction-based protocol. A SIP transaction involves a request and a response transaction by the two User Agents (UA) intending to communicate. A communication session is initiated by an INVITE transaction and terminated by a BYE-transactions. Figure 3.20 shows two UAs A and B intending to start a communication session.

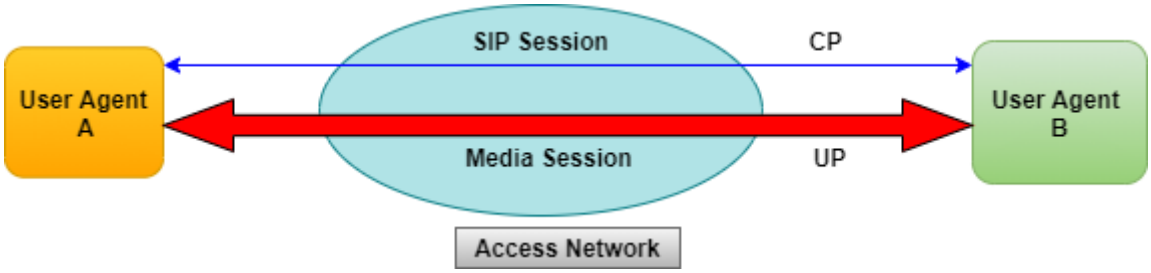


Figure 3. 20: SIP Session and Media Session [49].

For two users to communicate, SIP session and Media session must be established. The SIP session carries the SIP messages while the Media session carries the UE voice/video traffic

The SIP session runs via the underlying CP while the media session is carried in the UP. SIP uses the Hyper Text Transport Protocol (HTTP) which is human readable. The major function of the SIP is to facilitate the transmission of a media session between two UAs. The end points communicating via SIP messages are

called SIP UAs. A SIP session starts by the originating UA sending a SIP request message to the other UA who may either accept or reject this request by sending SIP acceptance or rejection message to the initiator UA of the SIP request message. The SIP communication between the two UAs is not direct, but runs via different SIP proxies residing in the signalling path of the SIP session. The SIP proxies play a vital role in establishing a SIP session, for example by routing the SIP messages from the initiator to the destination.

3.4.4 The IMS Core Network

Recall that in principle the IMS network is independent of the AN which facilitates the IMS network to be accessed via various IP Carrier Access Network (IP-CAN). The IMS network applies CP signalling and UP signalling towards the IMS CN through the IP-CAN. In general, the path taken by the media is usually as short and the media shall be transported end-to-end without the use of media proxies [49].The IMS core network is comprised of various entities called Call State Control Functions (CSCF) namely: Serving-CSCF (S-CSCF), Proxy-CSCF (P-CSCF) and Interrogating-CSCF (I-CSCF) as shown in Figure 3.21.

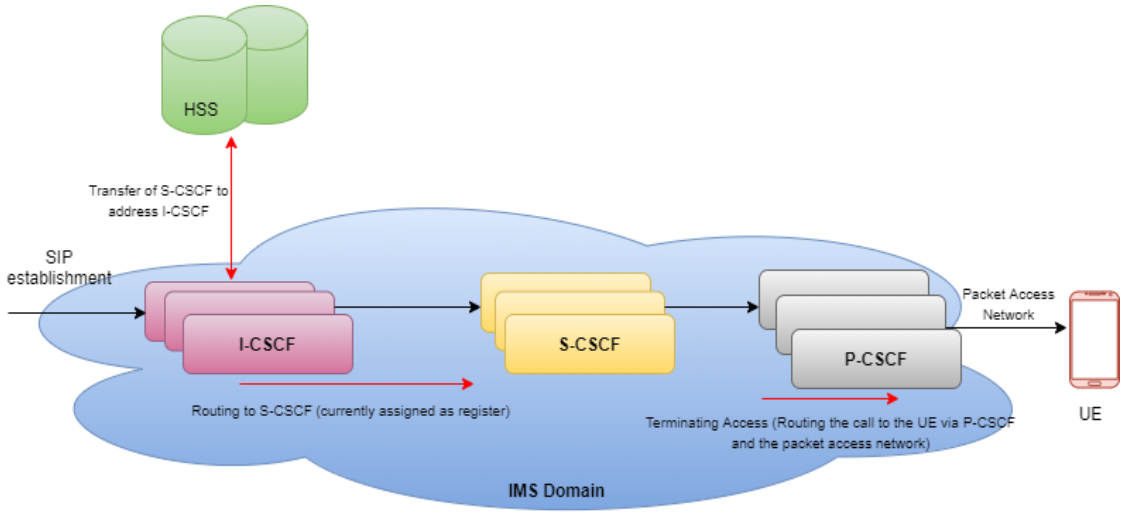


Figure 3. 21: IMS Core Network [49].

Once the SIP messages are transmitted to the IMS core network, the I-CSCF routes the messages to the S-CSCF which then routes the call to the P-CSCF. The P-CSCF the routes the call to the called UE.

These CSCF entities are usually co-located which does not structurally distort the architecture in the network. Multiple functional entities of the same kind can be deployed in an IMS network for various reasons such as redundancy, geographical distribution and capacity. The CSCFs mentioned communicate or 'talk' SIP to each CSCF as well as to the border gateways. When communication outside the IMS network is needed, for example when one IMS network needs an additional service of another IMS network in order to provide the agreed services, then the traffic is routed via an Interconnect Border Control Function (IBCF) [50]. The main core network entities are described as follows:

The Serving-CSCF: This is the main SIP control node in the IMS network. When a subscriber registers in the IMS network, the subscriber's details are registered in the S-CSCF and can provide their contact details including their IP address to the I-CSCF. In other words, the S-CSCF functions as the register of an IMS network which is analogous to the AMF of the 5GCN. There can be multiple S-CSCFs in the IMS network and one S-CSCF needs to be selected for a particular UE. Once a given S-CSCF is selected, the subscriber address is stored in the Home Subscriber Server (HSS) which is a data base for storage of user profile. In general, the S-CSCF acts as a SIP proxy for the establishment of SIP session for authentication.

The Proxy-CSCF: All SIP signalling and media transmissions run through the P-CSCF. The P-CSCF can reside either in the home or visited IMS network.

The Interrogating-CSCF: This entity is used for forwarding SIP requests to the S-CSCF when the sender of the SIP request does not know which S-CSCF to use. In such a case, the I-CSCF connects to the HSS to obtain the address of the S-CSCF to receive and process the SIP request. In general, an incoming SIP request or message coming from another IMS network will first arrive to the IBCF from where it is then be forwarded to an I-CSCF.

Other entities in the IMS network are described below:

A session Border Gateway (SBG): This is the CP entity in the Access Border Gateway (A-SBG) when the IMS network is connected to a wireline AN. SBG is a SIP entity for the UE to access the IMS network and it can adjust and conform the SIP headers to suite the IMS network operators' requirements. Other additional functions of the SBG includes topology hiding, access security, SIP firewall and so on.

IMS Access Gateway (AGW): This entity is responsible for ensuring that the address associated with incoming and outgoing media streams (traffic) are correct. Usually, AGW is incorporated into the functionality of the IBCF.

Interconnect Border Control Function (IBCF): This entity acts as a boundary control for different network service providers by providing IMS network security. Its functions also include topology hiding and session screening based on source and destination signalling address.

Multimedia Resource Function (MRF): This entity is a powerful multimedia service that gives access to applications on the IMS network.

The Session Gateway (SG): This is the user plane entity in the A-SBG. It is the entity to which the media transmitted is steered to the UE from the IMS network and which may be subject to bandwidth management. The RTP media stream carried in the SG may contain confidential information and under the control of the SBG can remove certain confidential information subjected to the operator's policy. The data is then transmitted via an IP AN which could be 4G or 5G depending on the RAT deployed.

3.4.5 IMS and the 5GCN

Figure 3.22 shows the architecture of the IMS network when connected to the 5GCN.The green line in Figure 3.22 shows the UP data path and the orange line indicates the SIP signalling path or the CP of the network. In this architecture, the entire IMS network is in a confined network [32] which means that the same network provider is providing both the AN and IMS network. The IMS AGW as already mentioned is used to carry media traffic and the HSS is located outside the local network. To establish a PDU session for an IMS service, the SMF shall select the intended UPF based on the subscription information or network configuration settings applied by the network operator. Based on the selected UPF, the local P-CSCF is located through the I-CSCF which allows IP connectivity to the UPF. The S-CSCF then uses the IMS Application Servers (AS) [50] to route the traffic for the intended media flow to a specified UPF. IMS AS is used to host various IMS applications for end user services such as gaming, content sharing,

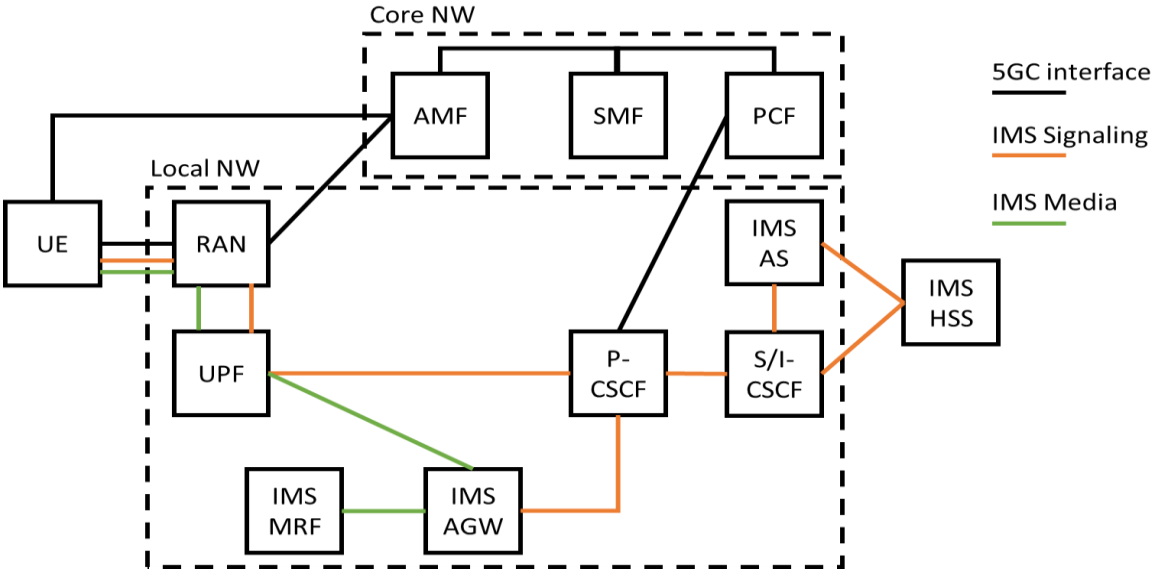


Figure 3. 22: Architecture of IMS Network Connected to 5GCN [32].

The IMS MRF is a multimedia service that provides access to IMS applications. The data is then routed to the IMS Access Gateway providing access to the 5G network. The data is then transmitted to the UE via the UPF as represented by the green line.

videoconferencing and so on. The PCF interconnects with the P-CSCF to influence the data for charging and QoS adaptation. The IMS signalling represents the CP while the IMS media represents the UP. Data is forwarded form the IMS AGW to the UPF via the RAN for example a gNB to the UE. Hence, the IMS data is judiciously carried through the UP from the MRF to AGW to the UPF of the 5CN down to the UE.

CHAPTER 4

THE ARCHITECTURE OF 5G NR WITH CUPS (NR-NR ARCHITECTURE)

This chapter presents an enhanced architecture of the 5G NR which decouples the CP signals and the UP data while using two gNBs in DC mode by introducing the principle of CUPS in the 5G RAN. The chapter is divided into six sections where Section 4.1 describes the architecture and signalling procedures in the NSA while Section 4.2 presents the design principles of the NR-NR architecture. Section 4.3 discusses the CP bearer configuration and the fourth section discusses UP bearer configuration. The fifth section describes the PDU session establishment procedure and detailed signalling procedures in the NR-NR architecture. Finally, the sixth section describes the UE mobility as it moves from one cell to another.

4.1 The Architecture of the Non-Standalone Architecture

EN-DC is a technology that allows the introduction of 5G services specifically the support of the higher 5G data rates in a predominantly 4G network. A UE supporting NSA can simultaneously connect to the LTE master node eNB (MN-eNB) and 5G Secondary Node gNB (SN-gNB) as shown in Figure 4.1.

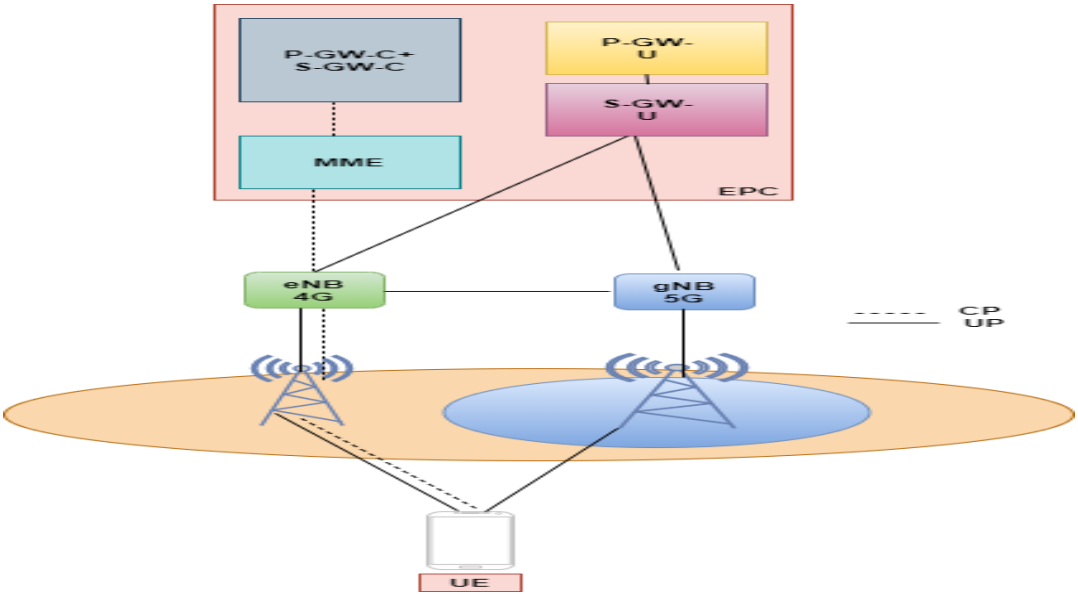


Figure 4. 1: Architecture of the NSA.

The UE is connected to both 4G network and 5G network simultaneously in DC mode. eNB handles both control signals and UP traffic while gNB only transmits UP traffic.

As shown in the figure, the eNB is connected to both MME and S-GW-U which operates both CP signals represented by dotted lines and UP data represented by solid lines. Besides, it is clearly seen that eNB and the gNB have a direct interface with the existing LTE EPC for UP functions (both are connected to S-GW-U) and only eNB has direct interface to the MME for CP signalling. The implication of a direct eNB interface to both MME and S-GW-U means that eNB handles both CP signalling and UP data while gNB is dedicated for carrying only user data. The MME is a CP entity in the EPC that transmits control messages inform of control signalling while S-GW-U is a UP entity in the EPC transmitting user data. In NSA, the UE first registers for 4G network service and EPC for data transmission. The EPC checks whether the UE is capable of connecting to the 4G network and 5G network simultaneously through the NAS signalling between the UE and MME during the initial attachment procedure. Then the UE also starts to perform measurements on the available 5G frequencies in order to start data transmission on 5G NR depending on whether the UE is enabled for 5G services. The LTE eNB communicates with the gNB to assign radio resources for a 5G data bearer establishment. The 5G radio resource allocation is then signalled to the UE and it is then simultaneously connected to both 4G and 5G networks.

Once the UE connects to the 4G LTE base station, a PDN connection is established between the eNB and the UE via the MME to establish a default bearer between the PGW-U and the eNB. From the UE's capability report which indicates its NR capabilities (ability to support 5G), an SN-gNB node addition (a message transmitted by eNB to add a gNB) is signalled in other to attach the UE to the gNB. Once the RRC reconfiguration complete message is received by the eNB from the UE, the eNB starts copying data to the gNB to ensure continuous media stream. In order to establish a data tunnel towards the gNB, the eNB triggers a path switch update procedure by sending an E-UTRAN Radio Access Bearer (E-RAB) modification indication to the EPC. The EPC, more specifically the S-GW-U, confirms this modification message through an end marker packet and then establishes a GTP PDU to gNB. The UE connects to 5G network and data flow is resumed over the 5G. The detailed signalling for the attachment procedure in NSA is shown in Figure 4.2.

4.2 The Design Principles of the NR-NR Architecture

The proposed NR-NR architecture is a typical example of a 5G standalone architecture that is made up of two gNBs. Recall that a standalone deployment means that the 5G network is made up of 5G NR providing NR functionalities to the UE and the 5GCN providing 5G CN capabilities. For HetNets, there are usually the small cells and the master cells with each cell operating on different carrier frequencies. The small cells are mainly used for UP data traffic. The master cells provide coverage and usually contribute insignificantly to the overall data transmission due to the fact they usually have low available bandwidth. For a master cell to ensure proper coverage, that is to avoid coverage gaps, the master cell needs to be of a lower frequency and the small cells should have high frequencies since they have short transmission range.

