UNIVERSITY OF THESSALY
SCHOOL OF ENGINEERING
DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

Design and implementation of mechanisms and policies for handover control in 5G wireless networks, using Machine Learning

Diploma Thesis

Dimitrios Kefalas

Supervisor: Athanasios Korakis

September 2022

UNIVERSITY OF THESSALY
SCHOOL OF ENGINEERING
DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

Design and implementation of mechanisms and policies for handover control in 5G wireless networks, using Machine Learning

Diploma Thesis

Dimitrios Kefalas

Supervisor: Athanasios Korakis

September 2022

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ
ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Σχεδιασμός και υλοποίηση μηχανισμών και πολιτικών ελέγχου μεταπομπής ανάμεσα σε σταθμούς βάσης 5ης γενιάς με χρήση τεχνικών μηχανικής μάθησης

Διπλωματική Εργασία

Δημήτριος Κεφαλάς

Επιβλέπων: Αθανάσιος Κοράκης

Σεπτέμβριος 2022

Approved by the Examination Committee:

Supervisor	**Athanasios Korakis** Professor, Department of Electrical and Computer Engineering, University of Thessaly
Member	**Antonios Argyriou** Associate professor, Department of Electrical and Computer Engineering, University of Thessaly
Member	**Dimitrios Bargiotas** Professor, Department of Electrical and Computer Engineering, University of Thessaly

Acknowledgements

This thesis is a tribute to everyone who helped me finish my studies and supported me through this beautiful 5-year journey.

First, I want to express my gratitude to my parents Giorgos and Spyridoula, my sister Eleftheria and niece Katerina, for their unwavering support and encouragement throughout my life.

I would also like to thank all my childhood and university mates for understanding and supporting me during my studies, hoping to repay them.

I want to thank my supervisor, Professor Athanasios Korakis for giving me the opportunity to come into contact with a research environment dealing with state-of-the-art projects, inspiring me to work in this field.

As my mentor, I feel blessed to have the Senior Researcher and Postdoctoral Nikolaos Makris, who was always available, guiding and feeding me with valuable advice and helping me to overcome the impasses I encountered in my projects.

Also, I am profoundly grateful to Professor Dimitrios Bargiotas and Associate Professor Antonios Argyriou for being a member of the examination committee of my thesis.

DISCLAIMER ON ACADEMIC ETHICS AND INTELLECTUAL PROPERTY RIGHTS

«Being fully aware of the implications of copyright laws, I expressly state that this diploma thesis, as well as the electronic files and source codes developed or modified in the course of this thesis, are solely the product of my personal work and do not infringe any rights of intellectual property, personality and personal data of third parties, do not contain work / contributions of third parties for which the permission of the authors / beneficiaries is required and are not a product of partial or complete plagiarism, while the sources used are limited to the bibliographic references only and meet the rules of scientific citing. The points where I have used ideas, text, files and / or sources of other authors are clearly mentioned in the text with the appropriate citation and the relevant complete reference is included in the bibliographic references section. I also declare that the results of the work have not been used to obtain another degree. I fully, individually and personally undertake all legal and administrative consequences that may arise in the event that it is proven, in the course of time, that this thesis or part of it does not belong to me because it is a product of plagiarism».

The declarant

Dimitrios Kefalas

Diploma Thesis

Design and implementation of mechanisms and policies for handover control in 5G wireless networks, using Machine Learning

Dimitrios Kefalas

Abstract

The fifth generation (5G) mobile network crosses over with the fourth generation (4G) network due to the exponential growth in mobile users. A key characteristic of the envisioned dense deployments is that the foreseen number of handovers (HO) scenarios and HO rates will significantly grow. The mobility of user equipment (UE) poses a significant challenge to maintaining a stable and dependable connection in current and future mobile networks. Mobile subscriber access to cellular networks and the accompanying traffic load are erratic, frequently out of balance, and uneven over time, resulting in unequal cell loads. While many UEs are connected in some cells overloading and congesting the networks, there are also cells with fewer UEs having underutilized resources. A handover load balancing policy promises to distribute clients across the cells to succeed in better network resource utilization. Towards contributing to the aforementioned challenges, in this thesis, three load-balancing policies are designed based on different criteria, such as the count of serving clients per cell (Round Robin) and the total uplink clients demand per cell. To give flexibility to the designed policies, we use Machine Learning methods to forecast each client's uplink demands based on a dataset collected from real large-scale systems. To validate and evaluate our work in a real environment, we utilize the NITOS Testbed to deploy a monolithic software-based LTE network. This architecture uses the nextEPC for the Core Network, the OpenAirInterface for the Radio Access Network, and FlexRAN for a Radio controller to get the corresponding metrics while static nodes are used as UEs. Also, a third-party application was developed to automate the handover decision and send handover commands to the LTE network. All the experiments happened in real-world environments without using any simulation tool. Our results denote that resources can be efficiently distributed among available cells, while the near-future traffic can be effectively forecast using our Machine Learning approach.

Keywords:

load balancing (LB), handover (HO), fourth generation (4G), fifth generation (5G)

Διπλωματική Εργασία

Σχεδιασμός και υλοποίηση μηχανισμών και πολιτικών ελέγχου μεταπομπής ανάμεσα σε σταθμούς βάσης 5ης γενιάς με χρήση τεχνικών μηχανικής μάθησης

Δημήτριος Κεφαλάς

Περίληψη

Το δίκτυο κινητής τηλεφωνίας πέμπτης γενιάς (5G) θα αντικαταστήσει το δίκτυο τέταρτης γενιάς (4G) λόγω της εκθετικής αύξησης των χρηστών κινητής τηλεφωνίας και των απαιτήσεων που δημιουργούνται. Τόσο ο αριθμός των σεναρίων όσο και τα επιτυχή ποσοστά μεταπομπής θα αυξηθούν σημαντικά. Η κινητικότητα των χρηστών θα αποτελέσει σημαντική πρόκληση για τη διατήρηση μιας σταθερής και αξιόπιστης σύνδεσης στα επερχόμενα δίκτυα κινητής τηλεφωνίας. Η πρόσβαση και οι απαιτήσεις των χρηστών των κυψελωτών δικτύων κινητής τηλεφωνίας συνεπάγονται φόρτωση του δικτύου, εξαιτίας των μεταδόσεων, οι οποίες δεν είναι πανομοιότυπες μεταξύ τους, έχουν διαφορετικές απαιτήσεις τόσο στον όγκο δεδομένων προς μεταφορά, όσο και στην ποιότητα εξυπηρέτησης τους, με αποτέλεσμα την τυχαία και συνήθως άνιση κατανομή των φορτίων στις κυψέλες. Ενώ πολλά UE είναι συνδεδεμένα σε ορισμένα κελιά τα οποία υπερφορτώνουν και στα οποία προκαλούν συμφόρηση, υπάρχουν επίσης κελιά με λιγότερους συνδεδεμένους συνδρομητές που υποχρησιμοποιούν τους διαθέσιμους πόρους τους. Μια πολιτική εξισορρόπησης φορτίου χρησιμοποιώντας μεταπομπές υπόσχεται να κατανείμει τους πελάτες σε όλες τις κυψέλες για να πετύχει καλύτερη χρήση των πόρων του δικτύου. Συμβάλλοντας στις προαναφερθείσες προκλήσεις, τρεις πολιτικές εξισορρόπησης φορτίου σχεδιάζονται με βάση διαφορετικά κριτήρια, όπως ο αριθμός των εξυπηρετούμενων πελατών ανά κελί (Round Robin) και η ταχύτητα μετάδοσης των πελατών στην ανοδική τους ζεύξη ανά κελί. Για να προσφέρουμε ευελιξία στις σχεδιασμένες πολιτικές, χρησιμοποιούμε μεθόδους μηχανικής εκμάθησης για να προβλέψουμε την ταχύτητα μετάδοσης των πελατών στην ανοδική τους ζεύξη με βάση ένα σύνολο δεδομένων που προέρχεται από πραγματικά συστήματα μεγάλης κλίμακας. Για να επικυρώσουμε και να αξιολογήσουμε την εργασία μας σε πραγματικό περιβάλλον, χρησιμοποιούμε το NITOS Testbed για να αναπτύξουμε ένα μονολιθικό βασισμένο σε λογισμικό δίκτυο LTE. Αυτή η αρχιτεκτονική χρησιμοποιεί το nextEPC για το Core Network, το OpenAirInterface για το

Radio Access Network και το FlexRAN ως διαχειριστή του συστήματος που θα λαμβάνει συγκεκριμένες μετρήσεις, ενώ οι στατικοί κόμβοι λειτουργούν ως UE. Επίσης, αναπτύχθηκε μια εφαρμογή ειδικά για τα πλαίσια της πτυχιακής μου εργασίας, με σκοπό την αυτοματοποίηση της απόφασης μεταπομπής και την αποστολή εντολών μεταπομπής στο δίκτυο LTE. Όλα τα πειράματα έγιναν σε περιβάλλοντα πραγματικού κόσμου χωρίς τη χρήση εργαλείων προσομοίωσης. Τα αποτελέσματα μας υποδηλώνουν ότι οι πόροι μπορούν να κατανεμηθούν αποτελεσματικά μεταξύ των διαθέσιμων κυψελών, καθώς η ταχύτητα μετάδοσης των πελατών στην ανοδική τους ζεύξη στο εγγύς μέλλον μπορεί να προβλεφθεί αποτελεσματικά χρησιμοποιώντας την προσέγγιση Machine Learning.

Table of contents

List of figures

Abbreviations

3GPP	3rd Generation Partnership Project
3G	3rd Generation
4G	4th Generation
5G NR	5G New Radio
5G	5th Generation
AF	Application Function
AKA	Authentication and Key Agreement
AMF	Access and Mobility Management Function
AUSF	Authentication Server Function
CQI	Channel Quality Indicator
EPC	Evolved Packet Core
EPS	Evolved Packet System
EUTRAN	Evolved Universal Terrestrial Radio Access Network
FDMA	Frequency-Division Multiple Access
GPRS	General Packet Radio Service
GSM	Global System for Mobile communication
HSS	Home Subscriber Server
IMSI	International Mobile Subscriber Identity
HO	Handover
IP	Internet Protocol
LAN	Local Area Network
LTE-A	Long Term Evolution Advanced
LTE	Long Term Evolution
MAC	Media Access Control
MIMO	Multiple Input Multiple Output

MME	Mobility Management Entity
Massive MIMO	Massive Multiple Input Multiple Output
NITOS	Network Implementation using Open-Source software
NRF	Network Repository Function
NSSF	Network Slice Selection Function
NOMA	Non-orthogonal multiple access
NWDAF	Network Data Analytic Function
OAI	Open Air Interface
OFDMA	Orthogonal Frequency-Division Multiple Access
OMF	Control and Management Framework
OS	Operating System
P-GW	Packet Data Network Gateway
PCF	Policy Control Function
PDCP	Packet Data Convergence Protocol
PDN	Packet Data Network
PHY	Physical Layer
QoS	Quality of Service
RAN	Radio Access Network
RF	Radio Frequency
RLC	Radio Link Control
RRC	Radio Resource Control
RSSI	Received Signal Strength Indicator
RSRP	Reference Signal Received Power: RSRP
SGSN	Serving GPRS Support Node
S-GW	Serving Gateway
SDR	Software Defined Radio
SMF	Session Management Function
SPGW-C	Control Plane of the Packet Data Network Gateway
SPGW-U	User Plane of the Packet Data Network Gateway
TDMA	Time-Division Multiple Access
UDM	Unified Data Management
UE	User Equipment

UMTS	Universal Mobile Telecommunication System
UPF	User Plane Function
USRP	Universal Software Defined Radio Peripherals

Chapter 1

Introduction

1.1 Motivation

The continuous and exponential growth of mobile subscriptions has caused an enormous surge in mobile data traffic and the wireless arena. Studies show that by 2027 the number of mobile subscriptions is expected to exceed 4.4 billion. Coupled with today's demands of our modern civilization to create smart cities, autonomous vehicles, drones, augmented reality (AR), virtual reality (VR), real-time Broadcasting, real-time Gaming, and the internet of things (IoT), the amount of mobile data to transfer increases even more. This situation leads the scientific community to create capable technologies that satisfy these mobile data transmission needs, such as high bandwidth, low latency, reliable and secure connections, numerous connected devices, and others.

The evolution of cellular networks can be briefly described as the transition from the only analog systems without data capability (1G) and with digital circuit-switched systems (2G), to broadband and multimedia systems (3G) with an all-IP network revolution, unified IP with broadband data transmission and broadcasting, in addition to very high-vol- ume of voice users (4G), up to massive and seamless end-to-end connectivity, mobility, net- work virtualization and slicing (5G), the network evolution . This evolution of cellular networks from 1G to 5G reflects the scientific community's effort to meet the ever-growing network demands.

In contrast to the previous generations of cellular networks, which focused on network management, the 5th generation networks shifted their interest to service management. The potential of integration of Machine Learning (ML) and Artificial Intelligent (AI) in 5G network architecture brings a massive revolution in network performance and systems since new possibilities to design much more efficient policies for numerous 5G procedures are created. Also, the integration of a hybrid cellular base station architecture (macrocell, small cell and femtocell) in 5G networks brings an exponential growth in the number of base station, creates new possibilities and perspectives in client's handover [1].

1.2 Thesis subject

Two primary issues in 5G networks are resource and handover management when maintaining seamless connectivity in mobile environments. In this diploma thesis, we address these two issues together and suggest three distinct load-balancing policies based on a Round Robin, UE's Uplink Load, and predicted UE's Uplink Load balance schemes. These schemes introduce new handover optimization methods for the 5G cellular network without considering UE mobility. All the experiments and evaluation steps took place in real-world environment. The last scheme uses machine learning techniques to predict UE's Uplink-generated traffic, and we propose its integration into the latest 5G Network architecture.

1.2.1 Contribution of Thesis

In this thesis, we did a brief historical analysis of the history of cellular networks in order to comprehend the differences and changes in network architecture from 1G to 5G. The numerous handover designs, topologies, and protocols used in LTE systems were presented in advance. Then, a fully working monolithic LTE network is implemented in NITOS-testbed alongside a third-party application to automate the handover decision and send handover commands to the LTE network. A brief experimental handover analysis took place, and three handover mechanisms based on the load balance policy were introduced. Finally, we integrate our mechanism into 5G Release 15 and Release 16 of 3GPP Technical Specifications.

1.3 Content Organization of Thesis

The study is divided into eight chapters. First, the evolution of the cellular networks from 1G through 5G, emphasizing on 4G protocol stack structure and network architecture, are shown in Chapter 2. Chapter 3 presents a handover overview alongside the reasons to trigger a handover. Chapter 4 gives a small introduction to Machine Learning Theory, while Chapter 5 introduces the design of the handover policy used for this diploma thesis. Chapter 6 presents the infrastructure and all the software tools used to deploy our experimental scenarios. Chapter 7 introduces the experimental system architecture, a real-environment handover analysis, and the experimental results from the proposed load balance policies. Finally, the overall architecture and proposed policies and their integration into 5G New Radio (NR) Rel. 15 are discussed in Chapter 8.

Chapter 2

Cellular Networks

2.1 Background

The so-called last mile of the Internet is implemented through the access network, which includes the cellular network. A cellular network is a radio network that is distributed in the form of cells. Each cell has a base station that is placed in a fixed position and receives/transmits in a particular geographical area, having a specific radio network coverage across in cell from/to numerous mobile devices that are into it [2] as seen in Figure 2.1.

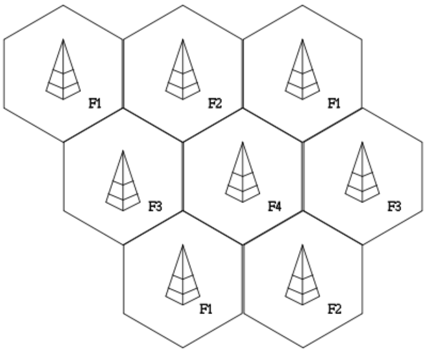


Figure 2.1: Cellular Network - BS topology

The area that is covered from each cell depends on many factors such as

- Transmission power of the base station
- Transmission power of the user equipment
- Base station antenna height and placement across the cell
- Base station antenna type

- Obstructing buildings in the cell

Although the traditional approach is that base stations are placed in the center of each cell, many systems nowadays locate base staions in corners where three cells intersect in order to a single BTS with directional antennas to coverage three cells (Figure 2.2).

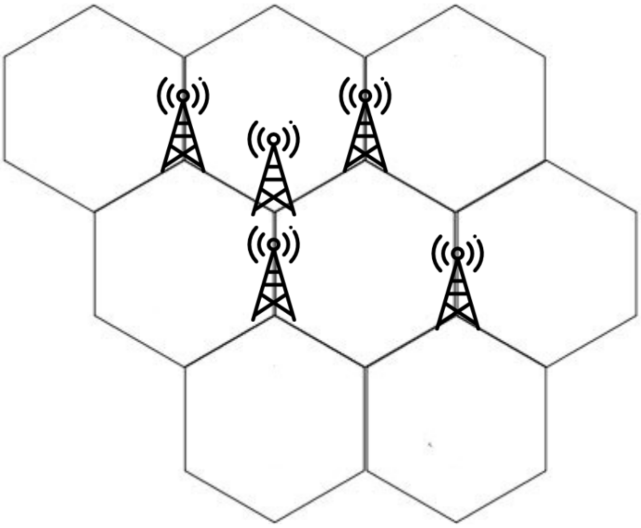
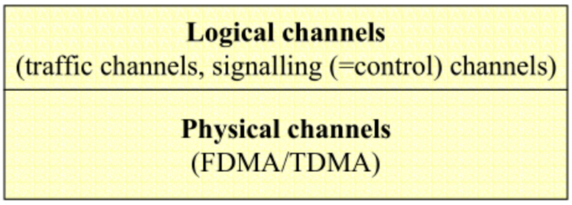


Figure 2.2: Cellular Network - BS placed at the corner of each cell

Before starting the presentation of the generations of cellular networks it is worth noting the differentiation of the physical and logical channels encountered in them. Physical channels deal with modulation, slot synchronization, multiple access, and others, whereas logical channels are built on top of physical channels and convey information transmitted between user equipment and the access network. These scheme is illustrated in Figure 2.3.



Logical channels
(traffic channels, signalling (=control) channels)

Physical channels
(FDMA/TDMA)

Figure 2.3: Logical and physical channels

2.2 1G & 2G

The first generation (1G) and the second (2G) of wireless cellular technology have different types of transmissions, analog and digital, respectively, but they share the same network architecture. 1G networks were analog FDMA systems intended only for voice communication; on the other hand, a combination of FDM and TDM having as operating area the band of 1.8 GHz is used as the air interface by the GSM standard for 2G cellular systems , providing additional services such as text messages (SMS), image messages (MMS), and multimedia messages (MMS).