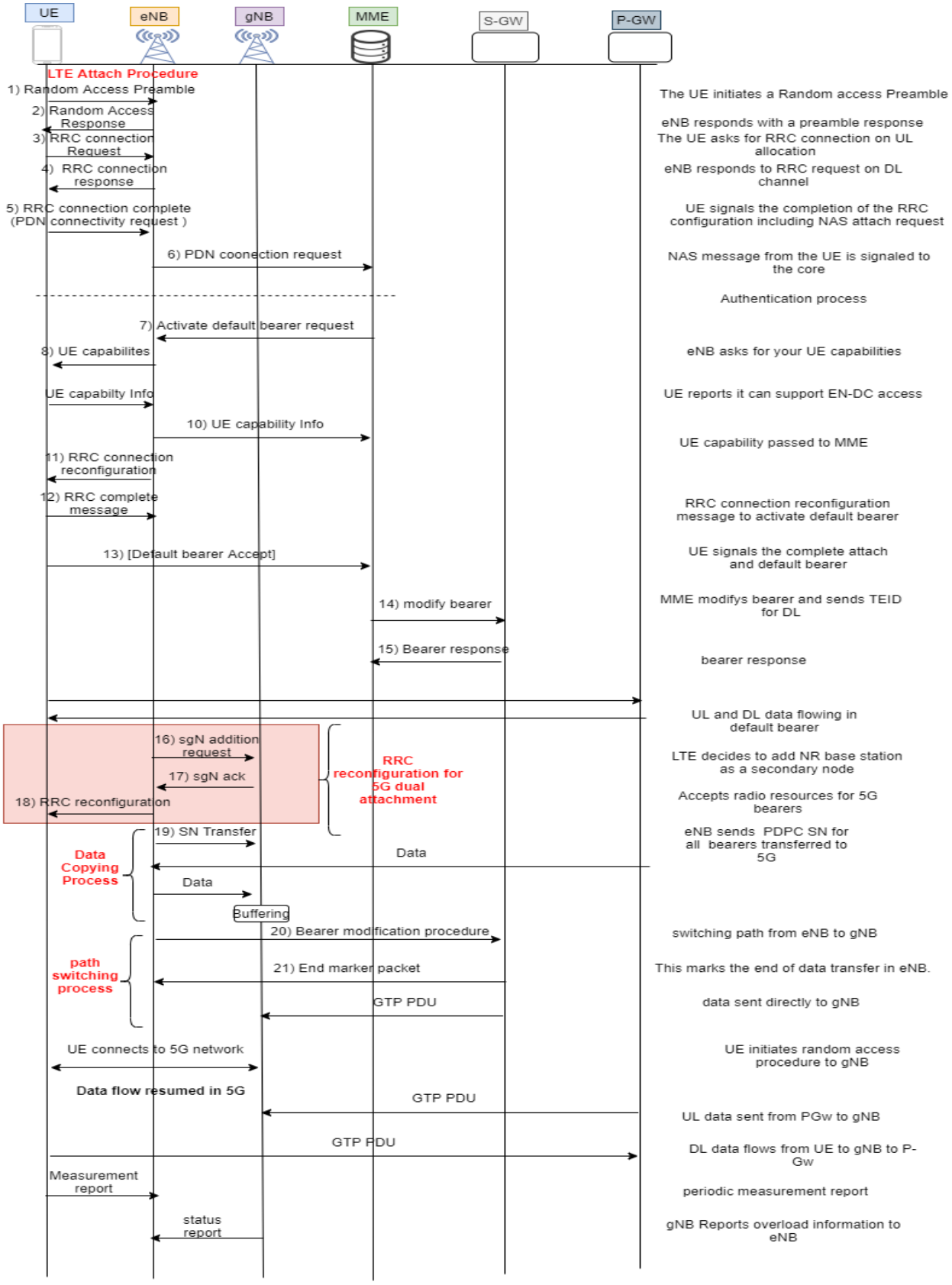


Figure 4. 2: Signalling in NSA [52].

The UE first connects to the 4G network and then to the 5G network for its data transmission to "enjoy" 5G data rates.

Decoupling the UP and CP allows for the scaling of the CP resources and thereby ensuring that small cells handle all UP data traffic. Designing the NR-NR architecture by introducing the CUPS principle does not only facilitate the benefits of CUPS as discussed in Chapter 3 but also minimizes the load handled by the master cells as all the UP transmissions are handled by small cells only by transferring all the UP transmissions to be performed by small cells only. This decoupling will facilitate an efficient scaling of each plane's resources as already emphasized. By designing the architecture in such a manner, it is also ensured that data is transmitted to the UE without a need for the UE to use radio resources in both cells. Secondly, the UE will only experience a handover whenever it leaves the coverage of the master cell. This is so because handover request message is only sent between two master cells. For movement of a UE within the coverage of a master cell, the "node addition message" is used to handover from a serving small cell to a target small cell. Considering the fact that the UE needs to be served simultaneously by the two gNBs as shown in Figure 4.3, the UE shall be configured in DC mode.

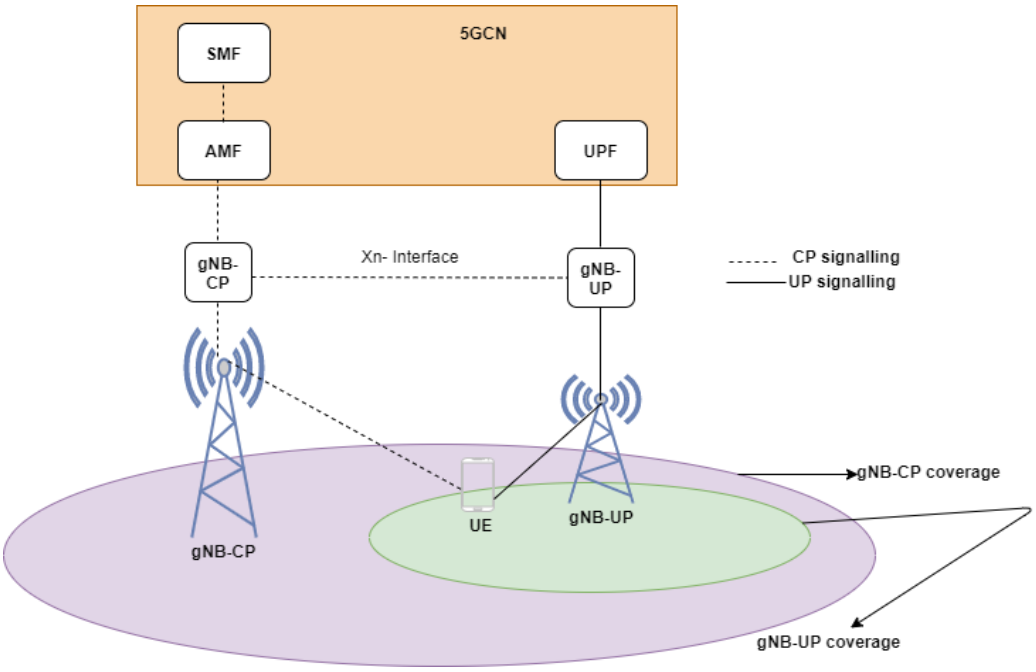


Figure 4. 3: CP/UP Split Overview of NR-NR Architecture.

UE is connected simultaneously to gNB-CP providing control signalling while gNB-UP only handles user data. The Xn interface interconnects the gNB-CP and gNB-UP.

The NR-NR architecture introduces a gNB-CP as a master cell to provide network coverage and acts as the main anchor point for transmitting only CP signalling. Also, the architecture introduces gNB-UP as a small cell dedicated in transmitting user plane data. By allowing gNB-UP to dedicatedly transmit user data enables CP/UP resources to be scaled as desired, and thereby it can increase the user throughput. Figure 4.4 shows the proposed architecture of the NR-NR standalone deployment.

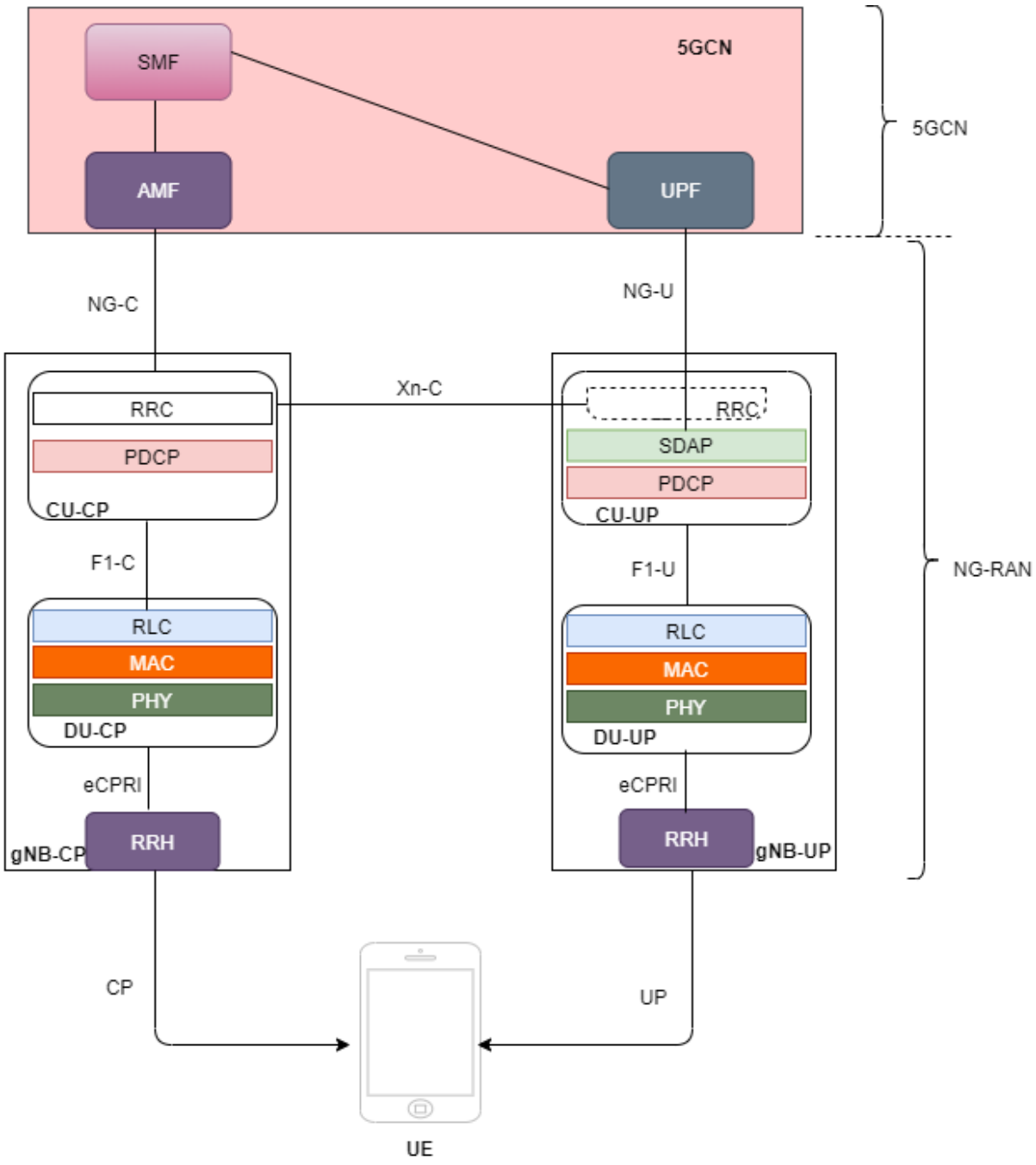


Figure 4. 4: The Architecture of the 5G NR-NR Deployment.

The gNB-CP is made of the CU-UP, DU-CP and a dedicated RRH. The CU-CP and DU-CP handles all control signalling. Signal radio bearers are established in the gNB-CP to carry control signals like an RRC connection message, and NAS messages.

All the network entities mentioned in Figure 4.4 are already discussed in the previous chapters. The F1-C interface, as well as the CU-CP and DU-CP network entities are all residing at the gNB-CP and use a dedicated RRH. Similarly, the F1-U interface, and the CU-UP and DU-UP network entities are residing at the gNB-UP and also have a dedicated RRH. The NG interface connects the RAN to the 5GCN. It is made up of NG-C and NG-U to transport control signalling and user data traffic respectively between gNB-CP/gNB-UP and 5GCN. As can be seen

from Figure 4.4, the gNB-UP has its own UP connection to the UPF. Thus, whenever a Data Radio Bearer (DRB) is configured to carry data for different applications, the corresponding data can only be transmitted to or from the gNB-UP. Considering that only the UP runs through the gNB-UP, there is a possibility that higher UE throughput can also be achieved by offloading gains. By offloading gains, it means that the UP data traffic can be offloaded to IEEE 802.11 (WIFI) to boost capacity and enhance network performance [53]. Also, there is a possibility that higher throughput can be achieved via Coordinated Multipoint Transmission/reception (CoMP). CoMP is a technique that is used to boost the throughput of cell edge users but it is only possible between two gNB-UPs because they are the only nodes that are involved in transmitting user data.

Figure 4.4 also shows the location of the various radio protocol stacks residing in the gNB entities. Specifically, for the gNB-CP, the RRC and PDCP-CP layers are located in the CU-CP while the other layers (RLC, MAC and PHY) are residing in the DU-CP. For the gNB-UP, the SDAP and PDCP-UP layers are residing in the CU-UP while the other layers (RLC, MAC and PHY) are residing in the DU-UP. The RRC layer of gNB-UP is depicted in doted lines to show that gNB-UP has its own pool of radio resources and allocates it accordingly when data transmission is needed. How RRC messages are transmitted between gNB-CP and gNB-UP is explained in detailed in section 4.3. Recall from Chapter 3 the reason why split option 2 is chosen for NR-NR architecture is to enable the introduction of CUPS principle in the CU. Figure 4.5 shows a view of the CU of gNB-CP and gNB-UP respectively.

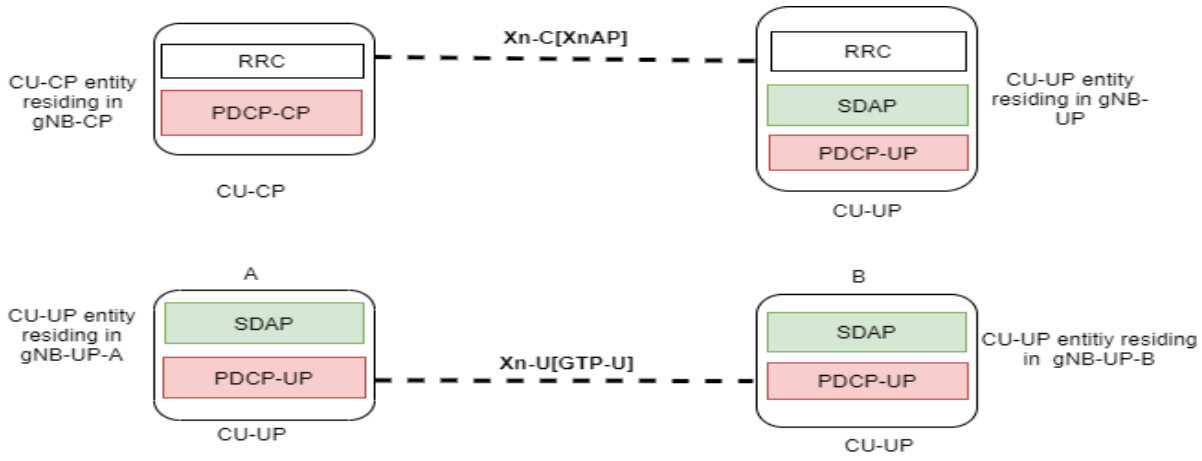


Figure 4. 5: The CU of the NR-NR Architecture.

The Xn-U interface is used to interconnect between two gNB-UPs.

The top part of Figure 4.5 shows the CU of gNB-CP and gNB-UP respectively. The interface between a gNB-CP and gNB-UP is the Xn-C interface which runs the XnAP protocol. The Xn-C interface terminates in the RRC layer of a gNB-UP and the interface is used for transmitting CP signals for activities like gNB-UP node addition message and transmission of handover messages. The bottom part of Figure 4.5 shows the CU of two gNB-UPs namely A and B. The interface between any two gNB-UPs is the Xn-U interface which runs the GTP-U protocol. The Xn-U interface

terminating the PDCP-UP layers of gNB-UP-B is used for copying data from gNB-UP-A to gNB-UP-B during handover procedures or during UE's mobility within the coverage of gNB-CP. It is important to mention that in the NR-NR architecture, Xn-C interface only exist between a gNB-CP node and a gNB-UP node while the Xn-U interface only exist between two gNB-UPs. The reason being that gNB-CP no longer has a UP part; all UP functions are now residing only in gNB-UP. In other words, the interface between gNB-CP node and gNB-UP is the Xn-C interface and between two gNBs is the Xn-U interface. The signal and data radio bearer configurations in gNB-CP and gNB-UP are described in section 4.3 and 4.4 respectively.

The major difference between the NR-NR architecture and the NR architecture described in Chapter 3 is as follows. In the NR architecture, the CU is separated into CU-CP and CU-UP entities where the CU-CP entity hosts the RRC and CP part of the PDPC layer and the CU-UP entity hosts the UP part of the PDCP layer and the SDAP protocol respectively. The CU-CP and CU-UP are interfaced using the E1 interface as they are residing in one gNB in order to allow both control signalling and UP traffic as already discussed in Chapter 3. In the NR-NR architecture, in order to ensure that gNB-CP strictly handles control signalling and is not involved in handling UP data, the CU-UP entity is residing only in gNB-UP and is strictly dedicated for UP data traffic. In other words, in NR-NR architecture, the gNB-CP has no UP part and is now migrated to gNB-UP. The difference is visualised on a high-level as shown in Figure 4.6 where on the right-hand side is NR-NR Architecture and on the left-hand side is the NR Architecture.

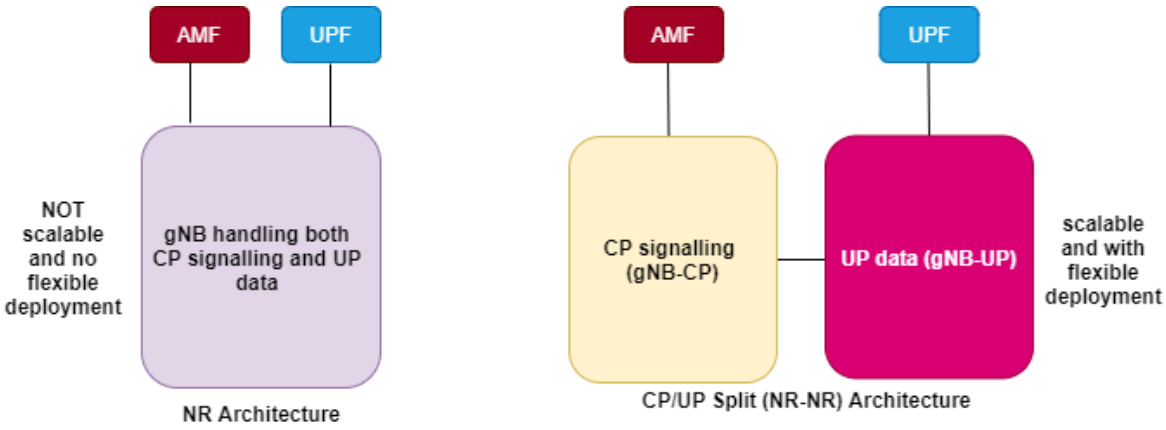


Figure 4. 6: Difference between NR-NR Architecture and NR Architecture.