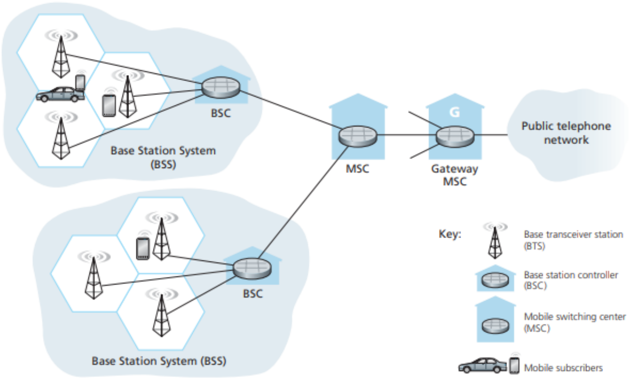


Figure 2.4: 2G network architecture

A GSM network's base station controller (BSC) typically manages a few tens of base stations. The BSC's responsibility is assigning mobile subscribers to BTS radio channels and performing paging and handovers. A GSM base station system (BSS) is made up of the base station controller and all base stations that are assigned. Following the network path of a GSM network, the mobile switching center (MSC) is responsible for user authorization and accounting, and gateway MSCs connect a cellular network to the larger public telephone network. These entitities compose the 2G network architecture as shown in Figure 2.4.

2.3 3G

The notable difference between the 3G and previous generations of cellular networks is the existence of a full TCP/IP protocol stack for cellular networks to connect with the Internet. The 3G standard analyzed in this section is the UMTS (Universal Mobile Telecommunications Service) 3G standard developed by the 3rd Generation Partnership project (3GPP).

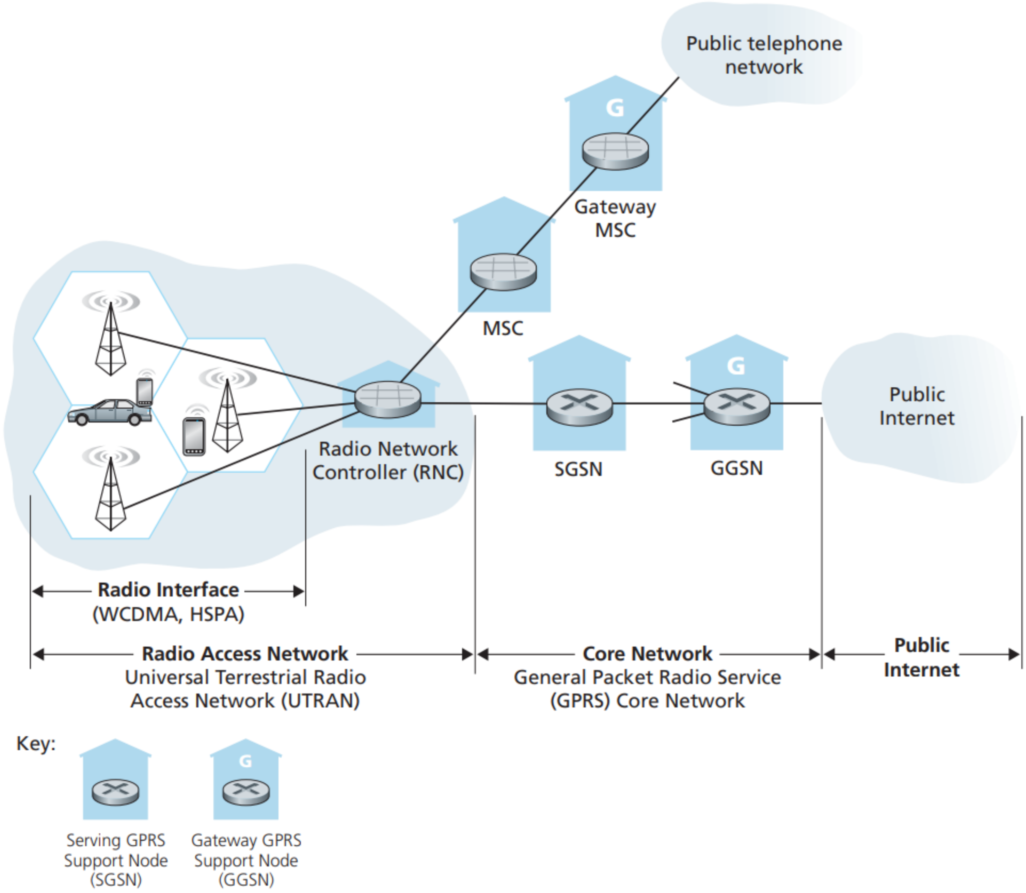


Figure 2.5: 3G system architecture

The 3G architecture is based on the 2G by inheriting the voice cellular network (MSC) and adding two new entities, the Universal Terrestrial Radio Access Network or simply Radio Access Network (RAN-UTRAN) and Core Network, to succeed internet connectivity (see Figure 2.5). These happened since researchers wanted to take advantage of the already existing infrastructure of the cellular networks and add additional cellular data functionality.

The core network is consisted by two entities, that of Serving GPRS Support Nodes (SGSNs) and Gateway GPRS Support Nodes (GGSNs).
SGSN has the following responsibilities:

- deliver datagrams from and to an UE that belongs to the connected radio access network
- interacts with the voice cellular network in order to authenticate users and perform handovers
- forwards datagrams from RAN to GGSN and vice versa

GGSN is the last piece of the 3G infrastructure with a gateway role connecting multiple SGSNs to the internet.

The radio access network (RAN) piece of a 3G system has three entities.

- User Equipment (UE): Any mobile or non-mobile device that can connect the user to the 3G network.
- Node-B: the name of BTS in 3G networks
- Radio Network Controller (RNC)

The Radio Network Controller (RNC) controls numerous BTS and is connected with MSC and SGSN to transmit voice via circuit-switched cellular voice network inherited from GSM and data via packet-switched Internet, respectively. The wireless link between Node BS and UEs operates the same way as in all cellular networks but uses a Direct Sequence Wideband CDMA (DS-WCDMA) scheme having as operating area the band of 2 GHz.

As mentioned above, for 3G to be able to transmit cellular data, a Radio interface protocol architecture would have to be created (see Figure 2.6).

Three protocol layers make up the radio interface:

- the physical layer (L1)
- the data link layer (L2)
  - Medium Access Control (MAC)
  - Radio Link Control (RLC)
  - Packet Data Convergence Protocol (PDCP)

  - Broadcast/Multicast Control (BMC)

- network layer (L3)

  - Radio Resource Control (RRC)

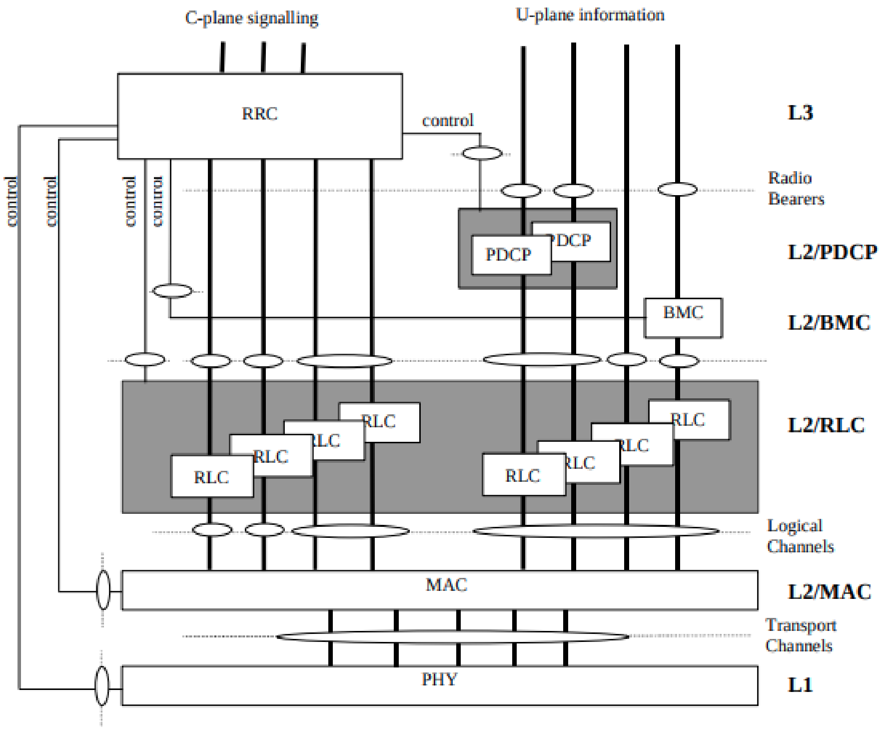


Figure 2.6: Radio Interface protocol architecture

Layer 3 and RLC are divided into Control (C-) and User (U-) planes, where the Control plane is separated into sublayers, with the lowest one (RRC) interacting with layer 2 and terminating in the UTRAN, in the U-plane PDCP and BMC exist.

The role of MAC, RLC, RRC, and PDCP will be described in detail in the 4G section. As for BMC offers a broadcast/multicast transmission function of common user data in an unacknowledged mode in the user plane on the radio interface.

2.4 4G

The use of the 3G network brought the possibility to browse the Internet and send and receive email, music, and video through UEs. Due to the high demand, 3G could not provide these resources since the peak speed of 14 Mbit/s was insufficient. The 3rd Generation

Partnership Project (3GPP) developed the 4th generation (4G) under the following principles [3]:

- It is necessary to maintain the 3G system's competitiveness in the long run.
- To satisfy user demands for higher data rates and quality of service.
- To create a Packet Switched optimised system with low complexity.
- Flexibility in frequency and bandwidth.

This new standard, named Long-Term Evolution (4G-LTE), sets peak requirements for data speed up to 100 Mbps in downlink (DL) with 50 Mbps in uplink (UL), operating at the band of 2-8 GHz having a channel's bandwidth of 100 MHz. Since LTE is a packet switch-oriented standard, the CN and RAN were designed purely IP based and bring notice difference in comparison with UMTS addressing UEs with IP, which outlines the transition from a cellular network capable of transmitting data packets (UMTS) to a completely oriented data-packet cellular network (LTE). In contrast to UMTS, where the UE is assigned an IP when a service is established and released when the service is terminated, in LTE, a UE is assigned an IP when switched on and released when switched off.

2.4.1 Orthogonal Frequency Division Multiple Access (OFDMA)

A revolutionary technology created to meet the needs of LTE is a new OFDMA modulation scheme that replaced the less efficient TDM, FDM, and DS-WCDMA schemes used by previous standard [4]. OFDMA is adopted in the downlink direction in LTE and divides the available frequency band into several orthogonal frequency subcarriers. It can be described as a mixture of FDMA and TDMA techniques, making it ideal for cooperating with environments with high RF interference. On OFDMA, one or more 0.5 ms time slots in one or more of the channel frequencies are assigned to each UE as seen in Figure 2.7.

The bit rate increases with the number of resource blocks a user receives and the type of modulation applied to the resource elements. This allows LTE to link adaptation through the assignment to UEs of different size resource blocks, which leads to the change of modulation and channel coding rate in order to maximize resource's utilization. Due to the low symbol rate, applying a guard interval between symbols is feasible, which eliminates inter-symbol interference (ISI). The OFDMA approach results in a high Peak-to-Average Power

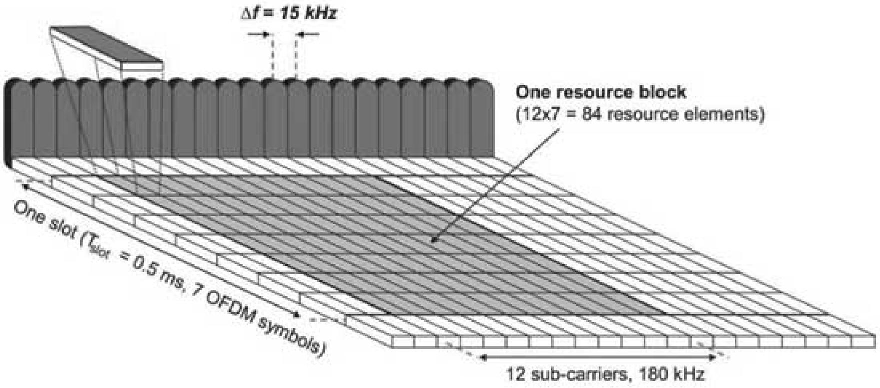


Figure 2.7: OFDM spectrum analysis

Ratio (PAPR), which raises the sender's power consumption by requiring expensive power amplifiers with strict linearity requirements. Due to obvious reasons, the UEs could not be equipped with such good quality amplifiers as required by OFDMA. So, a modulation scheme called SC-FDMA is used in the uplink part of LTE that has lower PAPR levels.

2.4.2 Multiple-Input and Multiple-Output (MIMO)

Another critical technology of the fourth generation of cellular networks is Multiple-Input and Multiple-Output (MIMO) technique [5]. MIMO is the key to improving the LTE's spectral efficiency, energy efficiency, and processing complexity. A MIMO system includes N transmission antennas and M reception antennas, as shown in Figure 2.8, to take advantage of the multi-path effects that are always there and transmit additional data rather than producing interference. Furthermore can achieve much better data rates than conventional single-input, single-output (SISO) channels because they use the spatial dimension of a communications link. Figure 2.9 illustrates the limitations of SISO transmissions concerning the capacity as a function of SNR against MIMO.

2.4.3 4G network architecture

LTE high network architecture adopts and evolves the packet-oriented part of UMTS's architecture as is shown in the Figure 2.10.

An overview of a high level 4G network architecture is split into two key components that are named Evolved UMTS Terrestrial Radio Access Network (E-UTRAN) and Evolved Packet Core Network (EPC).

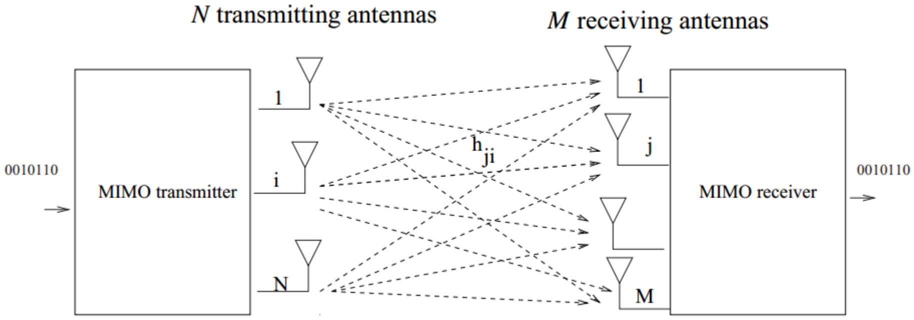


Figure 2.8: Mimo Transmitter and receiver

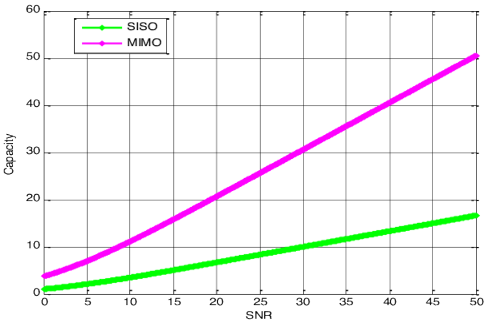


Figure 2.9: Capacity SISO vs MIMO

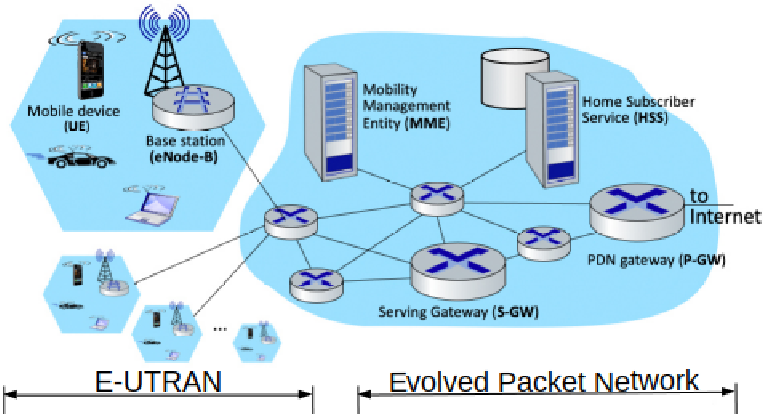


Figure 2.10: LTE network architecture

Evolved UMTS Terrestrial Radio Access Network (E-UTRAN)

As its name describes, the E-UTRAN part of an LTE network is the evolution of UMTS's RAN and is the access part of the EPC. The main parts of E-UTRAN are listed below:

- User Equipment (UE): As in 3G, UE is any mobile or non-mobile device that can connect the user to the 4G network.

- eNode-B: Stands for evolved NodeB (eNB) and is the evolution of 3G's NodeB to 4G.

Each UE is connected with one eNB through Uu interface. E-UTRAN adopts a flat distributed architecture since there isn't any centralized controller, in which all eNBs are interconnected via the X2 interface and connected with the core network via the S1 interface, as shown in the Figure 2.11. The reason behind this decentralized architecture is to minimize the time needed for essential 4G procedures like connection set-up and handovers. The dynamic allocation of resources to UEs in uplink and downlink, the IP header compression and encryption of user data stream, the MME[1] selection, routing of User Plane data towards Serving Gateway, more Radio Resource Management and other operations are hosted by eNBs.

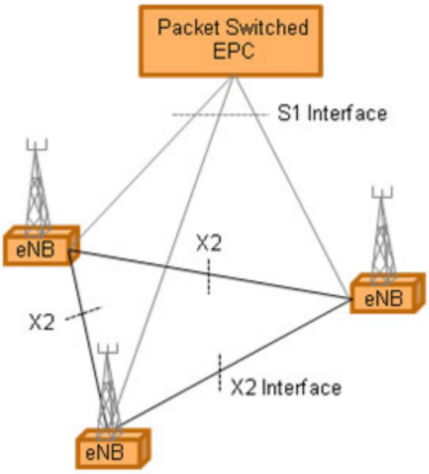


Figure 2.11: LTE-RAN architecture

Evolved Packet Core (EPC)

EPC is also based on flat network architecture. Its responsibility is to handle the data traffic efficiently concerning performance and cost perspectives. One of the key ideas is the separation between the user data and the signaling for scalable reasons (see Figure 2.12). EPC offers various services, such as mobility management, authentication, session management, and the use of different Quality of Service levels. The EPC has four network elements: the Serving Gateway (Serving GW), the PDN Gateway (PDN GW), the MME, and the HSS.

---

[1] The MME stands for Mobile Management Entity is a part of LTE Core Network and will be described in the following section.