A single gNB connected to both the AMF and UPF in the NR architecture while gNB-CP and gNB-UP connects to the AMF and UPF respectively in Dual Connectivity mode.

4.2.1 Full Protocol Interface in NR-NR Architecture

The detailed protocols running in the various reference points in the NR-NR architecture are shown in Figure 4.7. A brief description of the protocols is highlighted in Table 4.1.

Table 4. 1: NR-NR Protocols

Protocol/ Signals	Description
GTP-C	This is a control protocol for tunnel management such as creation, modification and deletion of tunnels per user for the CP. It carries control and signalling messages used to set-up GTP-U tunnels.
GTP-U	This is used to carry encapsulated user payload. They are identified by TEID which is visible in the GTP header.
NGAP	Protocol used for NAS messages, PDU session management and mobility procedures.
XnAP	Control protocol used between gNB-CP and gNB-UP to support various RAN procedures such as DC, SN transfer, RRC messages and so on.
NR[CP]	Signalling connection for carrying CP signals represented by SRBs.
NR[UP]	UP logical connection represented by DRBs.

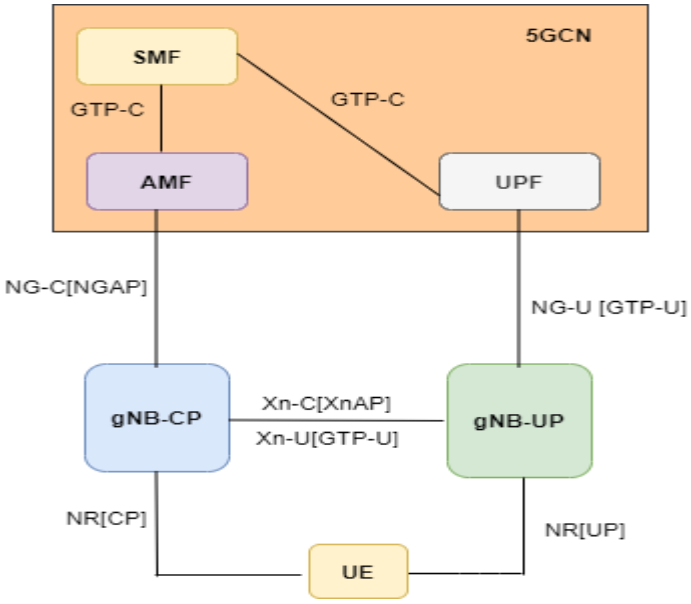


Figure 4. 7: Protocols running on NR-NR Architecture.

4.2.2 UE Cell Selection Procedure

In the NR-NR architecture, the UE shall select the strongest received Synchronization Signal Reference Signal Received Power (SS-RSRP) for gNB-CP and gNB-UP respectively. Then the UE selects the cells with the strongest SS-RSRP for both gNB-CP and gNB-UP.

4.2.3 UE Initial Attachment Procedure

The UE shall attach to the gNB-CP and also to the gNB-UP under the control of AMF when a PDU session is established. The gNB-CP performs UE registration and all authentication processes with the AMF via NAS signalling. Then a PDU session with default bearer is established by the UE and gNB-UP. The default bearer established shall have 5G QoS Indicator (5QI) value of 8 in anticipation for UE's data and another default bearer shall be established with 5QI value of 5 per PDU session in anticipation of the UE make an IMS Call with the IMS network. The default bearer with 5QI value of 5 is to be used to carry SIP signalling and shall be maintained as long as the UE is registered in the IMS network.

4.3 The gNB-CP Architecture Bearer Configuration

The function of gNB-CP is to specifically transmit control signals such as RRC messages and NAS signalling. The gNB-UP is also involved in establishing PDU connectivity between UE, gNB-UP, UPF and the AMF. Therefore, gNB-CP is responsible for generating, sending and maintaining all RRC connections to the UE for CP. In the same manner, the UE receives and responds to all RRC messages to/from via gNB-CP. This architecture does not support the transmission of RRC messages via the gNB-UP but rather through gNB-CP which relays the RRC messages of gNB-UP to the UE. In the NR-NR architecture, a gNB-UP possess an RRC entity because it has its own radio resources and are not shared with gNB-CP. This layer was utilized in this architecture in order to allow gNB-UP transmit its RRC messages to the UEs, since gNB-UP is responsible for allocating its own radio resources to the connected UEs. Having said that, the gNB-UP does the transmission of RRC messages indirectly via the Xn-C interface to the gNB-CP and then to the UE. Hence, whenever gNB-UP wants to modify or release its own part of the RRC configuration, it sends the related RRC message to the gNB-CP via the Xn-C interface. The gNB-CP then transmits the RRC messages it received from the gNB-UP to the UE via the air interface as shown in Figure 4.8.

There are three main types of radio bearers identified in the DC architecture namely the Master Cell Group (MCG) bearer, Secondary Cell Group (SCG) bearer and Split Bearer and they are already standardized for DC. In particular, for the NR-NR architecture only the MCG signalling radio bearer is supported in the gNB-CP while the MCG data bearer and MCG split bearers are not supported. This is because the gNB-UP only handles the UP traffic.

The MCG Bearer is a Signalling Radio Bearer (SRB) served only by the gNB-CP for the transfer of gNB-CP RRC messages and NAS signalling to the UE. This bearer can also be called SRB1or SRB2. The difference between SRB1 and SRB2 bearer is that SRB2 has a lower priority and is mostly configured by the network after security activation. Both SRB1 and SRB2 are usually transported via the Dedicated Control Channel (DCCH) in the air interface.

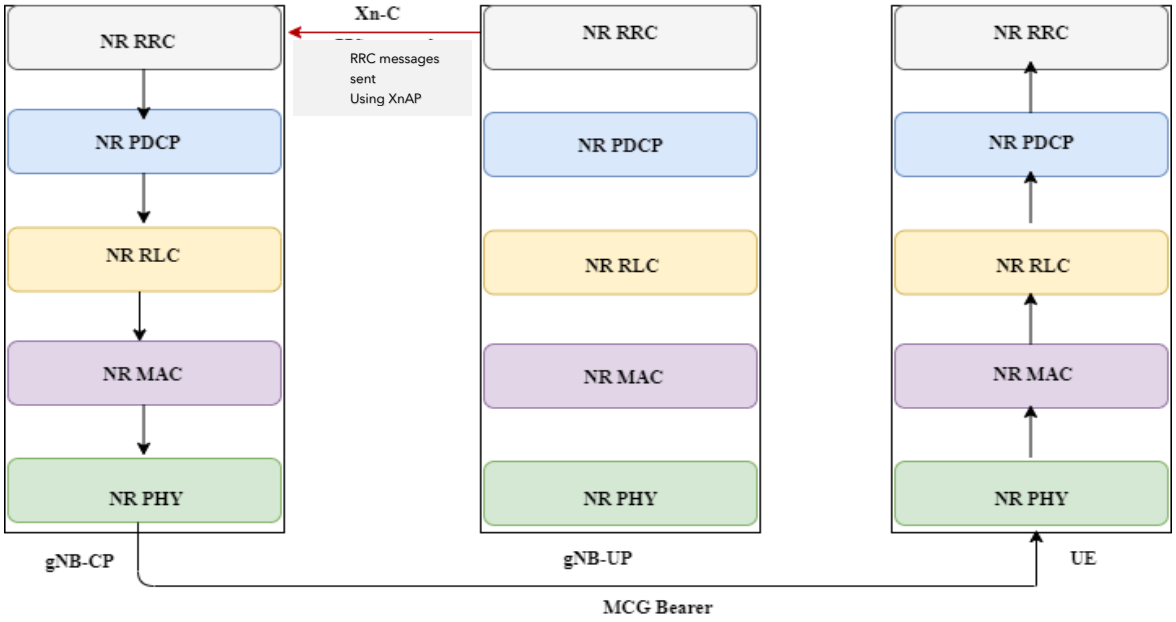


Figure 4. 8: MCG Bearer configured for NR-NR protocol stack.

Control signals are transmitted via the MGC bearer. The MCG bearer is only configured in gNB-CP.

The establishment of the SRBs for the CP is illustrated in Figure 4.8. The figure shows the protocol layers of the gNB-CP, the gNB-UP and the UE. The RRC layer of the gNB-UP generates RRC messages and forwards them via the Xn-C interface running XnAP to gNB-CP which relays the received RRC messages to the UE encapsulated in the MCG bearer. Hence, from a UE perspective, there is only one radio link via the MCG bearer transmitting RRC signals terminated at the gNB-CP. From the network perspective, there are two RRC radio links, one from the gNB-CP and the other form the gNB-UP.

4.4 The gNB-UP Architecture Bearer Configuration

The function of the gNB-UP is to transmit/receive the UP data to/from the UE. The NG-U interface which runs GTP-U carries the UP data between the gNB-UP and the UPF entity in the 5GCN.The user data are encapsulated and transmitted to the UE via the Data Radio Bearer (DRB) established in the gNB-UP. There are two main DRBs established in the NR-NR architecture, namely:

- **The SCG Bearer:** This DRB is used to transmit UP and it terminates at the gNB-UP. The UE data is then delivered from the gNB-UP to the UE via the air interface.
- **The Split Bearer**: This bearer is also used to transmit UP data. This bearer is configured for the purpose of copying data between two gNB-UPs within the coverage of a single gNB-CP especially during handover procedures to transmit data from a serving cell to a target cell. In this case, the data is forwarded via the PDCP-UP layer of the one gNB-UP to the RLC layer of the other gNB-UP via the Xn-U backhaul interface.

The flow of the DRBs is illustrated in Figure 4.9. The black arrows running through gNB-UP represent the SCG bearer running via the NG-U interface to the SDAP layer and then down to the lower layers. The SCG bearer only flows in the gNB-UP which is different compared to the NSA where both CP signalling and UP data are transmitted from/to the macro node (eNB). The blue arrows, in Figure 4.9, represent the split bearers. With the split bearers, the UP data is forwarded from the PDCP-UP layer of the one gNB-UP (e.g. the left-hand side gNB) to the RLC layer of the other gNB-UP (e.g. the right-hand side) via the Xn-U interface. It should be mentioned that Figure 4.9 illustrates the case when the UE is receiving data as a result of UE's movement within the gNB-CP coverage. On the right-hand side of Figure 4.9, the UE side is shown where the SCG and Split bearers are delivering the UP data to the UE.

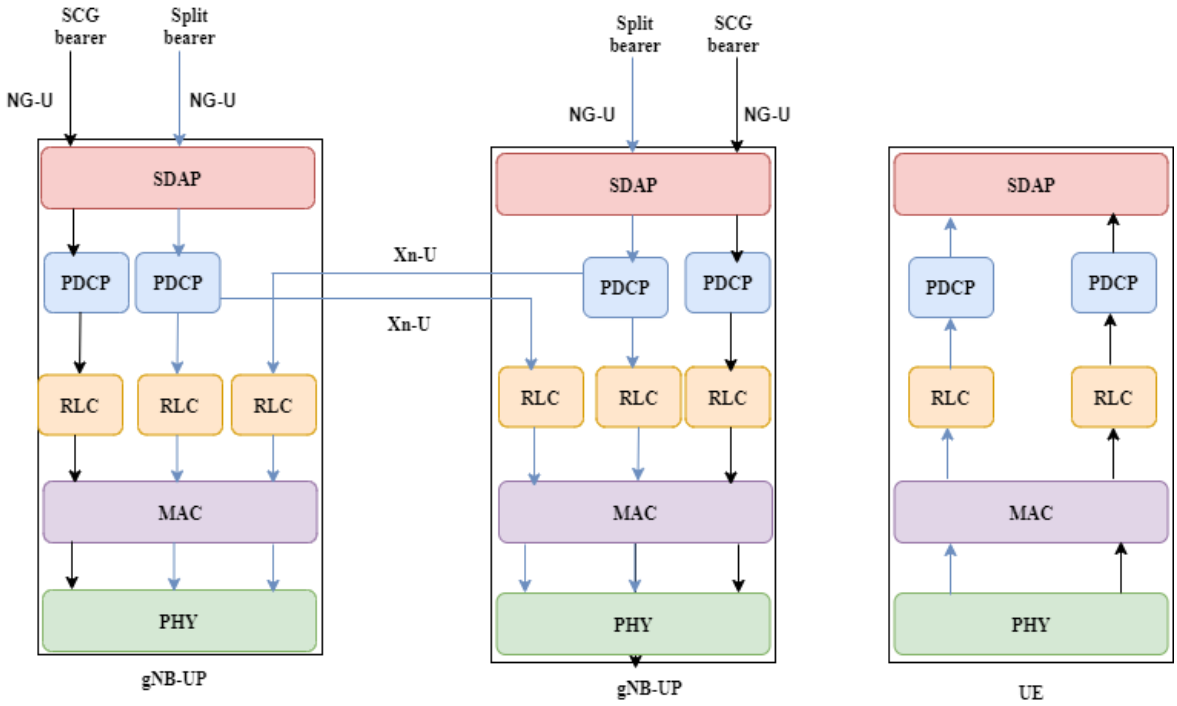


Figure 4. 9: SCG Bearer and Split Bearer flow in gNB-UPs and UE.

The SCG and Split bearers are both configured in gNB-UP to carry UP traffic. Split bearers are used to copy data from one gNB-UP to another gNB-UP via the Xn-U interface.

In the UE, it is necessary to configure different radio bearers. In particular, for the intended IMS voice call to be considered, the MCG SRB, SCG DRB and SCG split bearer shall all be configured per UE. Another benefit of the NR-NR architecture is that it allows for reduced UE computational resources since only the NR protocol stack is configured per UE when compared to the NSA.

4.4.1 Bearer Comparison Between NR-NR Architecture and NSA

A comparison between the NR-NR architecture and the NSA now follows. In the NR-NR architecture, the gNB-CP and gNB-UP nodes are used to differentiate between the CP and UP termination nodes over the NG interface. On the other hand, the CP/UP differentiation does not strictly apply for NSA. Besides that, in the NR-NR

architecture, the gNB-CP is specifically used to transmit CP signals over the air interface which is different from NSA where eNB can transmit both CP and 4G UP signals. Thirdly, in the NSA architecture, the MCG bearer functions both as SRB and DRB while in the NR-NR architecture, the MCG bearer is strictly a SRB that transmits CP signalling. Fourthly, in the NSA architecture the split bearer terminates at the eNB (macro node) while it is terminated at the gNB-UP in the NR-NR architecture. Finally, each architecture involves packet duplication at the PDCP layer especially during handover procedures. In the case of NSA for the split bearer, it involves packet duplication which refers to a duplicate copy of the same message sent over two nodes. A more extensive comparative study of the NR-NR Architecture and the NSA is discussed in Chapter 5 while the differences in terms of SRBs and DRBs are summarized in Table 4.2.

Table 4. 2: Comparison of Bearers between NSA and NR-NR Architecture.

	NSA	NR-NR
1.	The eNB and the gNB are defined to provide both EUTRAN and NR access to UE.	The gNB-CP and the gNB-UP are defined to differentiate between the CP and UP respectively.
2	The master node (in this case eNB) transmit both CP and UP data.	The master node (in this case gNB-CP) strictly transmits CP signals.
3	The MCG bearer can function both as SRB and DRB. Also, it uses the MCG split bearer to boost data rates for the secondary node.	The MCG bearer is strictly SRB for CP signalling.
4	Split bearer terminates between eNB and gNB.	Split bearer terminates between gNB-UPs.
5	Packet duplication between eNB and gNB.	Packet duplication between two gNB-UPs.

4.5 PDU Session Establishment in NR-NR Architecture

For the NR-NR architecture, the PDU session is established immediately after the RRC connection configuration complete messages has been sent from the UE to gNB-CP. To this effect, the AMF and the gNB-CP establish a PDU session to create a default bearer for the purpose of UE data transmission in anticipation for the UE data. Figure 4.10 shows the PDU session establishment in the architecture. Once the UE successfully attaches to the both the gNB-CP and the gNB-UP, the gNB-CP in collaboration with the AMF triggers the establishment of the PDU session in gNB-UP. The AMF sends a PDU session creation message to the gNB-CP for the purpose of creating a PDU session between the UE and the gNB-UP which also includes QoS parameters for the anticipated default bearer.

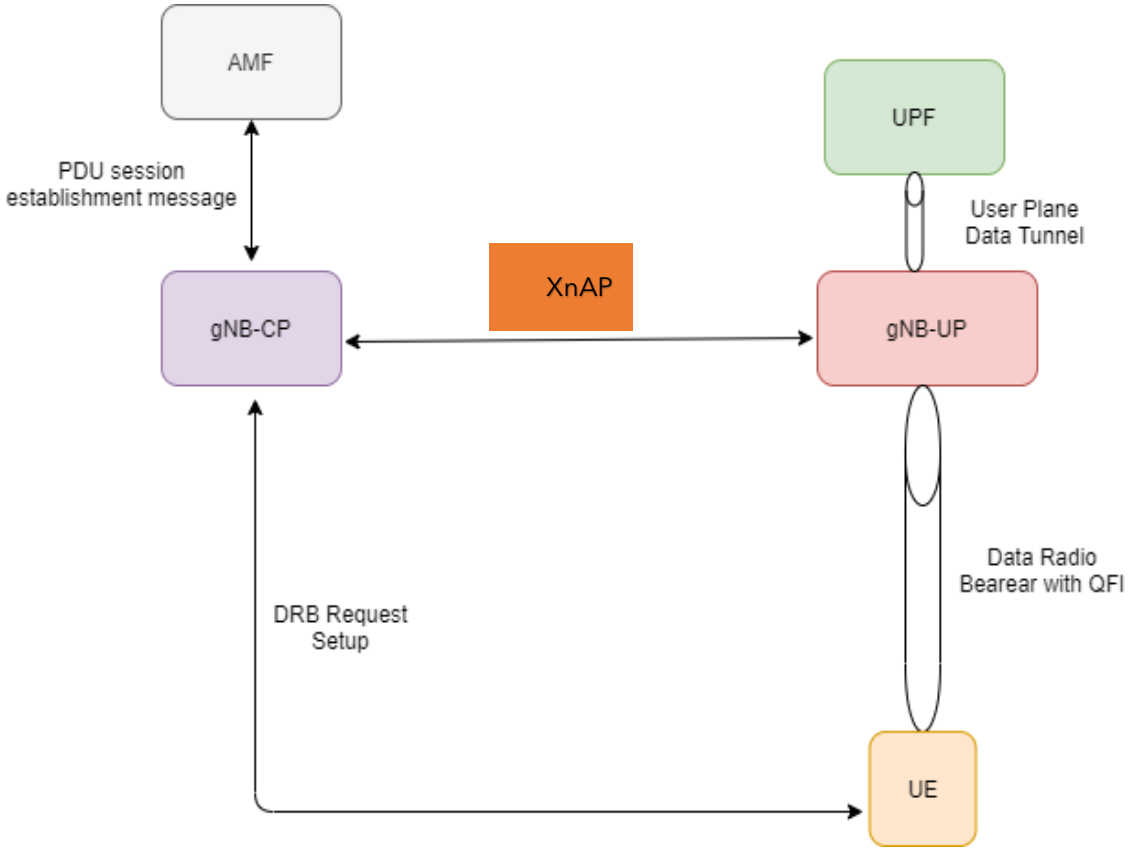


Figure 4. 10: PDU Establishment in gNB-UP.