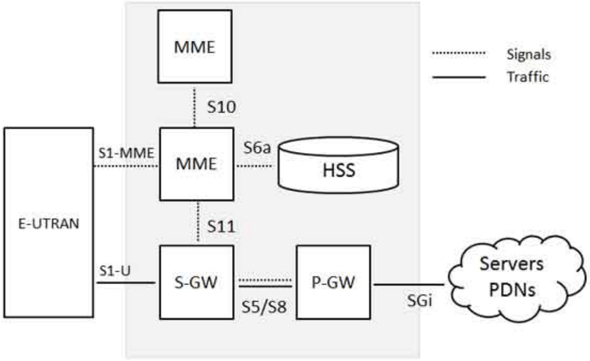


Figure 2.12: LTE-EPC architecture

- Home Subscriber Server (HSS): The HSS belongs to the control plane of the Core Network. It is a database that contains useful information about mobile devices (e.g., IMSI, Subscriber Key (K), Authentication Management Field (AMF), Operator Key (OPC/OP)) and their integration into the network. In combination with the MME, they carry out the authentication of the users.

- Packet Data Network Gateway (PGW): Packets coming from users and leaving the LTE network to encounter the PGW as their final hop before entering the PDN via SGi interface. It belongs to both the Core network's control plane and user data plane. For the remainder of the PDN, it performs the functions of a standard router and also some others functionality like IP address / IP prefix allocation, policy control and charging.

- Serving Gateway (SG-W): SGW also belongs to both the Core network's control plane and user data plane, and routes the data traffic to PGW while concealing the UEs' mobility.

- Mobile Management Entity (MME): MME belongs to the control plane of EPC. It is responsible in conjunction with the HSS for performing the authentication of a newly connected UE in the 4G network. The path of the data from the UE to the P-GW follows the path UE, eNB, S-GW, and P-GW. Between the entities eNB, S-GW, and S-GW, P-GW, the communication takes place with the use of a tunnel. These tunnels are configured by the MME so that in case a UE moves from one cell to another, only the tunnel between the eNB and the S-GW needs to be changed, thus offering flexibility to our network. Cell location via the paging process, which locates idle devices that have switched connection cells, is another crucial MME competency.

2.4.4 LTE protocol stack

As already mentioned, LTE architecture is divided into data and control planes. The same strategy is introduced in the protocol stack. Depending on the plane we're referring to, the protocol stack for each interface changes. There is even an additional separation between the protocols used on air interfaces (Uu) and those used on wired interfaces (S1, S10, S6a, S11, S5, S8, SGi).

The air interface protocol stack for the control and user plane is shown in Figure 2.13.

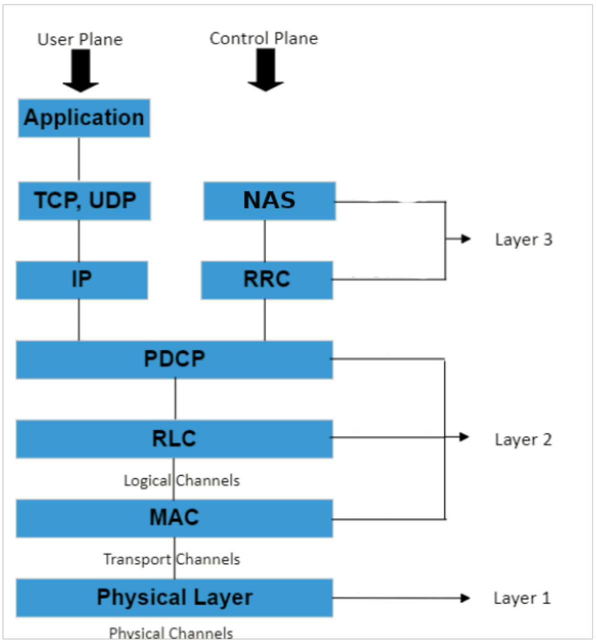


Figure 2.13: LTE User and Control Plane Air interface Protocol Stack

The protocols are illustrated in a top down approach:

- Radio Resource Control (RRC): RRC belongs in the control plane protocol stack of air interface of LTE, its functionality is to to support the flow of information to the upper levels. The RRC layer supplies the UE-UTRAN component of signaling links. Upper layer data is transferred using the signaling connection between user equipment and the core network.

- Packet Data Convergence Protocol (PDCP): The Packet Data Convergence Protocol, as its name implies, performs the header compression of IP packets as its primary function. In the user plane, PDCP is located between IP and RLC layers, when in the control plane, is between the RRC and RLC layers.

- Radio Link Control (RLC): The functionality of the RLC layer is to transfer PDUs in different modes in a way that can control the packet flow, for example, using Acknowledgments. RLC belongs to both user and control plane and in the layer stack is between PDCP and MAC layer.

- Medium Access Control (MAC): Radio resource allocation and data transmission services are essentially provided by the medium access control (MAC) layer to the upper layer and belongs to both user and control plane.

The non-air interface protocol stack utilizes better the OSI layers. The control plane protocol stack across X2, S1-MME and s6a interfaces uses the SCTP protocol that belongs to the transport layer and guarantees message delivery across eNB-eNB, MME-eNB and MME-HSS. There are several other protocols that belongs to the upper layer of SCTP in LTE:

- The X2AP protocol belongs to the control plane and it is used to handle UE's mobility in the RAN. Its met in the protocol stack on the X2 interface between two interconnected eNBs (Figure 2.14).

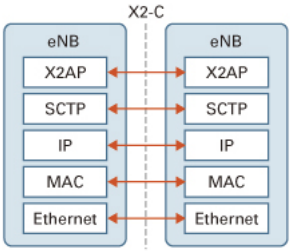


Figure 2.14: eNB-eNB protocol stack

- Between enB (placed at E-UTRAN) and MME (placed at EPC), the S1 Application Protocol (S1AP) provides the control plane signaling via the S1-MME interface (Figure 2.15).

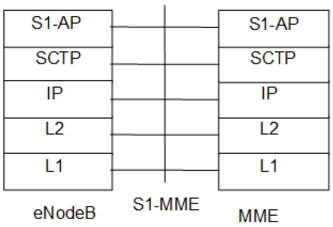


Figure 2.15: MME-enB protocol stack

- For network access and data mobility applications, the Diameter Protocol offers authentication, authorization, and accounting message services and is used above SCTP via s6a interface (Figure 2.16).

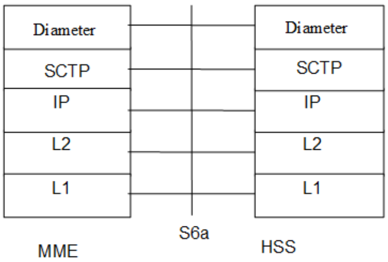


Figure 2.16: MME-HSS protocol stack

Another control plane protocol that is encapsulated in UDP is the GTPC protocol that is used for signalling between the pairs of (MME, SGW) and (SGW,PGW) (Figure 2.17, 2.18).

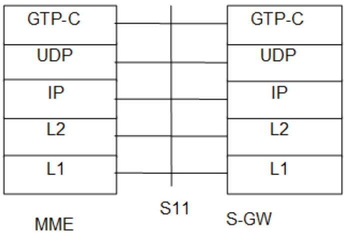


Figure 2.17: MME-S-GW protocol stack

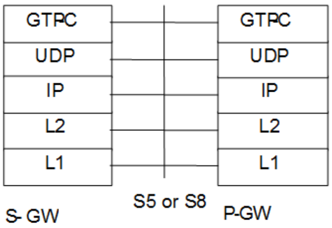


Figure 2.18: S-GW - S-PGW contol plane protocol stack

Last but not least, GTP-U (GPRS Tunnelling Protocol for the user plane) is the non air interface user plane protocol that is used. This protocol tunnels user data between eNodeB and SGW via S1 interface and between SGW and PGW via s8 interface (Figure 2.19).

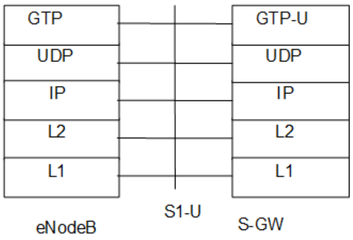


Figure 2.19: MME-S-GW protocol stack

A global overview of the protocol stack used among all the LTE entities are shown in Figure 2.20

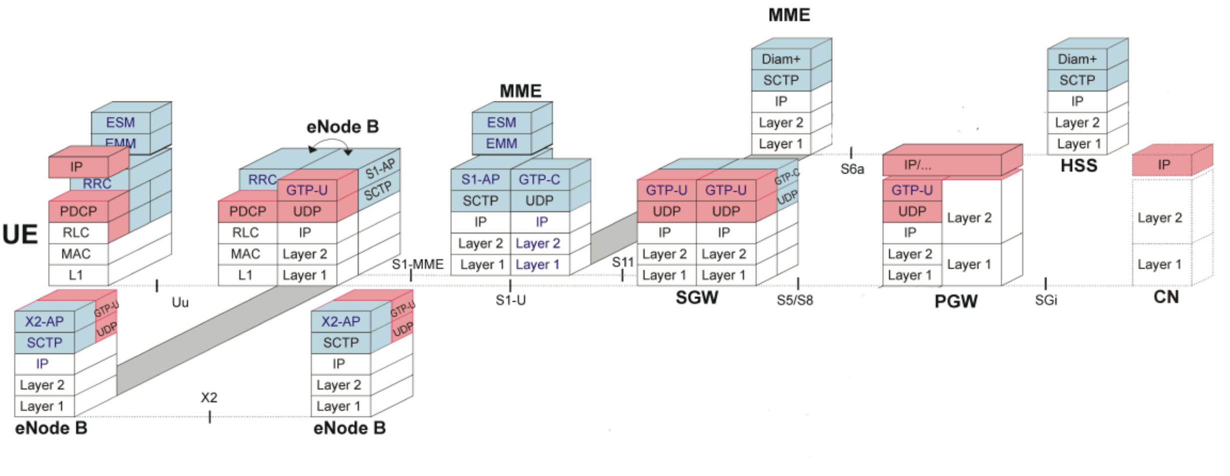


Figure 2.20: Global view of the LTE protocol stack

2.4.5 5G

The cellular network evolution leads to 5G. Based on the previous developments of cellular networks, we expect the 5th generation networks to focus mainly on increasing the data rate. Nevertheless, 5G provides many additional possibilities. 5G is a total re-architecture effort of the access network based on the shift away from a single access service (broadband connection) towards a more robust set of edge services and devices. The basic requirements set for the design of the new generation of cellular networks are pretty demanding and can be outlined in the following:

- Massive Internet-of-Things: Support devices with ultra-low energy, ultra-low complexity, and ultra-high density.