In order for a UE to transmit or receive data, a PDU session shall be established between the UE and gNB-UP under the control of the AMF. Once the PDU session is successfully established, default bearer is created. UE then receives or transmits data via the default bearer.

The gNB-CP on receiving the PDU session creation message, sends a DRB set-up message to the UE with the parameters for the DRBs to be established in the UE via the SRB already created during RRC attach procedure. The UE then establishes a default DRB to the gNB-UP including a QFI tag to be used to map the data radio bearer with the UP-data tunnel in order to meet the 5G QoS requirements. After the establishment, the UE sends DRB set-up complete message to the gNB-CP and the gNB-CP sends a PDU session acknowledgment message to the AMF to indicate that the PDU session was successfully established in the gNB-UP. UE's data path is now successfully created and data can be transmitted from the data network to the UPF down to the DRB to the UE. The bearer established is a default bearer and as such adequate resources are provided as long as the UE is attached to the network. The same procedure is repeated to set up another default bearer with 5QI value of 5 to carry IMS SIP messages for the purpose of setting up IMS call. Hence, at the end of the PDU session, two default bearers are created. One default bearer with 5QI value 8 is created for UE's data and a second default bearer as already mentioned is created for SIP signalling. Other dedicated bearers in a designated PDU session shall be created upon request to carry UE's various applications such as for IMS voice or video call. In other words, one default bearer is used for SIP signalling and another dedicated bearer is used to carry the IMS voice call. For instance, for any ongoing IMS call, two bearers are active one default bearer to carry SIP messages

and a dedicated bearer for carrying voice. For an IMS video call, three bearers are active, a default bearer for carrying SIP signalling and two dedicated bearers carrying voice and video data respectively.

4.5.1 Signalling Procedures in NR-NR Architecture for Data Transmission

The detailed signalling that occurs in the NR-NR architecture from initial UE attachment to the transmission of UL and DL data is shown in Figure 4.11. The brief description of the messages and signals are shown on the right-hand side of the figure. The signalling involves RRC connection for 5G DC, NAS signalling, gNB-UP node addition and UE data transfer. Recall from Chapter 3 TEID is used to identify which tunnel a particular PDU/bearer belongs to and is usually identified in the GTP header. Generally, a default bearer or dedicated bearer is identified by the GTP-U TEID carrying the IP address of both the source and the destination nodes.

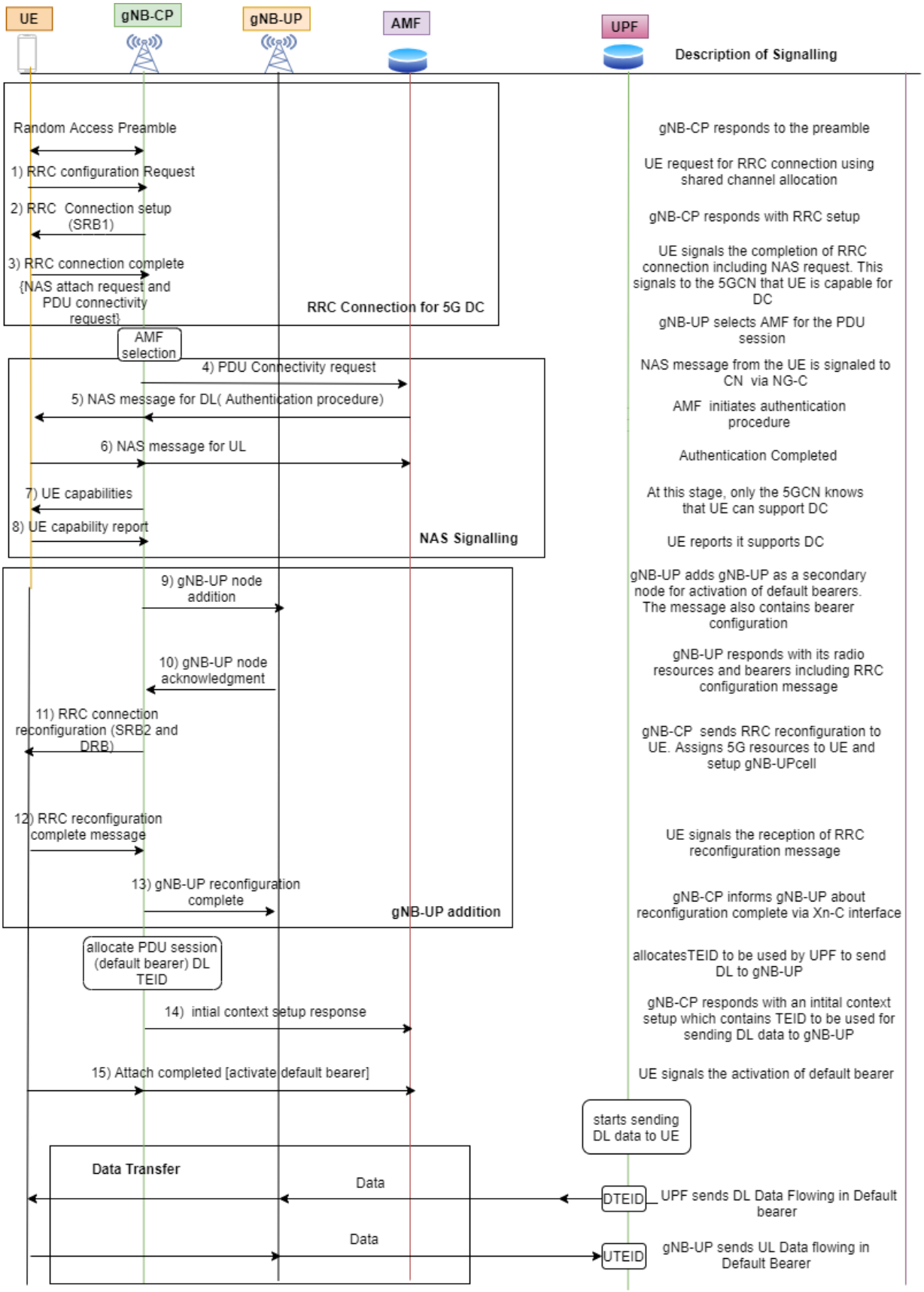


Figure 4. 11: Signalling Procedures in NR-NR Architecture (UE attachment).

4.6 UE Mobility

In this section, the gNB-UP node addition and release procedures of the proposed architecture is described. For this study mobility refers to the movement of the UE from one gNB-UP to another gNB-UP and for a scenario where both gNB-UPs are in the coverage of a single gNB-CP as shown in Figure 4.12. The RRC entity resides only in the gNB-CP from the UE's perspective since the RRC messages from gNB-UP are relayed to gNB-CP and transmitted to the UE. Based on the Measured Report (MR) reported by the UE, the gNB-CP makes necessary decision to add the target gNB-UP for the continuation of UE data. The MR contains Radio Resource Management (RRM) measurements and the gNB-CP acts accordingly based on the received MR from the UE. RRM refers to a series of techniques or procedures for power control, scheduling, handover, radio link monitoring, cell selection etc. The UE is usually configured to make these measurements and report via them via MR to the gNB-CP.

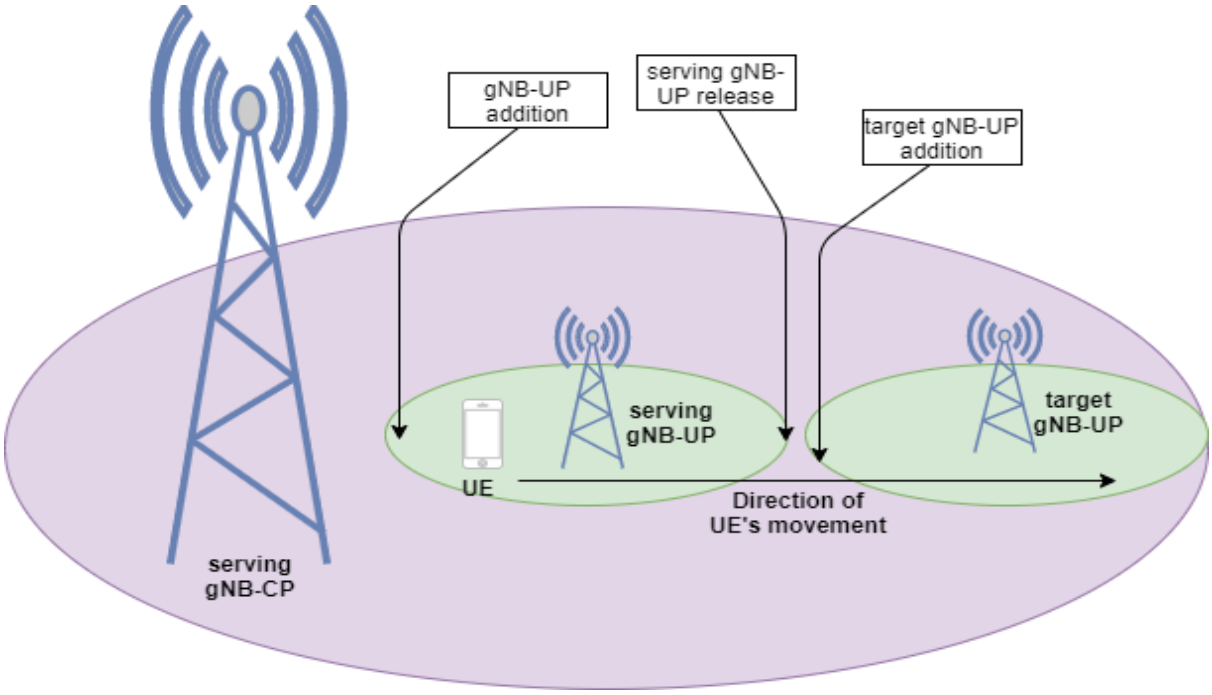


Figure 4. 12: UE Mobility.

As UE moves in the direction shown, the ongoing voice call is handed over from the serving gNB-UP to the target gNB-UP. The gNB-UP release means that the radio resources allocated to the UE by the serving gNB-UP is released.

It is worth mentioning that the UE also makes measurement such as Channel Status Information (CSI) report through gNB-CP to indicate its channel quality in order for the gNB-UP to assign a radio beam adopting Massive Multi User Multiple-Input Multiple-Output (MU-MIMO) technology to the user. Now each beam transmitted at the downlink is carrying a data with a rate commensurate to the required Modulation and Coding Scheme (MCS). MCS is a modulation and coding scheme used by gNB-UP to transmit UE data depending on UE's channel quality such as Quadrature Amplitude Modulation (QAM) and so on while beamforming refers to

techniques used by gNB-UP to assign radio beam to users to boost throughput especially for cell edge users.

From Figure 4.12, it shows the UE mobility as it moves from the serving gNB-UP to the target gNB-UP. Whenever a UE is in the coverage of a gNB-UP, gNB-UP addition procedure is used to configure the UE for DC and after the configuration procedure, that gNB-UP becomes the serving gNB-UP. The gNB-CP makes the decision to initiate gNB-UP change procedure based on the MR reported by the UE, while the UE moves from the serving gNB-UP to the target gNB-UP as shown in Figure 4.13. It must be emphasized that only the gNB-CP can initiate the gNB-UP node addition procedure thereby avoiding any data interruption or delay during g-NB-UP node addition and handover procedures. The gNB-UP change procedure involves the release of the current RRC connection to the serving gNB-UP and initiates a new RRC connection to the target gNB-UP. Once the UE attaches to the target gNB-UP, a path switching signal is trigged by gNB-CP via the AMF to modify the bearer from the serving gNB-UP to the target gNB-UP. Once bearer modification is done, UP data traffic is now forwarded directly from the UPF to the target gNB-UP. The steps discussed below describes the signalling messages involved for a target gNB-UP addition procedure and a summarized version is shown also in Figure 4.13.

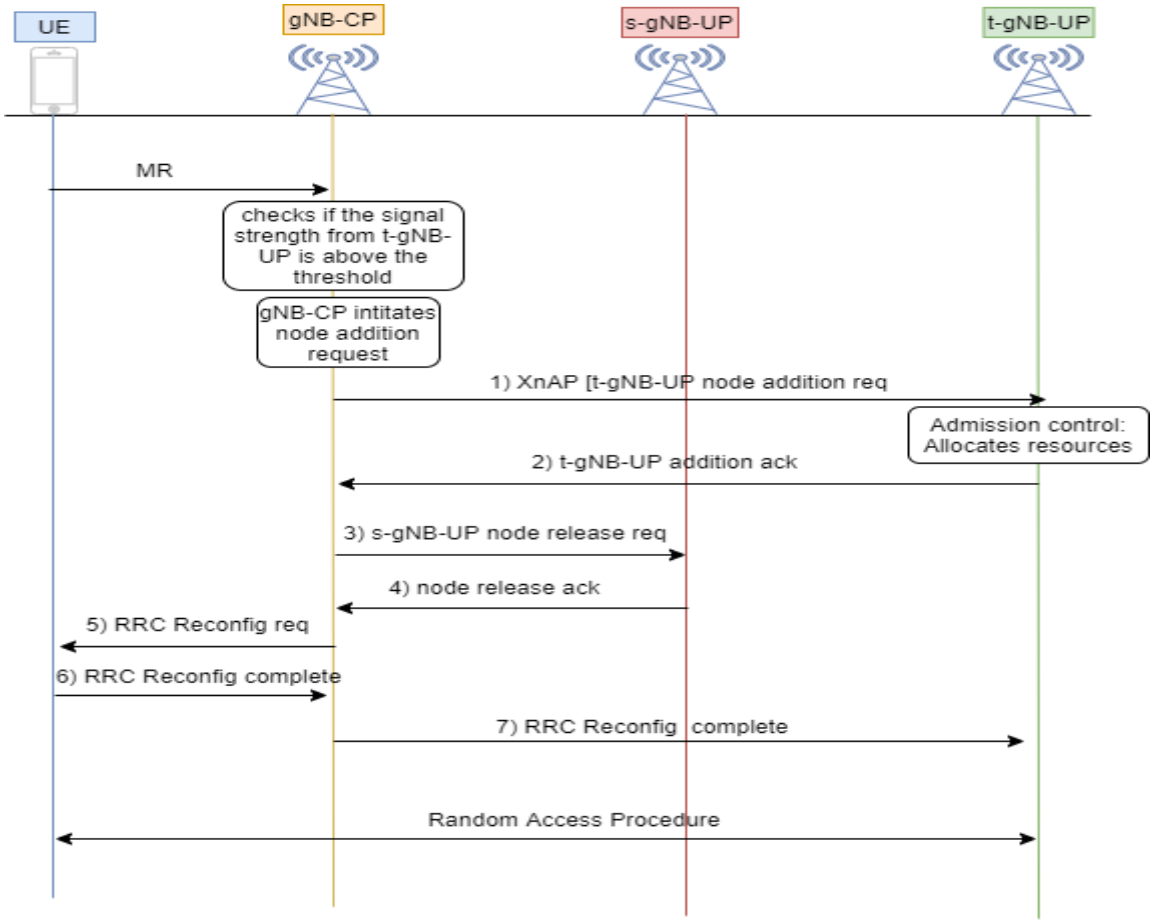


Figure 4. 13: The t-gNB-UP Node Addition.

The gNB-CP adds a target t-gNB-UP within its coverage in order to handover user data from the serving s-gNB-UP.

Step 1: As UE is in motion and moving away form s-gNB-UP, the received signal strength (SS-RSRP) from serving (s-gNB-UP) gradually becomes below a given threshold. The node addition decision is made by gNB-CP based on the signal received by the UE from the target (t-gNB-UP) and reported to gNB-CP. In addition, the gNB-CP extracts the UE's capabilities to infer the required amount of radio resources that the UE is allowed to use. Then the gNB-CP requests the target t-gNB-UP to allocate radio resources for the relevant UE and includes the UE's capabilities in the request.

Step 2: The t-gNB-UP sends to the gNB-CP a reject or acceptance message. For the scenario where the gNB-UP have the required available resources, it allocates the radio resources needed to the UE.

Step 3: The gNB-CP sends RRC connection release request to s-gNB to request the release of radio resources.

Step 4: The s-gNB-UP replies with a node release request acknowledgment confirming the release of the radio resources

Step 5: The gNB-CP then sends the RRC reconfiguration message to the UE.

Step 6: The UE applies the new RRC configuration and replies back to the gNB-CP with an RRC reconfiguration complete message.

Step 7: The gNB-CP informs the t-gNB-UP that the UE has accepted the configuration messages by forwarding the RRC reconfiguration complete message. The Random Access Procedure between the UE and the t-gNB-UP is performed. The UE attaches successfully to the t-gNB-UP for purpose of UP data transmission.

As the UE moves in the direction shown in Figure 4.13, based on the Measured Report (MR) which contains RRM information, the gNB-CP makes a handover decision. The following signalling steps takes place during the handover procedure while a summarized version is shown in Figure 4.14. Steps 1-7 shown in Figure 4.13 are equivalent to steps 1-8 shown in Figure 4.14 and is already discussed for node addition and hence, they are not described again to avoid repetition.

Step 9: The serving gNB-UP will forward and deliver buffered packets to the target gNB-UP by forwarding the Sequence Number (SN) status to the target gNB-UP. The reason for this is to ensure continuous data transmission without interruption.

Step 10: The buffered and in-transit packets in the serving gNB-UP are forwarded to the t-gNB-UP. The UE detaches from the serving gNB-UP and attaches to the target gNB-UP.

Step 11: The UE now synchronizes with the target gNB-UP and by so doing completes the attachment process to t-gNB-UP.

Step 12: The gNB-CP informs the AMF that the UE has changed cell through a path-switch request message. In this case, there is bearer modification in the UPF.

Step 13: The AMF forwards the path-switch request message to SMF.

Step 14: The SMF sends bearer modify request message to the UPF.

Step 15: The UPF completes the bearer modify signalling and replies with a bearer modify response to SMF.

Step 16: The UPF acknowledges the path switch request via the end-marker packet and sends it s-gNB-UP. The end-marker packet is in the header of the GTP which indicates that the PDU session should be tunnelled towards the t-gNB-UP.

Step 17: The s-gNB-UP forwards the end marker packet to t-gNB-UP. This informs the t-gNB-UP that data bearer shall be tunnel towards it.

Step 18: The SMF sends the path switch complete message to AMF.

Step 19: The AMF sends a path-switch complete acknowledgement to the t-gNB-CP to indicate a successful modification (tunnelling) of the bearer towards the t-gNB-UP.

Step 20: The gNB-CP sends UE context release message to s-gNB-UP to release its radio resources allocated.

Step 21 The s-gNB-UP releases the allocated radio resources and respond with a UE context release acknowledgement message.

The connection is now established and user data is transmitted to/from the UE to the t-gNB-UP via the UPF end-to-end.

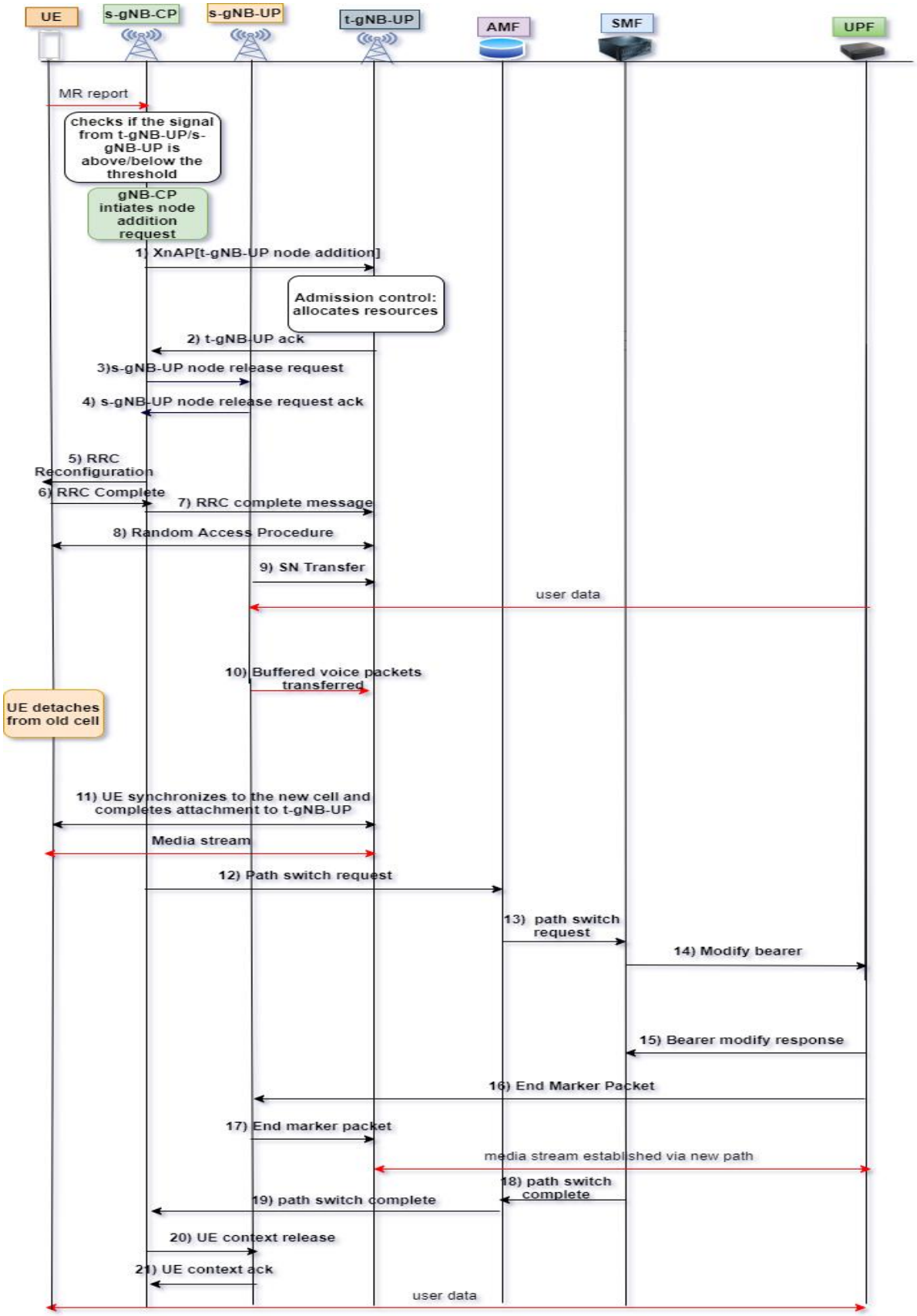


Figure 4. 14: UE Mobility Signalling in NR-NR Architecture.

CHAPTER 5

NR-NR ARCHITECTURE AND IMS VOICE CALL

This chapter shows the signalling messages that are transmitted in the NR-NR architecture in order to demonstrate the workings on how the proposed architecture handles control signalling for an IMS voice call. The chapter is divided into four sections specifically, the Section 5.1 presents the definition of control messages and the second section analyses the signalling messages that are transmitted in order to establish an IMS call specifically a VoNR call and the comparison with NSA is discussed. The third section analyses the signalling messages transmitted when a UE experiences handover during an ongoing IMS voice call and the comparison with NSA is also discussed. Finally, the fourth section discusses the setbacks envisaged in deploying the NR-NR architecture.

5.1 Definition of Control Signalling

To avoid any form of confusion, the thesis defines control messages as control information that are transmitted or communicated by the UE and various network nodes namely the gNBs, the AMF, SMF, UPF and the IMS Gateway in order to set up a connection. The connection considered in this study are VoNR call set up and VoNR handover connection. Also, the total number of control messages transmitted is to be used to determine the control overhead generated in order to start up an IMS voice call and perform VoNR handover in the NR-NR architecture. Since the design principle of the NR-NR Architecture is focused on the RAN and the 5GCN, the signalling analysis does not focus on the IMS network which is outside the scope of the study.

5.2 IMS Voice Call Setup

In this section the signalling procedure for setting up an IMS voice call for NR-NR architecture is discussed. Then a comparison is made on how VoNR call is established between the NR-NR Architecture and NSA.

5.2.1 Signalling Procedures for Setting up IMS Voice Call in NR-NR Architecture

In order to analyse how the NR-NR Architecture handles control signalling, an IMS voice call setup is discussed. Figure 5.1 shows the IMS Call set up considered for this study, for the purpose of a voice call between two UEs. The figure shows the various reference points and the interfaces between the nodes. The figure shows

two UEs where the UE on the right-hand side is the called UE and the UE on the left-hand side is the calling UE.

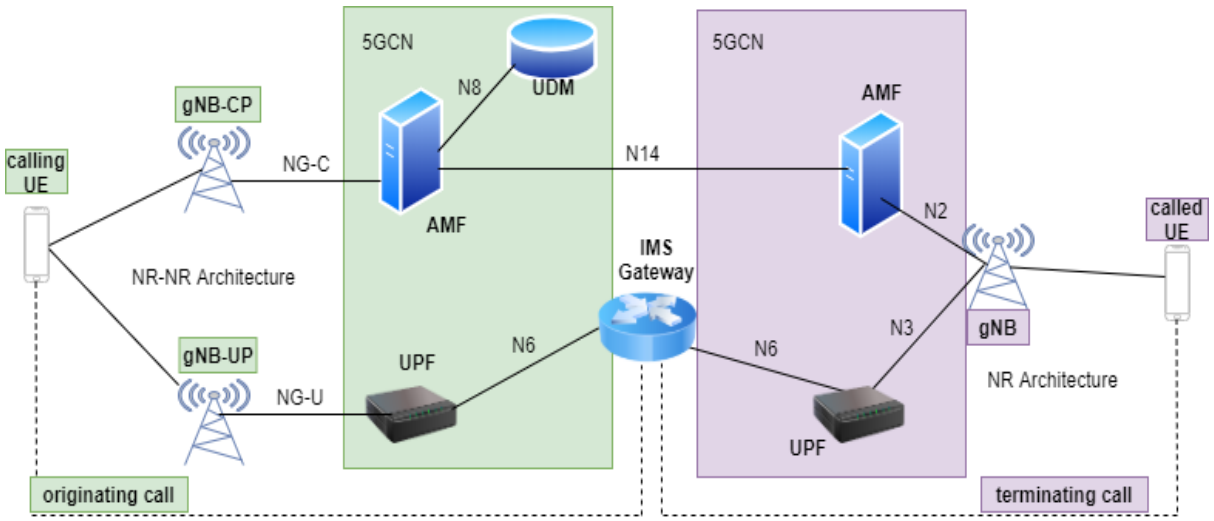


Figure 5. 1: Overview of a VoNR call.

The IMS Gateway provides access and connects the call to the called UE attached to its own network. Once the connection is established, the voice calls are transmitted through the UPFs via the gNB-UP to the gNB providing access to the called UE.

The calling UE is attached to the NR-NR architecture consisting of two base stations. The two base stations including the AMF and UPF of the 5GCN transmit signalling messages for the purpose of setting up an IMS voice call. In the same manner, the called UE is attached to a NR Architecture consisting of a single base station and the 5CN. The IMS network which is a network of its own, consists of an IMS Gateway that connects or routes the call to the other network (NR network architecture). Recall that the architecture of the IMS CN was already discussed extensively in Chapter 3. For the purpose of this research, it is assumed that the called UE and the calling UE are both subscribed to the same IMS network. The two UEs intending to communicate are both under the coverage of their base stations. When a UE registers to NR-NR network, the UE's profile is stored in the UDM and the subscriber's profile is probed for authentication and confirmation purposes whenever the UE attaches to the network. Also, during network subscription, in order for a UE to make an IMS call, the UE is usually subscribed to IMS services when it registers to an IMS network. In order words, the UE performs two registrations, one registration belonging to the NR-NR network and the second registration at the IMS network. The IMS registration is in the form of IP Multimedia Services Identity Module (ISIM) which is an application residing in the Universal Integrated Circuit Card (UICC) in the UE. This application contains parameters for UE authentication and identification in the IMS network such as IP Multimedia Private Identity (IMPI). The originating call and terminating call represent the calling UE network (NR-NR architecture) end and the called UE network (NR architecture) end, respectively. Figure 5.2 shows the signalling messages that are exchanged by various network nodes for the process of starting up an IMS call between the two mobile subscribers.

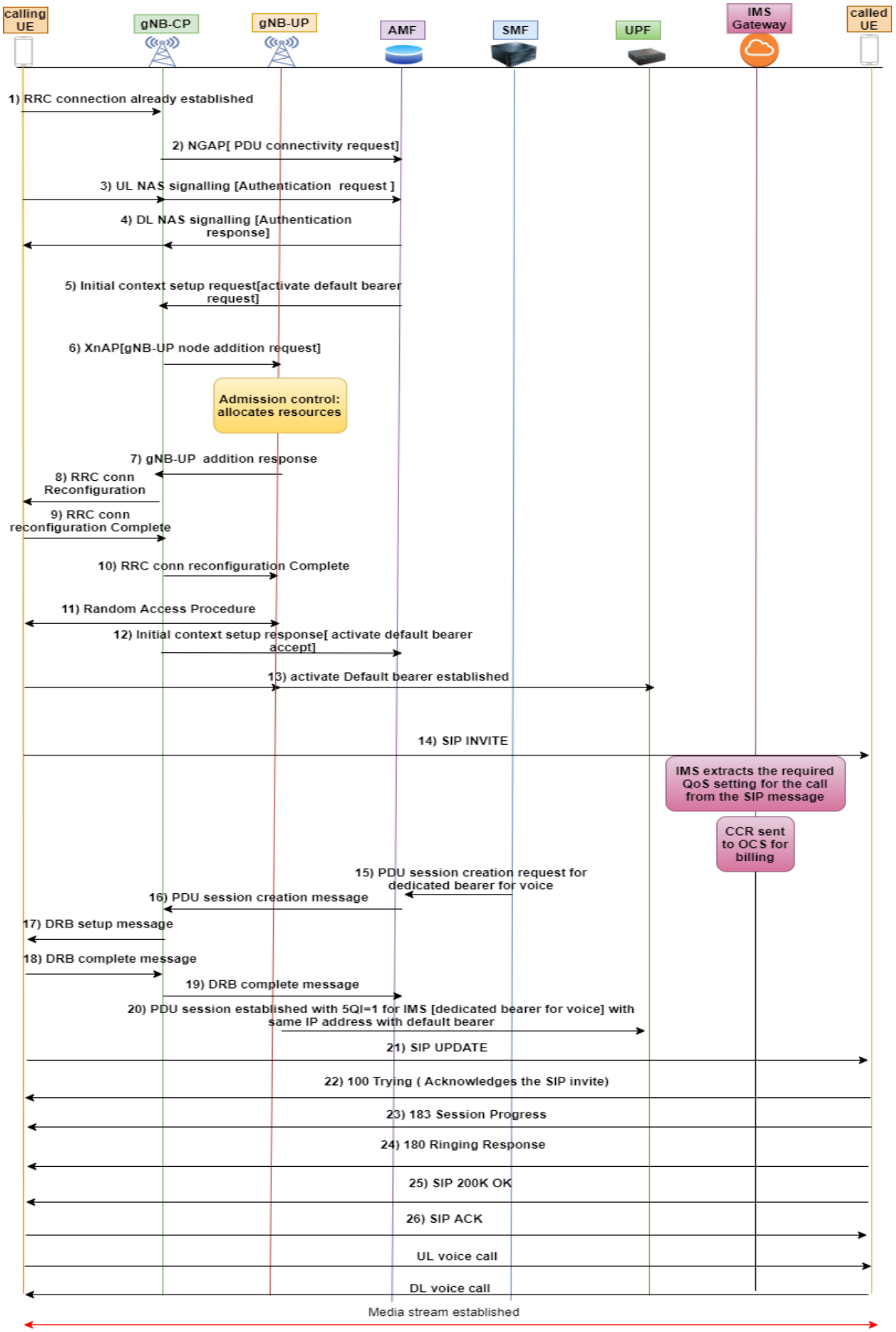


Figure 5. 2: Signalling Messages for IMS Voice Call Set-up.

Firstly, the UE attaches to a gNB-CP and together with the AMF perform UE authentication to gNB-UP in order to establish default and dedicated bearers which will handle the call traffic. The default bearer established is used to carry SIP messages to set the VoNR call. Whenever the UE makes a request to access other applications such as voice or video call, separate dedicated bearers are also established upon request to carry the UL and DL voice/video traffic. The step by step signalling messages are described as follows:

- **Step 1:** The subscribed UE attaches to gNB-CP and RRC connection is established.
- **Step 2:** The NAS message from the calling UE is signalled to the AMF via gNB-CP to indicate the request to establish PDU connectivity.
- **Step 3:** The UE initiates authentication process with the AMF via NAS signalling to validate the calling UE.
- **Step 4:** The AMF performs authentication of the UE and replies back to the UE to indicate that it was successful.
- **Step 5:** The AMF responds back to gNB-CP with an initial context setup request which also contains an activate default bearer request message via NGAP.
- **Step 6:** The gNB-CP adds gNB-UP as a secondary node to handle all UE data traffic (voice call) by sending gNB-UP node addition request via XnAP.
- **Step 7:** The gNB-UP performs admission control and allocates resources for the UE following the UE capabilities extracted from the node addition request. The gNB-UP acknowledges the allocation of UE's resources via node addition response message and sends it to gNB-CP. The node addition response message contains the RRC connection message of gNB-UP.
- **Step 8:** The gNB-CP sends an RRC connection message of gNB-UP via the RRC reconfiguration message to the calling UE.
- **Step 9:** UE extracts the new RRC connection message of the gNB-UP and sends an RRC reconfiguration complete message to the gNB-CP.
- **Step 10:** The gNB-CP forwards the RRC reconfiguration complete message to the gNB-UP.
- **Step 11:** The random access procedure is performed and the calling UE successfully attaches to gNB-UP.
- **Step 12:** The gNB-CP responds to the AMF signalling the attachment of the UE to the gNB-UP and the acceptance for the activation of a default bearer.
- **Step 13:** Default bearer signalling is established between the UE and gNB-UP. In this stage the default bearer to the IMS network is activated which has a unique IP address.
- **Step 14:** The calling UE makes a call and SIP INVITE signalling is sent to the called UE via the active default bearer which is already established. The SIP INVITE message contains the codec to be negotiated for the intending call including all SIP parameters such as bandwidth details, QoS parameters etc

for the High Definition (HD) call. The IMS network (Application Server) will extract the required QoS parameters for the call. The IMS network sends Charging and Control Rules (CCR) to the Online Charging System (OCS) through diameter signalling to enforce the charging and billing of the intending voice call. Diameter is a signalling protocol for billing, charging and authentication used by the IMS network.

- **Step 15:** The SMF sends a dedicated bearer request message to the AMF to create a PDU session for the dedicated voice bearer with 5QI= 1 with the same IP address as the default bearer already established during UE's initial attachment to network.
- **Step 16:** The AMF forwards the PDU session creation message to the gNB-CP.
- **Step 17:** The gNB-CP sends a dedicated bearer request set up message to the calling UE.
- **Step 18:** The UE confirms the establishment of the DRB and the voice call can now be performed in the NR.
- **Step 19:** The DRB complete message is sent to AMF from gNB-CP to notify the completion of DRB.
- **Step 20:** The dedicated bearer is activated and ready to carry voice traffic.
- **Step 21:** The calling UE sends the SIP UPDATE message to the called UE to confirm that the 5G network can support the IMS call request for the intending call.
- **Step 22:** The called UE sends the **100 Trying** message to acknowledge the **SIP INVITE** message for the intending call.
- **Step 23**: The called UE also sends the **183 [Session Progress]** confirming that the resources have been reserved for the terminating call side.
- **Step 24:** The called UE sends the **180 Ringing response** message towards the calling UE for alerting.
- **Step 25:** If the called UE answers the call, a **SIP 200 OK [INVITE]** message is towards the calling UE to indicate that the called UE has answered the call.
- **Step 26:** The calling UE sends a **SIP Ack** message to the called UE to acknowledge the reception of the **200 OK [INVITE]** message.