- Mission-Critical Control: Provide ultra-high availability, ultra-low latency, and extreme mobility.

- Enhanced Mobile Broadband: Demands like extreme data rates and capacity.

Since 5G networks are on an evolutionary path and not a point solution, they represent the first attempt to create a fully disaggregated, virtualized, and software-defined cellular network architecture oriented to enable a wide range of future uses beyond those currently fully known.

Non-Orthogonal Multiple Access (NOMA)

One of the most exciting features of 5G over LTE Networks is the usage of the Non-Orthogonal Multiple Access (NOMA) method [6]. Orthogonal multiple access (OMA) methods such as time division multiple access (TDMA), frequency division multiple access (FDMA), or Orthogonal Frequency Division Multiple Access (OFDMA) are commonly used in all pre 5G cellular networks, as seen in the previous sections. The Noma scheme adopts a different technique for allocating available channel resources. Each user in NOMA works in the same band simultaneously, and they may be identified from one another by their power level as is illustrated in Figure 2.21. In the LTE scheduler, the user with a good channel condition is given a higher priority to be served, while the user with a bad channel state must wait for access, so fairness, latency, and connectivity issues are concerned. NOMA can serve numerous users simultaneously with various channel circumstances, improving user fairness, lowering latency, and increasing massive connectivity.

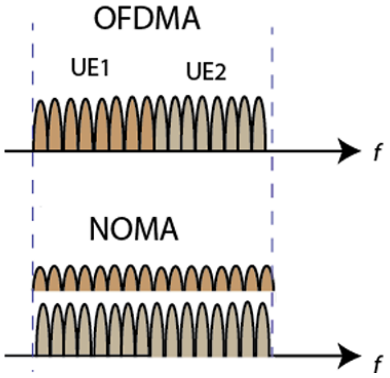


Figure 2.21: Oma vs Noma resource allocation

Family bands of 5G

Due to bandwidth shortage, the 5G new standard makes an effort to expand the available bandwidth since the previous generation of cellular networks has fully occupied bands below 3GHz. 5G networks optimize each new waveform family to a particular radio spectrum band. The first family band, which has its carrier frequencies below 1GHz having a maximum bandwidth of 50 MHz, is dedicated to mobile broadband and IoT services focusing on range performances. The band between 1-6GHZ offers wider bandwidths and is specially created for mobile broadband and mission-critical applications having a bandwidth up to 100 MHz. Last but not least is the third family band, which has carrier frequencies above 24GHz (mmWaves) and provides enormous bandwidth with limited short covering. This increased spectrum bandwidth in combination with massive parallel communications and ultra-dense networks are expected to significantly boost the performance of the 5G cellular networks.

Massive MIMO

Exploiting new spectrum families that involve very high carrier frequencies has led to energy per symbol management problems in traditional MIMO systems, such as those used in LTE. Massive MIMO systems manage the high energy requirements for symbol transmission by increasing the number of antennas used, which brings drastic improvements in throughput and efficiency. Massive MIMO is generally understood as a physical-layer technology that provides each BS with many active antennas that may be used to multiplex numerous UEs on the same time-frequency resource spatially.

Network Function Virtualization (NFV)

NFV is one of the 5G's most critical vital components since it moves functionality traditionally embedded in hardware appliances into VMs and containers that can run on a cluster. Thus, NFV is regarded as a network architectural concept that uses virtualization techniques to create flexible and scalable shared classes of network functions into building blocks that may also be linked together to produce and provide communication services and utilities. Network slicing and end-to-end connectivity are some new entities enhanced into 5G by NFV adaptation. 5G's virtualization can lead to effective and efficient deployment, scaling, and management systems due to its adaptability since it can move functions across distributed hardware resources. Besides, the NFV network architecture gives a full cluster-based character to the 5th generation networks, allowing them to adapt to needs and requirements that we have not yet met.

2.4.6 5G network architecture

Although the 5th generation cellular network's architecture is based on virtualized architecture techniques and function entities, its basic architecture remained the same as that of LTE, dividing the network into two parts: the radio access network (RAN) and the mobile core. There is also the separation of the network into the control and user plane.

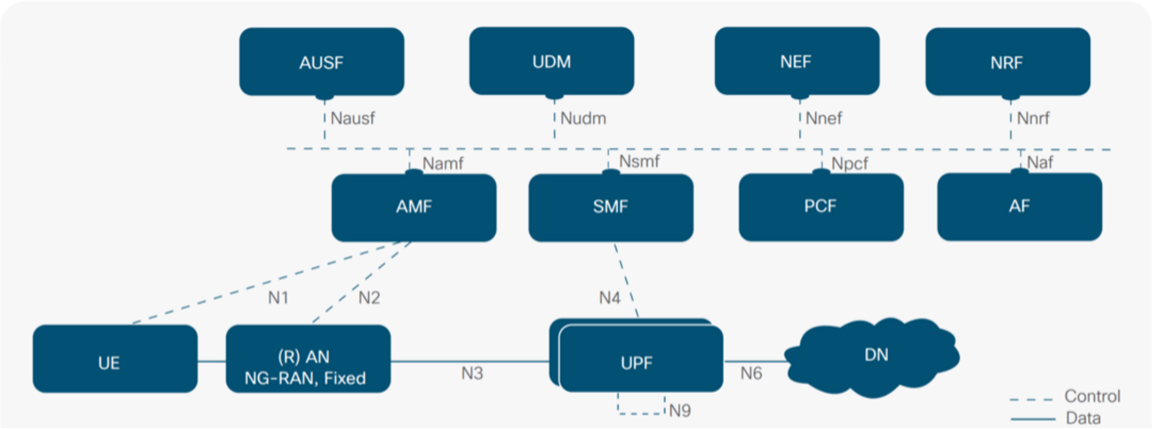


Figure 2.22: 5G system architecture

Radio Access Network in 5G

RAN adopts the same functionality as in LTE, enriched with additional mechanisms such as link aggregation and load balancing. The gNB, a radio node, is the 4G architecture's equivalent of the eNB. The split functionality is a notable difference between 5G and 4G RAN. By maintaining the same protocol stack as 4G and separating it into three different entities, 5G makes a separation between physical and logical elements by separating the protocol stack (Figure 2.22) and further splits RAN functionality into centralized and distributed entities (Figure 2.23).

This results in a RAN cluster-based and distributed version, in which a single Central Unit (CU) serves multiple Distributed Units (DUs), each of which DU serves multiple Radio Units (RUs) as seen in Figure 2.24.

The only constraint in this approach is that since scheduling decisions take place at the

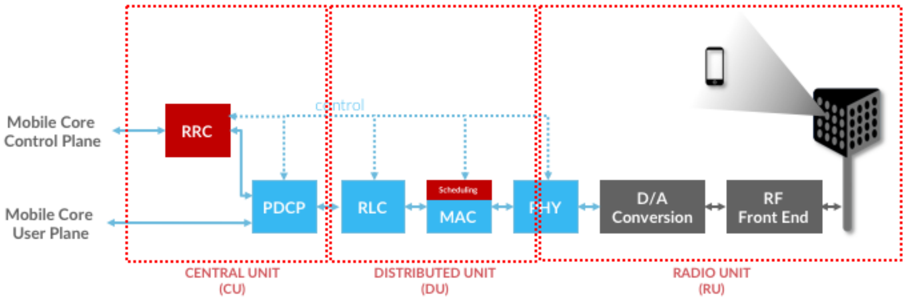


Figure 2.23: RAN-Split into CU, DU, RU

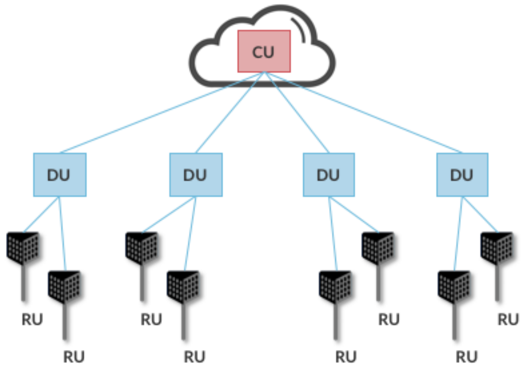


Figure 2.24: RAN-Split hierarchy

MAC layer which is placed in DU, DUs and RUs needs to be near in orderd to have real time communication.