The RTP voice packets (voice call) are then transmitted via the dedicated bearer from the calling UE to gNB-UP via the UPF to the IMS gateway. From the IMS gateway, it is then routed to the terminating network end via its UPF to the called UE. Both the CP connection and UP connection are maintained for the duration of the voice call. At such the gNB-CP handles all the control signalling needed to established an IMS voice call and gNB-UP transmits all the UE's voice call.

5.2.2 IMS Voice Call Signalling in NSA

Figure 5.3 shows the signalling procedures for starting a VoNR call in the NSA. For the NSA as shown in the figure, the UE first establishes a PDN connection in the eNB

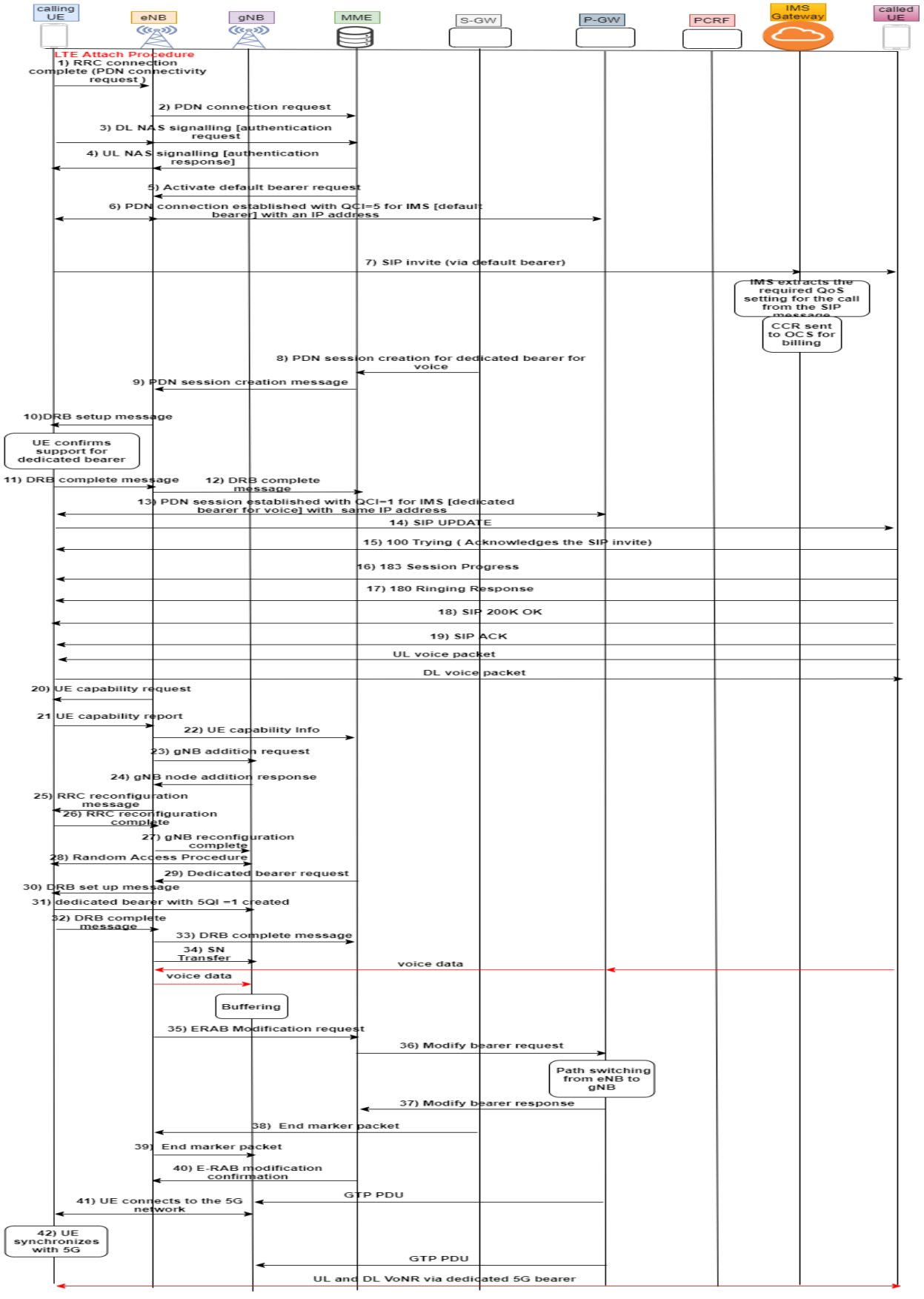


Figure 5. 3: IMS Voice Call Start Up Signalling for NSA.

for the purpose of performing a voice call. The default bearer for SIP signalling is established through the eNB. Recall from the NSA, the establishment of a PDN connection in eNB is made possible because eNB is connected not just to the MME but also the S-GW-U. Once the SIP signalling is transmitted, a dedicated bearer for voice is established to carry the voice data. This whole signalling procedure for setting up a VoLTE call is illustrated in Figure 5.3 from step 1 to step 19. For the case the UE makes a request for a VoNR call based on UE's capabilities, the call request has to be switched from the eNB to gNB. In doing so, a PDU connection (dedicated bearer with 5QI value of 1 with a new IP address) has to be established in gNB to carry the voice traffic. The reason for a new IP address is because the destination node has changed from the eNB to the gNB. This switching process triggers additional control messages to be transmitted in order to switch the UE from a VoLTE call to a VoNR call. The control signalling messages for switching to a VoNR is illustrated as shown in Figure 5.3 from Step 20 to Step 42.

5.2.3 Comparison of IMS Voice call between NR-NR Architecture and NSA

When the NR-NR architecture is compared with the NSA, it shows that for NR-NR architecture, the UE is capable of making a VoNR call directly without the need of first setting up a VoLTE call and depending on UE's capabilities for a VoNR call, the call is switched from a VoLTE call to a VoNR call. This means that the control signalling need to first establish a VoLTE call is no longer needed in the NR-NR architecture which is shown from Step 1 to Step 13 excluding Step 3 and 4 of Figure 5.3. The reason why Step 3 and Step 4 are excluded is because they are NAS authentication signalling which are also transmitted in the NR-NR architecture. In addition, the switching from a VoLTE call to a VoNR call introduced additional control signalling from Step 35 to Step 40 also including UE capability signalling illustrated from Step 20 to Step 23 as shown in Figure 5.3. These additional signalling are not transmitted in the NR-NR architecture since a VoNR call is directly established via the gNB-UP. As a result, 26 signalling messages are transmitted in setting up a VoNR call in the NR-NR architecture. This reduction is achieved as follows, first, in NR-NR architecture, control signalling needed to start a VoLTE call are not performed and secondly, the additional control signalling needed to switch from a VoLTE call to a VoNR call (Step 35-40 and Step 20-23) are not also performed. In simple terms, for NSA to perform a VoNR call, it needs to switch the UE from a VoLTE call to a VoNR call while in NR-NR Architecture, a VoNR call is directly performed.

From the extensive signalling procedures discussed for the purpose of setting up an IMS voice call, it follows that all control signalling for the purpose of establishing PDU connectivity between the UE and AMF is handled only by the gNB-CP whereas the gNB-UP is strictly handling all UE call traffic. By decoupling the UP plane from the gNB-CP has ensured that the gNB-CP is not involved in handling UE data traffic.

It also allows for a flexible deployment of either the gNB-CP or the gNB-UP by the network operator thereby preventing coverage gaps especially in a densely distributed small-cell environment where coverage and capacity is usually a challenge and we use small cells to boost throughput for UEs. Besides, by splitting the CP and UP, it reduces the load handled in gNB-CP cells which helps to ensure efficient delivery of CP signals.

5.3 Signalling Procedures for Handover During an Active IMS Voice Call

5.3.1 UE's Mobility between gNB-UPs within the Coverage of gNB-CP

In this section, to avoid repetition the UE's mobility between gNB-UPs within the coverage of gNB-CP has already been discussed extensively in Section 4.5. The same controls signals are transmitted in order to handover the voice call from s-gNB-UP to t-gNB-UP, also as shown in Appendix A. The difference between Figure A-1 in Appendix A and Figure 4.14 is that an active IMS voice call is ongoing and data traffic is referred as voice call instead of user data as illustrated in Figure 4.14. There is no significant difference in terms of control signalling achieved during UE's mobility within the coverage of gNB-CP when compared with the NSA.

5.3.2 Handover signalling when UE moves out of Coverage of gNB-CP

In this section, we consider the handover procedure of a UE when it moves out of range of a gNB-CP. When a UE moves out of coverage of a serving gNB-CP, the s-gNB-CP of the NR-NR architecture triggers a handover request to disconnect the UE from s-gNB-CP and connects the UE to a target t-gNB-CP. Figure 5.4 shows the control signalling messages while the details of each step are discussed as follows:

- **Step 1**: Based on the periodic MR, the s-gNB-CP triggers a handover request. The handover request message shall only be transmitted if the signal strength of the s-gNB-CP is below a threshold and if the signal strength received by UE from the t-gNB-CP is above a threshold. The s-gNB-CP sends the handover request to the t-gNB-CP based on the signal strength received by the UE that is above a threshold. The t-gNB-CP performs admission control and allocates the available required resources for the incoming UE.
- **Step 2:** For a case when the t-gNB-CP can accommodate the new UE, the t-gNB-CP sends an RRC configuration node request to a t-gNB-UP which will handle the UE's voice traffic and also allocates resources for the incoming UE within its own coverage.
- **Step 3:** The t-gNB-UP sends a node addition acknowledgment to the t-gNB-UP confirming allocation of radio resources.
- **Step 4:** The t-gNB-CP sends a handover acknowledgment message to the s-gNB-CP confirming the allocation of resources for the handover.
- **Step 5:** The s-gNB-CP sends node release request to the s-gNB-UP to release its radio resources.

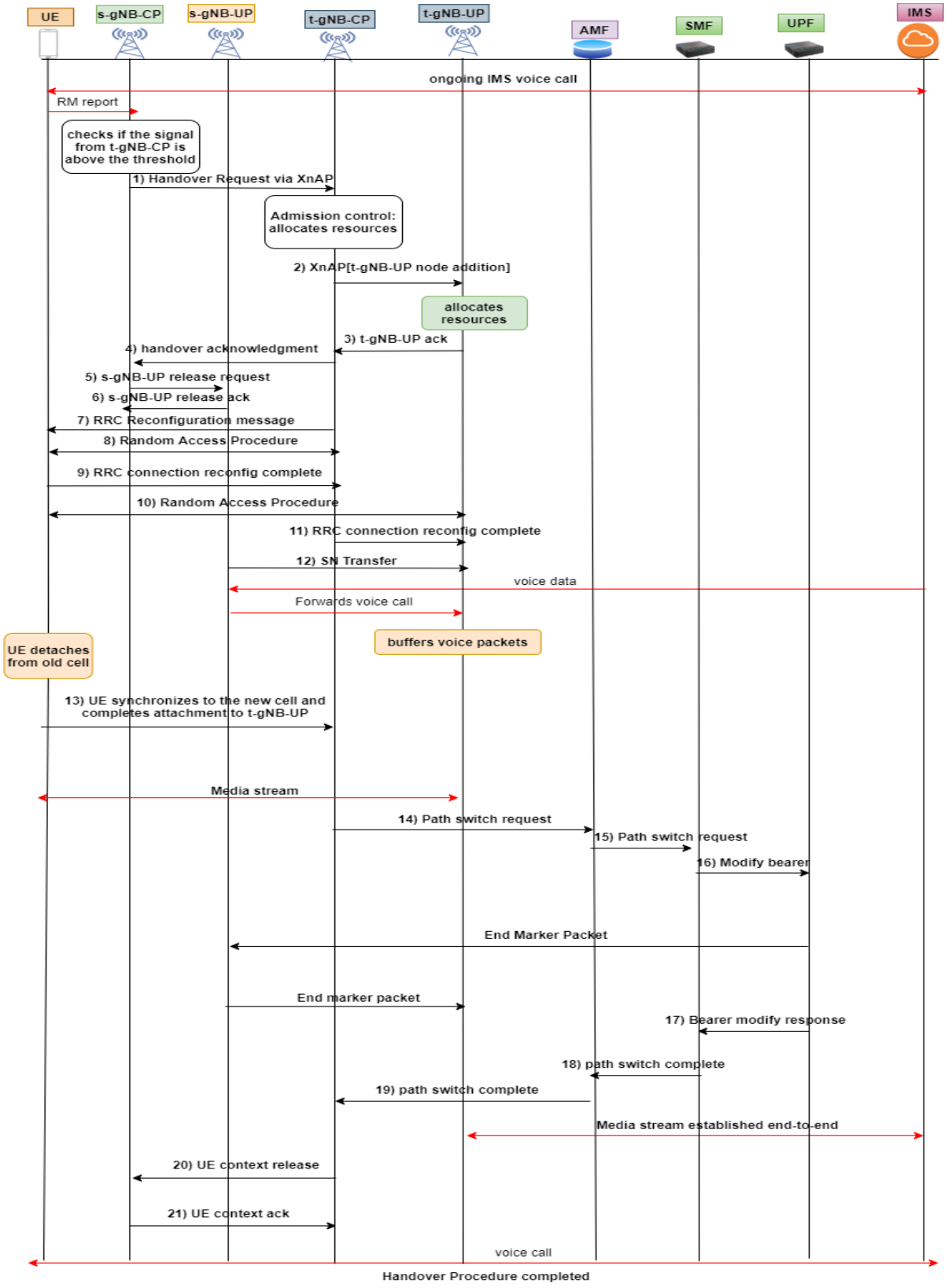


Figure 5. 4: Handover Signalling when UE moves out of Coverage of gNB-CP.

- **Step 6:** The s-gNB-UP responds back to the s-gNB-CP with a node release acknowledgment message.
- **Step 7:** The t-gNB-CP sends RRC reconfiguration message to the UE to configure the new radio resource configuration.
- **Step 8:** The Random Access Procedure is performed between the UE and the t-gNB-CP.
- **Step 9:** The UE sends an RRC reconfiguration complete message to t-gNB-CP to signal the completion of the RRC connection to the t-gNB-CP.
- **Step 10:** The UE performs Random Access Procedure with the t-gNB-UP.
- **Step 11:** The t-gNB-CP sends the RRC reconfiguration complete message to the t-gNB-UP.
- **Step 12:** The s-gNB-UP will forward and deliver buffered packets to the t-gNB-UP by transferring Sequence Number (SN) status to the t-gNB-UP. The reason for this is to ensure continuous voice transmission without interruption. At this stage, the UE detaches from the s-gNB-CP.
- **Step 13:** The UE synchronizes with the new cell and completes the attachment to t-gNB-CP. The Media stream (carrying RTP packets) is now established at this stage. The same UE's IP address is used for the voice call transmission and the handover procedure is completed.
- **Step 14**: The t-gNB-CP sends a path switch message to the AMF.
- **Step 15:** The AMF forwards the path switch request to the SMF.
- **Step 16:** The SMF sends a modify bearer request to the UPF. The UPF transmits a path switching signalling for DL. The Media stream is now established between the UE and the t-gNB-UP. An end marker packet is forwarded to the old bearer path towards the s-gNB-UP to indicate the termination of the UE data in the old path (s-gNB-CP).
- **Step 17**: The UPF sends a bearer modify response to the SMF to acknowledge the modification of the bearer.
- **Step 18**: The SMF forwards the path switch complete message to the AMF.
- **Step 19**: The AMF sends the path switch response reply message to the t-gNB-CP. At this stage the Media stream is established end-to-end between the UE and the UPF.
- **Step 20**: The t-gNB-CP sends the UE a context release message to the s-gNB-CP to release its radio resources.
- **Step 21**: The s-gNB-CP releases UE's context and responds with a context release acknowledgement message to the t-gNB-CP.

At this stage, end-to-end media stream is established between the UE and the UPF resulting to a completed handover procedure. The red line in Figure 5.5 shows the optimized data path for the UP traffic. The data coming from the s-gNB-UP is forwarded directly to the t-gNB-UP without being forwarded to the s-gNB-CP.

5.3.3 Handover Signalling when UE moves out of Coverage of eNB in NSA

Figure 5.5 shows the handover signalling transmitted in NSA when UE moves out of coverage of the s-eNB.

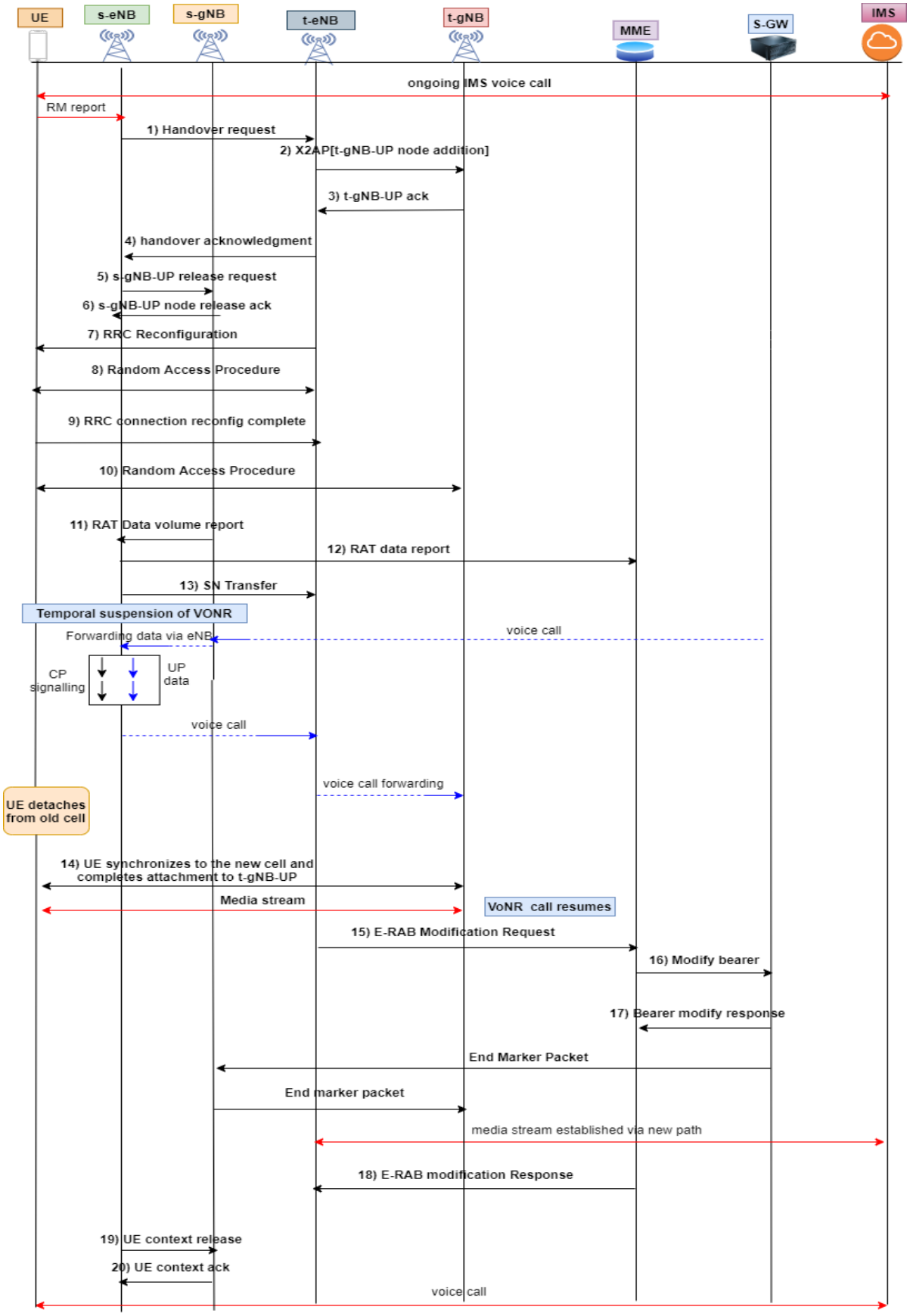


Figure 5. 5: Handover Signalling in NSA (Out of Coverage of s-eNB).

The box shown in Figure 5.5 shows that for a VoLTE call, the eNB both performs CP signalling and transmits UP data represented by black and blue arrows as shown respectively. From the MR, as long as the signal received from the t-eNB is higher than a threshold, handover procedure can begin. The RAT Data volume message report is also signalled as shown in Step 10 from the s-gNB to the s-eNB which shows the volume of data (voice call) delivered to the UE over the NR for the ongoing VoNR call. This message is transmitted to the MME in step 11 to provide information on used NR resources. In the NSA whenever a UE moves out of coverage of the s-eNB, the VoNR call is temporarily suspended and the voice traffic coming from the S-GW to the s-gNB is forwarded to the s-eNB. The s-eNB then forwards the voice traffic to the t-eNB which forwards the voice traffic to the t-gNB to resume VoNR call. Recall that bearers for VoLTE are already established during initial call set up. The reason for this temporal suspension of VoNR during handover is because of the architectural design of the NSA. In the NSA, for handover between two eNBs, it is designed that only the s-eNB can handover data (voice call) to the t-eNB. Afterwards, both data and voice traffic can then resume via the gNB. The second reason being that as UE is moving closer to the cell edge of the s-gNB, the VoNR call has a high tendency of failing when compared to VoLTE. To prevent this situation, the UE is forced to temporarily suspend the VoNR call and switch to a VoLTE call. Once the handover to the t-eNB is completed, the VoNR call can then resume. Besides, for NSA, the early deployment of 5G base stations are not well optimized to handle VoNR calls. This is because in NSA, the CN is still the EPS unlike the standalone architecture where NR and 5GCN are better optimized to support a VoNR call.

5.3.4 Comparison of Handover between NSA and NR-NR Architecture

The major difference between the NR-NR architecture and the NSA is as follows. In the NR-NR architecture, by introducing the principle CUPS in the 5G RAN, all control signalling messages transmitted for the purpose of handover procedure are all handled by the gNB-CP while UE voice call is maintained only by the s-gNB-UP. This is not the case for the NSA as shown in Figure 5.6 where the eNB both handles control signalling and UP traffic. Secondly, whenever UE moves out of coverage of the s-gNB-CP, there is no interruption of the UE voice call unlike the NSA. As a result, the voice call is forwarded directly from the s-gNB-UP to the t-gNB-UP. Unlike in the NSA, where the call has to be forwarded first from the s-gNB to the s-eNB and then to the t-eNB and the VoNR call finally resumed when the call is forwarded to the t-gNB. The disadvantage of switching from a VoNR call to a VoLTE call is that it temporarily causes interruption of the UE voice call in switching from 5G to 4G and then back to 5G. It is also believed that there is a possibility that this switching can introduce additional end-to-end latency to the voice call experienced by users.

In terms of similarity, it is expected that both the NR-NR architecture and the NSA reduces the rate of the gNB-CP/eNB cell change. This is because the gNB-CP/eNB cells size is usually large providing wider network coverage. Hence, there is a possibility of lower occurrence of the gNB-CP/eNB handover compared to the handover within the coverage of gNB-CP/eNB. Secondly, for handover procedures both gNB-CP and eNB initiate the handover processes. It is observed that the

movement of the UE within the coverage of a gNB-CP/eNB does not trigger handover request message. Only the gNB-UP/gNB node addition request message is used to handover the UE call as it moves from the s-gNB-UP to the t-gNB-UP within the coverage area of a gNB-CP.

By introducing the CUPS principle in the RAN, has ensured a continuous uninterrupted transmission of the UE voice traffic during handover.

5.4 Overall Architectural Comparison with NSA

In general, both NSA and NR-NR architecture have different goals. While the goal of NSA was to expedite the deployment of 5G network for UEs, the goal of NR-NR architecture is for a complete CP/UP split in the 5G RAN which allows for the scaling of each plane's resources. The aim of the comparison is not to present a case that NR-NR architecture is better in terms of performance because they are not really competing alternatives. Having said that, the comparison is made to strictly show the architectural differences between the two architectures because they have similar characteristics in terms of; both utilizes two base stations, operate in DC mode, targeted for HetNets and a deployment option for 5G networks.

The architectural comparison is presented as follows. In terms of technology involved, the NR-NR architecture is comprising of two 5G base stations supporting a UE that is configured in DC mode. On the other hand, the NSA encompasses two technology namely 4G and 5G to provide the UE access to both 4G and 5G network. The UE usually configured in DC mode attaches to a 4G base station (eNB) which provides 4G coverage and also attaches to a gNB whenever it needs 5G data rates. In the NR-NR architecture, the MCG bearer is strictly a signal bearer where SCG and SCG split bearers are used only for carrying user data unlike in NSA where MCG bearer is both a signal and data bearer. This is so because the eNB handles both control signalling and user data. For the case of NR-NR architecture only NR protocol stack is configured in the UE which saves the need for more computing resources. Besides, in the NR-NR architecture the CN is the 5GCN providing MTC, URLLC services and so on while in the NSA, the UE is configured for both LTE and NR protocol stacks; as a result, there is a need for more computing resources for the UE. Moreover, the CN of NSA is the EPC which allows for EPS fallback which on the other hand is not supported yet in the NR-NR architecture. The decouple of the CP and the UP in NR-NR architecture needs good co-ordination to achieve optimal network functionality and in terms of cost of deployment, it is envisaged that the NR-NR architecture might be expensive to be deployed because of the new 5GCN added and also based on the fact it uses two gNBs. Finally, the NR-NR architecture is expected to be the future deployment framework for a 5G network from 2024 to 2025 and beyond. However, for the NSA and in terms of ease of management, it is complex to manage because two RATs are used at the same time; finest co-ordination is needed for optimal network functionality. In addition, since the gNB node is added to the already exiting LTE base station, the cost of deployment is less

expensive and it is designed to be the immediate deployment option of 5G by accelerating the launch of 5G networks. A compact comparison between the NR-NR architecture and NSA is presented in Table 5.1.

Table 5. 1: Architectural Comparison of NR-NR with the NSA.

	Features	NR-NR Architecture	NSA
1.	**Generation Technology**	Uses two NR base stations (only 5G).	Uses both the LTE and the NR base stations. (4G and 5G).
2.	**Attachment Procedure**	UE attaches to gNB-CP for coverage and to gNB-UP providing UP data.	UE attaches to the eNB for both coverage and data. It also attaches to the gNB whenever it needs higher bit rates.
3.	**Signal Radio Bearer Configuration**	Only the MCG signal radio bearer is configured to carry control signalling.	The MCG signal bearer is not configured only for carrying control signalling.
4.	**Data Radio Bearer Configuration**	SCG and Split bearers are both configured and terminated at gNB-UP.	Same as NR-NR Architecture but the bearers are terminated at the macro node.
5.	**UE bearer Configuration**	Only NR protocol stack is configured in the UE.	Both LTE and NR protocol stack are configured in the UE
6.	**Resource Allocation (Default bearer establishment)**	Allocates resources for default bearer during initial UE attachment.	Same as NR-NR architecture.
7.	**Default bearer for IMS call**	Default bearer is established in gNB-UP	Default bearer is established in eNB.
8.	**Type of Core Network**	Strictly the 5GCN.	Uses the 4G EPS.
9.	**UE Cell Selection Procedure**	The UE shall select the strongest received signal for both gNB-CP and gNB-UP.	The UE selects the strongest received signal from both the eNB and the gNB bands.
10.	**Xn-Interface**	Uses the Xn Interface to connect between two gNBs.	Uses the X_2 Interface to connect between gNB and eNB.
11.	**CUPS Principle**	Strict separation of CP and UP in both RAN and CN.	Only applicable in EPC.

12.	**Connection with Legacy base station**	No inter-connection with LTE; Strictly NR.	Inter-Connection with other technology such as 4G LTE.
13.	**IMS Support**	Supports IMS Network.	Also, Supports IMS Network.
14.	**Dual Connectivity**	Supports DC.	Supports DC.
15.	**UE Computing Resources**	UE only supports NR protocol stack configuration. (lesser computing resources).	UE supports both LTE and NR protocol stack configuration (more computing resources).
16.	**Cost of Deployment**	Might be expensive as a result of the new 5GCN.	It is cheaper (gNB is only deployed alongside the existing LTE. In so doing expedites the deployment of 5G network).
17	**Deployment Option**	For HetNet deployment specifically for densely distributed small-cell environment.	For HetNet deployment.
18	**Ease of Management**	The decouple of UP and CP needs good coordination to achieve optimal network functionality.	Complex to manage because of two RATs being involved.
19	**EPS Fallback**	Not Supported yet.	Supported; call established via the EPS of the 4G LTE for IMS call.
20	**Services Offered**	Supports all cases including MTC, eMBB and URLLC because of the new 5GCN.	Supports only eMBB since it utilizes the EPC.
21	**Service Dependence**	Does not depend on the LTE coverage	Depends on the LTE network for coverage
22	**Deployment Time Frame**	Designed to be the future of 5G network framework (2024 and beyond).	Designed for immediate deployment of 5G network (2020).

5.5 Envisaged Setbacks

One major drawback envisaged in the NR-NR architecture is that a gNB-UP must always be in the coverage area of gNB-CP. Hence, the NR-NR architecture is not suitable for a homogenous deployment. This means that by decoupling the UP of gNB-CP makes the architecture not suitable for UEs that are configured for mono-

connectivity. Secondly, we foresee that the NR-NR architecture has the possibility of increasing the signal load, that is the total number of control messages transmitted is more when compared to the NR architecture but then the benefits of a complete split CP/UP in the RAN is sacrificed. This is because the NR-NR uses two base stations and excellent harmonization between the two base stations is needed to achieve the most appropriate network functionality. Finally, the NR-NR architecture presented is an extra option to the already existing NR architecture and not an alternative and hence, network operators might be slow to embrace this extra option.

Despite these setbacks, we see potential in the NR-NR architecture because DC in the NR-NR architecture is as a result of the introduction CUPS in the RAN, but since the NR-NR architecture uses both the NR and the 5GCN, the UE experiences 5G data rates. In addition, 5GCN together with NR is better optimized for IMS call. To further boost the UE throughput, the UE shall also be configured for Multi Connectivity (MC) in order to receive data simultaneously from multiple gNB-UPs, something especially important for cell edge users and for avoidance of inter cell interference. Hence, the NR-NR standalone architecture provides the benefits of 5G network consisting of the 5GCN with a complete CP/UP split in the RAN.

CHAPTER 6

CONCLUSION AND FUTURE RESEARCH

This chapter presents the conclusions of the research study and also provides recommendations for future research. The chapter is divided into two sections where Section 6.1 presents the summary and conclusions of the research study and Section 6.2 provides recommendations and directions for future work.

6.1 Summary and Conclusion

In this research study, an enhanced architecture for the standalone 5G network called NR-NR architecture was presented which decouples the CP and UP by introducing the principle of CUPS in the 5G RAN up to the UE. The proposed NR-NR architecture is comprising of two gNBs in DC mode. One of the gNBs, called the gNB-CP, provides coverage and handles all control signalling while the other gNB, called the gNB-UP, is dedicated for the transmission of UE data. The gNB-CP architecture of the NR-NR architecture is described including SRB configurations, the gNB-UP architecture and all DRBs configurations enabled in the architecture. This work also describes how a PDU session is established in the proposed architecture for a UE configured in the DC mode and how UE mobility is also handled.

To analyse how the proposed architecture handles control signalling, an IMS Voice call is chosen as an application for the research study. Two cases where considered; the first case analysed the control signalling involved in starting up an IMS voice call (VoNR) and the second case analysed the control signalling involved for the handover procedure during an active IMS call. Two mobility procedures were discussed. The first case considers the situation when the UE changes gNB-UP within the coverage area of gNB-CP. The second case considers the situation when the UE moves out of coverage region of the gNB-CP. A comparative study was carried out between the NR-NR architecture and the NSA based on the overall architectural features, control signalling messages procedures for starting up an IMS voice call and control signalling messages procedures for handover procedures when the UE moves from one gNB-CP cell to another gNB-CP cell.

The NR-NR architecture proposed in the research study is able to decouple the UP data from CP signalling in the RAN. By decoupling the 5G RAN allows for the optimization of the UP functions since the UP data traffic are no longer tightly coupled with the control signalling. The decoupling was accomplished by designing the architecture with two gNBs and with UE in DC mode. Besides, the

decouple allows the scaling of each plane's resources. The gNB-CP is carrying all control signals, and we foresee there may be a possibility of an increase in signalling load in the NR-NR architecture as a result of the optimal co-ordination between gNB-CP and gNB-UP needed to ensure excellent network functionalities but with the benefits of fully decoupled CP/UP split in the 5G RAN.

Furthermore, for all handover procedures considered, the gNB-CP initiates and handles all the control signalling messages needed to successfully handover the UE call. For the handover involving change of gNB-CP, the VoNR call is continuously ensured by the NR-NR architecture by forwarding the voice traffic directly from the serving gNB-UP to the target gNB-UP without the need of temporarily suspending the VoNR call as in the case of the NSA. Finally, the NR-NR architecture only allows for the NR protocol stack to be configured per UE thereby eliminating the need for UE additional computing resources which might save UE energy.

6.2 Future Research

To successfully design a network architecture is very challenging and time demanding just like any other research study. It requires passion, patience and in-depth research study. The conducted research has inspired the following research areas or topics for future work.

- **Performance Analysis between NR-NR Architecture and NR Architecture**
  The NR-NR architecture proposed was compared to the NSA based on the observed architectural differences. The natural evolution foreseen or anticipated is the progression from the NSA to the NR architecture and then to the NR-NR architecture presented in this research study. Therefore, it becomes natural to evaluate the performance gain between the NR-NR architecture and the NR architecture. It is highly recommended for future study to carry out a comparative study in terms of signalling load, throughput, latency, mobility and so on.
- **Interfacing NR-NR Architecture with NR Architecture**
  Recall that the NR-NR Architecture designed is targeted for a non-uniform densely populated small cell environment were UE is configured in DC mode. This is not generally the case since there are some UEs that are not configured for DC especially for a sparsely populated environment. It does not make so much sense to deploy the NR-NR architecture since data traffic in a sparsely populated environment is quite low. Furthermore, to investigate how a full CP/UP split in the RAN such as the NR-NR architecture be interconnected to NR architecture serving a sparsely populated environment to perform procedure like mobility. In other words, it is recommended for future work to look into the design principles of NR-NR architecture to investigate how the architecture can be modified or enhanced to support UEs that are not configured or supported for DC.
- **Designing a Network Slice for a NR-NR Architecture**

Decoupling the CP and UP in the 5G RAN is one approach we anticipate may facilitate the concept of network slicing. It will be interesting to explore and investigate how different gNB-UPs can be supported to dedicatedly handle specific network slices for URLLC applications such as self-driving cars, AR and robots used in industrial automation. Since these applications require stringent latency requirement, scaling CP resources of the gNB-CP accordingly might help to meet the URLLC requirement target. It is interesting to investigate if by having a dedicated gNB-UP for URLLC will reduce latency significantly.

- **Frequency Sharing in NR-NR Architecture**
  In the course of designing the NR-NR architecture, it is designed that both gNB-CP and gNB-UP use different carrier frequencies for operation. We predict that the ultimate solution could be for a frequency sharing between gNB-CP and gNB-UP using the Almost Black Subframe (ABS) [55] to achieve frequency sharing. We recommend that how ABS assignment is done in LTE should be extensively studied in order to draft the workings and apply the technique for an enhanced NR-NR architecture.
- **Latency Analysis of IMS Call During Active Multiple Applications**
  It is suggested to carry an investigative study to analyse how multiple ongoing applications in a UE affects the UP latency of an IMS voice call and specifically for a VoNR call. To this effect, it is suggested to use network emulators that can emulate the 5G NR and 5GCN architecture in order to carry out such latency analysis. In addition, there are quite a few available commercial phones yet for 5G network to carry out such testing. Building a network simulator tool for 5G RAN and 5GCN with a CP/UP split needs modification of some of the already available tools such as NS3. Secondly, the non-availability of 5G test network in TU Delft could also pose a challenge but it is believed that it will be deployed soon. Hence, it is recommended to look for other companies offering 5G test work like Ericsson and so on for the research work.
- **EPS Fallback in NR-NR Architecture for UE not Configured for NR Capabilities**
  Finally, it is also recommenced to investigate how EPS fallback can be performed in the NR-NR architecture for the case a UE is not configured for NR capabilities or for the case where 5GCN is temporarily unable to establish a VoNR call. EPS fall back refers to the procedures on how a VoNR can be established via the EPS of the LTE [54]. Also, how migration from EPS fallback to VoNR [56] shall be handled in the NR-NR Architecture is also a good case study. To further understand how the procedures shall be worked out, it is recommended to first understand how the NR-NR architecture can be interconnected with EPS of the 4G network. It is worth mentioning that EPS fallback is imperative to avoid call establishment failure when UE is trying to perform a VoNR call.