Also, this split-Ran architecture completely alters the nature of the Backhaul Network as described in LTE (eNB - core network connectivity) inasmuch the gNB is now separated into CU, DU, and RU. The corresponding connections are now the following:

- RU-DU: Fronthaul
- DU-CU: Midhaul
- CU-Mobile Core: Backhaul

5G core Network

The new standard decomposed the 5G Core Network's architecture into a set of Network Functions divided into control and user planes. These functions are the following:

- Access and Mobility Management Function (AMF): The AMF function of 5G is the evolution of the 4G MME, without some features such as Sessions Management which is now a separate function of the SMF. It is responsible for RAN termination, access authentication, authorization, and mobility management. The supported interface list between AMF and the other functions of 5G Core is:

  - N1 - Reference point between UE and AMF.
  - N2 - Reference point between R(AN) and AMF.
  - N8 - Reference point between AMF and UDM.
  - N11 (Namf) - Reference point between AMF and SMF.
  - N11 (Nsmf) - Reference point between AMF and SMF.
  - N12 - Reference point between AUSF and AMF.
  - N14 - Reference point between AMF and AMF.
  - N15 - Reference point between AMF and PCF.

- Session Management Function (SMF): The SMF is one of the most critical elements of 5G's service-based architecture. It is responsible for session management, allocating, and managing the IP addressing of UEs and quality of service. The supported interface list between SMF and the other functions of 5G Core is:
  - N11 - Reference point between SMF and AMF.
  - N10 - Reference point between SMF and UDM.
  - N4 - Reference point between SMF and UPF.
  - N7 - Reference point between SMF and PCF.
- Authentication Server Function (AUSF): AUSF provides authentication and authorize functionalities such as authenticate the UE for each network function. The supported interface list between AUSF and the other functions of 5G Core is:
  - N12 - Reference point between AUSF and AMF.
  - N13 - Reference point between AUSF and UDM.
- Network Exposure Function (NEF): NEF offers a possibility to access internal data of the core network to third parties via exposing services. The supported interface list between NEF and the other functions of 5G Core is:
  - N30 - Reference point between NEF and PCF.

- Policy Control Function (PCF): It controls the behavior of the network, providing policy rules for other entities in the control plane to follow. The supported interface list between PCF and the other functions of 5G Core is:
  - N15 - Reference point between PCF and AMF.
  - N30 - Reference point between NEF and PCF.
  - N7 - Reference point between SMF and PCF.
  - N60 - Reference point between PCF and UDM.
  - N5 - Reference point between PCF and AF.
- Unified Data Management (UDM): Network user data is managed via unified data management (UDM) in a single, centralized component. UDM is the evolution to the home subscriber service (HSS) on a 4G network, and it was created expressly for 5G and is cloud-native. The supported interface list between UDM and the other functions of 5G Core is:
  - N60 - Reference point between PCF and UDM.
  - N8 - Reference point between UDM and AMF.
  - N10 - Reference point between SMF and UDM.
  - N13 - Reference point between AUSF and UDM.
- Application Function (AF): Any new control plane function that need to be created since is required by network slices and usually provided by third parties. The supported interface list between AF and the other functions of 5G Core is:
  - N5 - Reference point between PCF and AF.
- User Plane Function (UPF): UPF is responsible for interconnecting the mobile infrastructure with the Data Network, PDU sessions, packet routing and forwarding, QOS handling and traffic reports for billing and more. The supported interface list between UPF and the other functions of 5G Core is:
  - N3 - Reference point between UPF and gnB.
  - N6 - Reference point between UPF and DN.
  - N4 - Reference point between SMF and UPF.

2.4.7 Chapter Conclusion

This chapter presents a brief yet detailed presentation of the evolution of the various generations of cellular networks. Thus we can propose changes and modifications both in the architecture of the 5th generation cellular networks and in the entity functions, which will aim to integrate mechanisms and policies for handover control with the Machine Learning usage. Although the proposed implementation will be integrated into entities and functions of the 5th generation network architecture in this thesis, its experimental implementation and evaluation use the LTE networks.

Chapter 3

Handover

3.1 Background

One of the essential elements of wireless and cellular networks is their customers' mobility. In the 5G era, where the enormous amount of transmitted data and real-time services have a primary role, 5G systems should guarantee different Quality of Service (QoS) and connections with varying characteristics across different scenarios. Customer mobility results in the dynamic change of the topological state of the radio access network and its conditions ass illustrated in the following Figure 3.1.

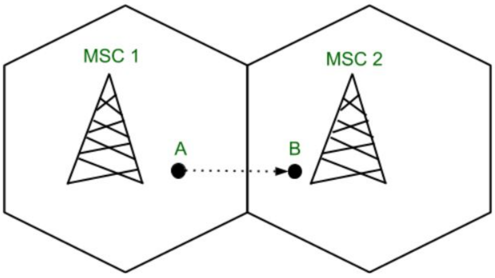


Figure 3.1: Handover Scenario

For example, if a mobile UE moves from inside to outside the cell of the base station it is assigned to, it is common to face different propagation conditions and interference levels. Suppose this situation becomes more acute while at the same time approaching the cell of a neighboring base station from which it has better network conditions. In that case, it is reasonable for the customer to prefer its assignment to the adjacent base station so that its network demands will be better satisfied since the cell serving the UE is not the best.

Handover is named the cellular network procedure in which one UE changes from its source assigned BS to another target BS while keeping its connection active. Generally, a handoff process has three steps. The mobile device, a network agent, or varying network conditions first trigger the handover to start. The second stage is to generate a new connection, during which the network must locate new assets for the handoff connection and carry out any other routing tasks. The data delivery from the old connection path to the new one must be maintained under data-flow control while adhering to the agreed-upon QoS assurances.

There exist two types of handovers, the hard and the soft ones.

Hard handovers

Before the UE connects to the target base station, the UE drops the link between the user and the source base station for a short time. The mobile device can only transmit on one frequency at once. Thus it must switch to a different channel where the connection can be made again (see Figure 3.2). During the break, there is no voice or data transfer. Another name for this style of handover is BREAK-BEFORE-YOU-MAKE. The interruption is so brief that the user rarely really notices it.

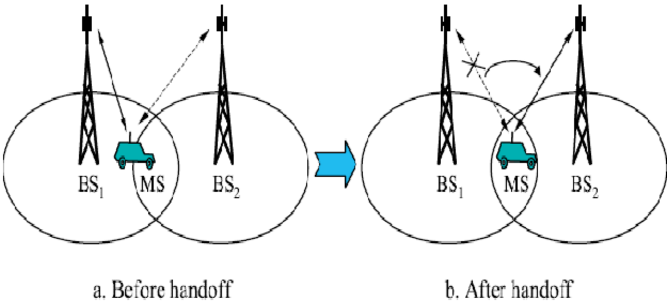


Figure 3.2: Hard Handover

Soft handovers

When a ue is transferred from one cell to another using Soft Handover, the radio link is neither interrupted or broken. The entire frequency spectrum is always available to all users. As a result, it is feasible for adjacent cells to use the same frequency, negating the need to sever the connection during handover. A soft handover is when a new connection is created before the previous one is released (see Figure 3.3). This characteristic has led to its other name, MAKE-BEFORE-YOU-BREAK. This type of handover results in seamless connection.

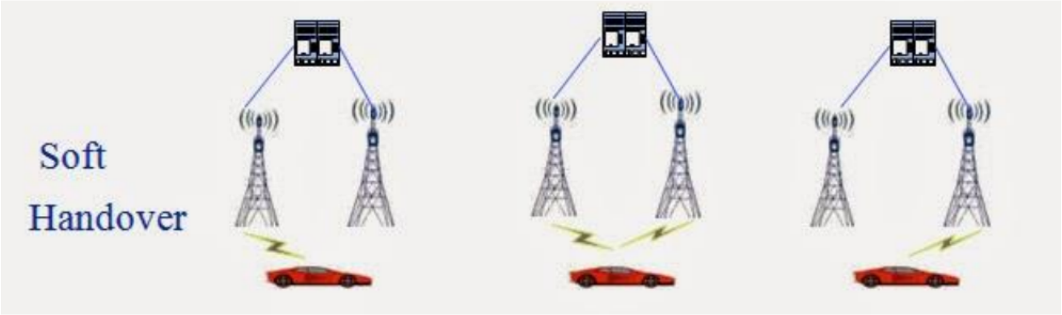


Figure 3.3: Soft Handover

There is also another division of the handover process into Intra and inter. In intra handover, the customer does not change the cell to which he is assigned. A change occurs in the Radio Unit part of the cell, which can be the operating frequency of the antenna or even the radio unit itself. Intra handovers seem to have a lot of room for development in 5G networks with the disaggregated version of split-Ran. Inter handovers are handovers where UE changes the cell to which he is assigned.

3.1.1 Handover in LTE

Handovers in LTE cellular networks can be divided into the following categories:

- Intra-LTE Handovers: Source and target cells in this scenario belongs to the same Core Network.
- Inter-LTE Handovers: With Inter-LTE handovers, we are referring to handovers across different components of the LTE core Network, such as:
  - Inter-MME Handover: Handovers that occur between source MME and target MME can be triggered when UE moves between two different MMEs but is connected to the same SGW.
  - Inter-MME/SGW Handover: UE must switch from one MME/SGW to another. Target eNodeB is a member of one MME/SGW, while source eNodeB is a member of another MME/SGW.

3.1.2 Intra-LTE handovers

Intra-LTE handovers are the most popular and often used handovers in the real world. There are two interfaces that can be used for handover signaling. As already described in section 2.4.3, each eNB in a 4G network architecture is interconnected with another eNB via X2 interface and with EPC via S1 interface.

X2-handover

In LTE, there is no Radio Network Controller, and the network intelligence is stored on the eNB side. If communication link exists between two eNBs, the established X2 link between them is used for signaling and handover and is controlled by the X2 Application Protocol interface (X2-AP).