- **Prediction or Anticipatory User Plane Management**
  Another approach for optimization for a densely populated environment with regard to mobility is through anticipatory UP management which can be used to reduce the UP-reconfiguration cost generated as a result of UPF reselection [58]. This challenge may benefit from deep-learning algorithms and SDN that utilizes the prediction of user's mobility behaviour over time to predict the UE's future mobility for improved post-handover process.

REFERENCES

[1] A. Mohamed, O. Onireti, M. A. Imran, and R. Tafazoli, "Control-Data Separation Architecture for Cellular Radio Access Networks: A Survey and Outlook", in IEEE Communication Surveys & Tutorials, June 2015.

[2] P. Rost, C. J. Bernardos, A. De Domenico, M. Lalam, D. Sabella, and D. Wubben, "Cloud Technologies for Flexible 5G Radio Access Networks", in IEEE Access, May 2014.

[3] O. Chabbouh and S. B. Rejeb, "Cloud RAN Architecture Model based upon Flexible RAN Functionalities Split for 5G Networks", in Advanced Information Networking and Application Workshops, June 2017.

[4] Ericson Mobility Report, November 2018.

[5] 3GPP TR 23.794, Release 17, December 2019.

[6] I. Parvez, A. Rahmati, I. Guvenc, A. Sarwat and H. Dai, "A survey on Low Latency Towards 5G: RAN, Core Network and Caching Solutions", in IEEE Communications Surveys & Tutorials, May 2018.

[7] P. Schmitt, B. Landais and F. Yong Yang, "Control and User Plane Separation of EPC nodes (CUPS)", in 3GPP News Letter, June 2017.

[8] H. Venkatarman and R. Trestain,"5G Radio Access Networks: Centralized RAN, Cloud RAN, and Virtualization of Small Cells", in IEEE Access, August 2017.

[9] K. Yu, T. Kurita, Q. Hua, T. Sato, T. Asami, and V. Kafle, "Information Centric Networking: Research and Standardization Status", in IEEE Access. August 2019.

[10] X. Lin, M. Samaka, A. H. Chan and R. Jain, "Network Slicing for 5G: Challenges and Opportunities" in IEEE Internet Computing, August 2018.

[11] S. Ahmadi, "5G New Radio, Architecture, Technology, Implementation, and Operation of 3GPP New Radio Standards" in Elsevier Inc, July 2019.

[12] X. Wang, C. Cavdar, L. Wag, M. Tornatore, and H. Lee, "Virtualized Cloud Radio Access Network for 5G Transport", in IEEE Communications Magazine, September 2017.

[13] K. Rabie, "Core Network Evolution- How Cups Changed the Call Flow" in Netmanias Technical documents, April 2018.

[14] 5G PPP Architecture Working Group, "View on 5G Architecture", Version 2.0, December 2017.

[15] P. Gupta, "5G Deployment Options", 18th August 2019. [Online]. Available: https://www.linkedin.com/pulse/5g-deployment-options-pallab-gupta/. [Accessed 20th July 2020].

[16] A. Aubida and M. Jasmin., "LTE Heterogenous Network: A case study ", in IEEE Journal of Computer Applications, January 2013.

[17] B. Svetlana, and L. Wolfgang, "Heterogenous Wireless Network Selection: Load balancing and Multicast Scenarios", in IEEE International Journal on Advanced Networks and Scenarios, December 2013.

[18] P. Anorld, N. Bayer, J. Belschner, and G. Zimmermann, "5G Radio Access Network Architecture Based on Flexible Functional Control/User Plane Splits" in European Conference on Networks and Communication, June 2017.

[19] 3GPP TS 29.274, "3GPP Evolved Packet System (EPS); Evolved General Packet Radio Service (GPRS) Tunnelling Protocol for Control Plane (GTPv2-C)", Release 16, March 2020.

[20] H. Ali-Ahmed "An SDN approach for DenseNets", in Euro Workshop on Software Defined Network, October 2013.

[21] J. Bartelt, N. Vucic, D. Camp-Mur, and E. Grass, "5G Transport Network Requirement for the Next Generation Fronthaul Interface", in EUARSIP Journal on Wireless Communication and Networking, May 2017.

[22] 5G Oriented OTN Technology, a White Paper by NGOF, March 2018.

[23] R. Guerzoni, R. Trivisonno, and D. Soldani, "SDN-Based Architecture and Procedures for 5G Networks", in 1st International Conference on 5G for Ubiquitous Connectivity (5DU), November 2014.

[24] L. Du, J. Chain, T. Yu, and X. Liu, "C/U split Multi-connectivity in the Next Generation New Radio System", in IEEE Access, November 2017.

[25] Huawei, Nokia and Ericsson, "The eCPRI Specification" Version 1.0 [Online]. Available: http:/www.cpri.info/ August 2017. [Accessed 13th March 2020].

[26] H. Rashmi and G. Ranjani, "5G New Radio & Cloud Radio Access Network", in International Journal of Engineering Research and Technology, May 2019.

[27] X. Xu, H. Gaoining, and S. Zhang "On Functionality Separation for Green Mobile Networks", in IEEE Communication Magazine, May 2013.

[28] N. Bhusjanetal, "Network Densification: The Dominant Theme for Wireless Evolution into 5G", in IEEE Communication Magazine, Feb 2014.

[29] Calnex Article "eCPRI and Networking the Fronthaul" [Online]. Available: https://www.calnexsol.com/en/article-display/114-archived-blog/870-ecpri-and-networking-the-fronthaul May 2018. [Accessed 20th March 2020].

[30] B. Bertenyi, R. Burbidge, G. Masini and Y. Gao, "NG Radio Access Network (NG-RAN)", in Journal of ICT, April 2018.

[31] 3GPP TS 38.420, "Xn General Aspects and Principles", Release 15, January 2018.

[32] 3GPP TS 38.145, "PDU Session User Plane Protocol", Release 15, March 2018.

[33] S. Homma, and T. Miyasaka, "User Plane Protocol and Architectural Analysis on 3GPP 5G System Draft", in IETF Internet Draft, November 2019.

[34] Techplayon: "Deployment Scenarios for 5G NR" [Online]. Available: http://www.techplayon.com/deployments-scenarios-for-5g-nr/ September 2017. [Accessed 5th June 2020].

[35] Ericsson, Huawei, and Nokia, "Common Public Radio Interface Specification V 7.0", [Online] Available: http://www.cpri.info 2015. [Accessed 30th March 2020]

[36] C. Simon, and M. Maliosz, "Design aspects of Low Latency Services with Time Sensitive Magazine", in IEEE Communications Standards Magazine, June 2018.

[37] ITU Technical Report, "Transport Network Support of IMT-2020/5G" in GSTR-TN5G, February 2018.

[38] G. Brown, "New Transport Network Architecture for 5G RAN", A White Paper by Fujitsu, June 2018.

[39] P. Shepherd in LinkedIn, "Learn QoS of 5G Networks" [Online]. Available: https://www.linkedin.com/pulse/learn-qos-5g-networks-paul-shepherd/. June 2019. [Accessed 19th April 2020].

[40] M. Mezzavilla, M. Polse, A. Zanella, A. Dhanajay, T. Rapapport and S. Rangan "Public Safety Communication above the 6GHz: Challenges & Opportunities", in IEEE Access, November 2017.

[41] T. Uchino, K. Kai, H. Takahashi, "Specifications of NR Higher Layer in 5G", in Technology Report on special Articles on Release 15 standardization, October 2018.

[42] Netmanias Artcile: "The 5G QoS", [Online]. Available: https://netmanias.com/en/post/oneshot/14105/5g/5g-qos Feb 2019. [Accessed 19th April 2020].

[43] 3GPP, "Control and User Plane Separation" [Online]. Available: https://www.3gpp.org/news-events/1882-cups, June 2018. [Accessed 13th March 2020].

[44] Huawei Technologies Co Ltd, "Vo5G Technical White Report Paper", July 2018.

[45] 3GPP TS 38.804, "Radio Interface Protocol Aspects"', Release 14, March 2017.

[46] NFON: The IMS System [Online] Available: https://www.nfon.com/en/service/knowledge-base/knowledge-base-detail/ims [Accessed on 14th April 2020].

[47] A. Eliasha, M. Yahia, and A. Mohamed, "Practical Guide to LTE-A, VoLTE and IoT: Paving the way towards 5G", Wiley, July 2018.

[48] R. Agrawal, A. Bedekar, T. Kolding and V. Ram "Cloud RAN Challenges and Solutions", in Innovations in Cloud, Internet and Networks Conference, March 2016.

[49] R. Noldus, U. Olsson, A. Ryde, C. Mulligan, M. Stille and I. Fikouras "IMS Application Developer's Handbook: Creating and Deploying Innovative IMS Applications", Academic Press, November 2016.

[50] ETSI Standard, ES 282 001, "Protocols for Advanced Networking", Release 1, 2005-2006.

[51] H. Klifi and J. Gregoire, "IMS Application Services, Roles, Requirements and Implementation Technology", in IEEE Computer Society, June 2018.

[52] EventHelix, "5G Non-Standalone Access", [Online]. Available: https://medium.com/5g-nr/5g-non-standalone-access-4d48cea5db5f. March 2019. [Accessed 2nd June 2020].

[53] K. Zeman, and Z. Tunka," User Performance gains by data Offloading of LTE Mobile Traffic into Unlicensed IEEE 802.11 Links", in International Conference on Telecom and Signal Processing, July 2015.

[54] Award Solution Online Course: "5G Voice Solutions – VoNR and EPS Fallback" [Online]. Available : https://www.awardsolutions.com/portal/ilt/5g-voice-solutions-vonr-and-eps-fallback June 2018. [Accessed 18th July 2020].

[55] F. Alfarhan, R. Lerbour, and Y. L. Helloco, "An Optimization Framework for LTE eICIC and Reduced Power eICIC", in IEEE Global Communication Conference, February 2016.

[56] R. Keller, "Migration from EPS fallback to support Voice in 5G", [Online]. Available: https://www.ericsson.com/en/blog/2018/10/migration-from-eps-fallback-to-support-voice-in-5g. October, 2018. [Accessed 3rd June 2020].

[57] TechPlayon,"5G NR gNB Higher Layer Split (HLS)", [Online]. Available: http://www.techplayon.com/5g-nr-gnb-higher-layer-split-hls/. April 2019. [Accessed 10th May 2020].

[58] S. Peters and M. A. Khan, “Anticipatory User Plane Management for 5G”, in IEEE 8th International Symposium on Cloud and Service Computing (SC2), July

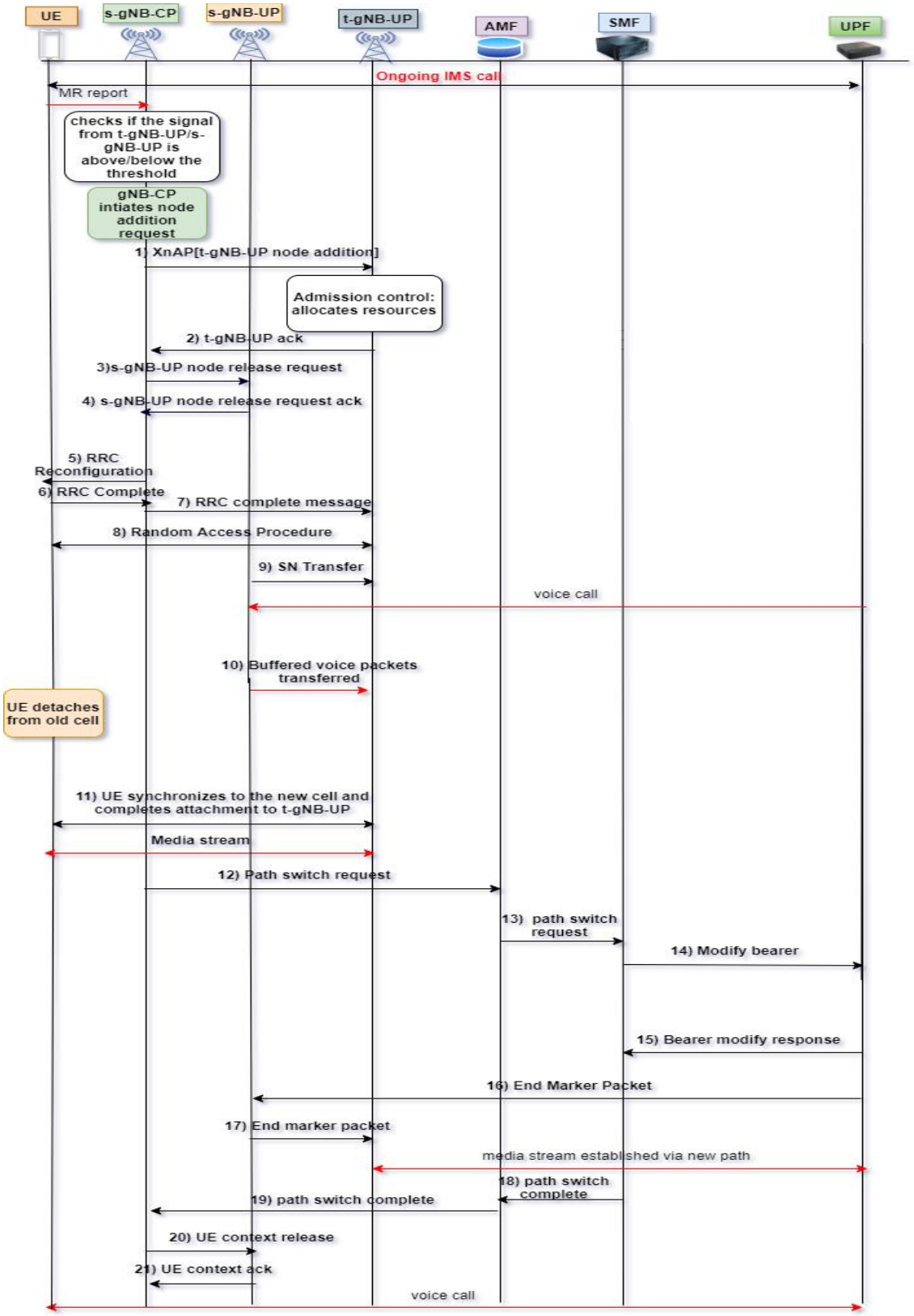
ABBREVIATIONS

3GPP	3rd Generation Partnership Project
4G	4th Generation
5G PPP	5G Infrastructure Public Private Partnership
5G	5th Generation
5QI	5G QoS Indicator
ABS	Almost Blank Subframe
AGW	Access Gateway
AKA	Authentication and Key Agreement
AMF	Access and Mobility Function
AP	Application Protocol
API	Application Program Interface
AR	Augmented Reality
ARQ	Automatic Repeat Request
AS	Application Server
BBU	Baseband Unit
BS	Base Station
CA	Carrier Aggregation
CAN	Carrier Access Network
CDSA	Control-Data Separation Architecture
CN	Core Network
CoMP	Coordinated Multipoint
CP	Control Plane
CPRI	Common Public Radio Interface
C-RAN	Centralized Radio Access Network
CSI	Channel State Information
CU	Centralized Unit
CUPS	Control and User Plane Separation
DC	Dual Connectivity
DL	Downlink
DN	Data Network
DNS	Domain Name Server
D-RAN	Distributed Radio Access Network
DRB	Data Radio Bearer
DU	Distributed Unit
eCPRI	enhanced Common Public Radio Interface
eMBB	enhance Mobile Broadband
eNB	Evolved NodeB
EN-DC	E-UTRAN New Radio-Dual Connectivity
EPC	Evolved Packet Core
EPS	Evolved Packet System
E-UTRAN	Evolved Universal Terrestrial Radio Access Network
FRAN	Fog Radio Access Network

gNB	new Generation NodeB
GTP	GPRS Tunneling Protocol
GW	Gateway
HARQ	Hybrid Automatic Repeat Request
HetNet	Heterogenous Network
HSS	Home Subscriber Server
HTTP	Hyper-Text Transport Protocol
ICIC	Inter-Cell Interference Coordination
I-CSCF	Interrogating-Call State Control Function
IMS	IP Multimedia Subsystem
IoT	Internet of Things
IP	Internet Protocol
LTE	Long Term Evolution
MAC	Medium Access Control
MC	Multi Connectivity
MCS	Modulation and Coding Scheme
MEC	Mobile Edge Computing
MIMO	Multiple Input Multiple Output
MME	Mobility and Management Entity
MNO	Mobile Network Operators
MR	Mobility Report
MRF	Mobile Resource Function
MTC	Machine-Type Communication
MU-MIMO	Massive User- Multiple Input Multiple Output
NAS	Non-Access Stratum
NF	Network Function
NG	Next Generation
NR	New Radio
OSI	Open System Interconnection
PCC	Policy Charging and Control
PCF	Policy and Control Function
PCM	Pulse Code Modulation
PCRF	Policy and Charging Rules Function
P-CSCF	Proxy- Call State Control Function
PDCP	Packet Data Convergence Protocol
PDN	Packet Data Network
PDU	Protocol Data Unit
P-GW	PDN Gateway
QFI	QoS Flow Indicator
QoE	Quality of Experience
QoS	Quality of Service
RAN	Radio Access Network
RAT	Radio Access Technology
RE	Radio Equipment
REC	Radio Equipment Control

RF	Radio Frequency
RLC	Radio Link Control
RRC	Radio Resource Control
RRH	Remote Radio Head
RRU	Remote Radio Unit
RSS	Received Signal Strength
RTCP	Real Time Control Protocol
RTP	Real Time Protocol
SBG	Session Border Gateway
S-CSCF	Serving-Call State Control Function
SCTP	Stream Control Transport Protocol
SDAP	Service Data Adaptation Protocol
SDN	Software Defined Networking
S-GW	Serving Gateway
SI	Signal Indicator
SIP	Session Initiation Protocol
SMF	Session Management Function
SON	Self Optimized Network
SRB	Signal Radio Bearers
SS-RSRP	Synchronization Signal Reference Signal Received Power
TCP	Transmission Control Protocol
TEID	Tunnel Endpoint Identifier
TLN	Transport Layer Network
UDM	Unified Data Management
UDN	Ultra Dense Network
UDP	User Datagram Protocol
UE	User Equipment
UP	User Plane
UPF	User Plane Function
URLLC	Ultra-Reliable Low Latency Communication
VNF	Virtual Network Function
VoIP	Voice over IP
VoLTE	Voice over LTE
VoNR	Voice over New Radio
VR	Virtual Reality

APPENDIX A HADOVER SIGNALLING



A1: Handover Signalling when UE is within the coverage of gNB-CP.