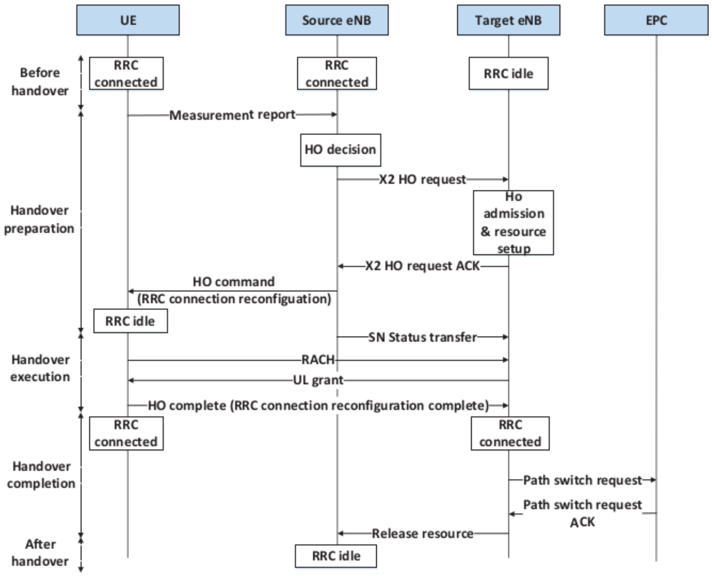


Figure 3.4: X2-Handover

Figure 3.4 illustrates the five steps that make up the X2 procedure.

Before the handover, the UE is connected to the source eNB. UL/DL traffic is sent between eNB and UE thanks to the established radio link between them. While the UE is connected to a corresponding eNB, E-UTRAN and EPC allocate all the resources the UE needs. During the handover preparation step, the UE continuously reports measurements to the source eNB for several metrics, including CQI, RSSI, DL/UL bandwidth, and others. Either the source eNB or the UE can decide whether to initiate a handover and select the target eNB based on the reported measurements.

Following that, the source eNB notifies the destination eNB via an X2 handover request. The necessary information to complete the handover is contained in this message. A handover request acknowledgment (HO-ACK) is sent to the source over the X2 inteface after the target eNB runs a call admission control and determines if it can allocate the requested resources for the new UE. This message, which contains certain setup and synchronization information for the target eNB, is delivered to the source eNB. The source eNB then transmits to the UE the HO command message with the RRC Connection Reconfiguration message inside. The source eNB receives an X2 failure message if the destination eNB is unable to accept the Ho request. The UE states are still unaffected at this phase.

During the execution of the handover, the UE gets the RRC Connection Reconfiguration message, enters the idle state, and then disconnects from the source eNB. Through the X2 interface, the source eNB transmits the Sequence Number (SN) status transfer message to the destination eNB, which contains the PDCP sequence numbers. The first missing data unit for UL is included, and the following sequence number is to be assigned to DL. When the handover to the target eNB has been successfully completed, the UE synchronizes with the target according to the parameters specified and sends the HO Confirm message, which includes the RRC Connection Reconfiguration Complete. As a result, with regard to the target eNB, the UE changes to the connected state.

The target eNB gets the RRC Connection Reconfiguration Complete message during the Handover Completion stage, at which point the path switch operation between the target eNB and the MME/S-GW is started. Prior to any new packets arriving from the Serving Gateway, the target eNB begins to forward all those packets received from the X2 interface to the UE (S-GW). After that, the target eNB's UE release context message is received in order to release the source eNB's UE context. The S1 bearer that was first created between source eNB and UE is also destroyed at this point.

During the After handover stage the UE is attached to the target eNB and follows the same functionality as the before handover stage.

S1-handover

When the X2-based handover cannot be used—for example, due to lack of X2 connectivity to the target eNodeB, an error indication from the Target eNB following an unsuccessful X2-based handover, or due to dynamic information learned by the Source eNB using the STATUS TRANSFER procedure—the S1-based handover procedure is used. By transmitting a Handover required message over the S1-MME reference point, the source eNB starts the handover. The decisions made by the Source eNB are not altered by the EPC. Based on

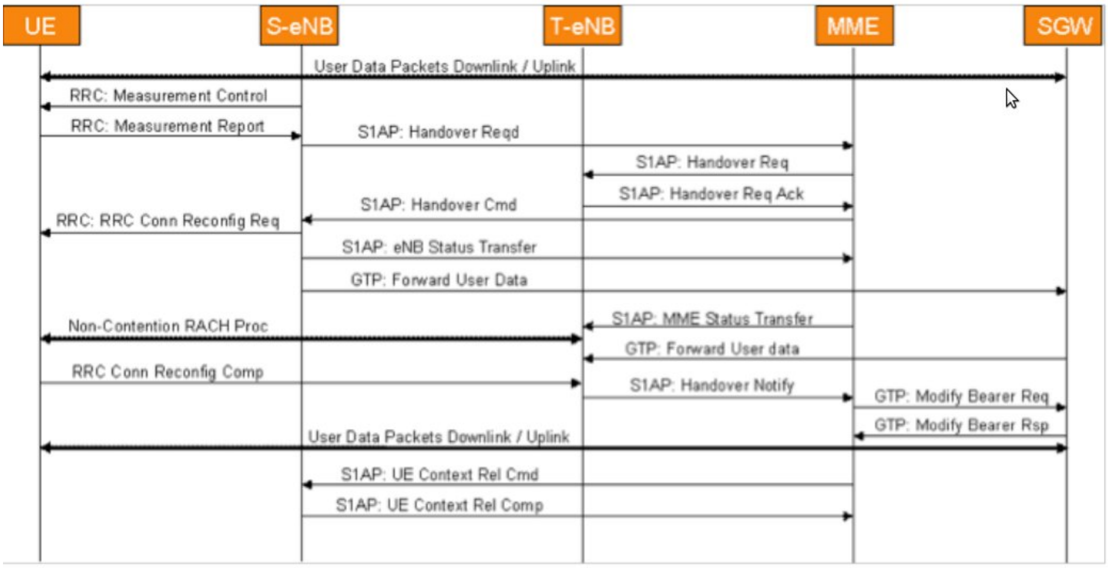


Figure 3.5: S1-Handover

the X2 connectivity with the Target eNB, the Source eNB determines the availability of a direct forwarding path and notifies the source MME of its findings. As show in Figure 3.5, the handover procedure S1 is exactly identical to x2, with the only difference being the mediation of the MME during the communication of the source with the target enB [7].

Intra-LTE Handover decisioning

Multiple reasons could trigger a handover in LTE networks with the main ones being the following [8]:

- Quality-based reasons: According to UE'S measurement report, the currently assigned base station cannot satisfy UE'S desired QoS levels, so the UE needs to switch to another eNB to enhance its QoS metrics.
- Coverage-based reasons: As already said, while a UE moves from inside to outside the cell of the base station it is assigned to, it is common to face different propagation conditions and interference levels. So, since the eNB's coverage is decreasing, the UE has to change the assigned eNB to another with better coverage conditions.
- Load-based reasons: These are optimization reasons related to the load on various eNBs. Can further be categorized to:
  - Load balancing: By taking into account the overall system capacity, this category

manages the load imbalance between two adjacent eNBs.

- Interference coordination: This category lists Radio Resource Management (RRM) activities that can be coordinated to reduce interference. Using this knowledge, the target eNB can decide on its scheduling strategy according to how sensitive it is to interference.

3.1.3 Chapter Conclusion

This chapter presents a brief and detailed presentation of the handover mechanism implemented in LTE networks. After understanding the handover process, it is possible to enforce various load balancing policies for 5G networks to increase the adaptability of the network to user requirements while emphasizing the maintenance of the quality of service.

Chapter 4

Machine Learning

4.1 Introduction to Machine Learning

Machine learning is the process of programming computers to maximize a performance criterion based on previous experience or example data. Learning is the execution of a computer program to optimize the parameters of the model using training data or previous experience. The model might be predictive to make future predictions, descriptive to learn from data, or both.

In simple words, ML is a type of artificial intelligence that extract patterns out of raw data by using an algorithm or method. The key focus of ML is to allow computer systems to learn from experience without being explicitly programmed or human intervention.

The main components of any machine learning algorithm are namely Task(T), Performance(P) and experience (E). In this context, we can simplify this definition as ML is a field of AI consisting of learning algorithms that improve their performance (P), executing some task (T), over time with experience(E).

Based on the above, Figure 4.1 represents a Machine Learning Model

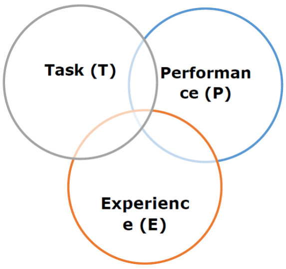


Figure 4.1: Machine Learning Model

- Task (T): From the perspective of problem, we may define the task T as the real-world problem to be solved. The problem can be anything like finding best house price in a specific location or to find best marketing strategy etc. On the other hand, if we talk about machine learning, the definition of task is different because it is difficult to solve ML based tasks by conventional programming approach. A task T is said to be a ML based task when it is based on the process and the system must follow for operating on data points. The examples of ML based tasks are Classification, Regression, Structured annotation, Clustering, Transcription etc.

- Experience (E): As name suggests, it is the knowledge gained from data points provided to the algorithm or model. Once provided with the dataset, the model will run iteratively and will learn some inherent pattern. The learning thus acquired is called experience(E). Making an analogy with human learning, we can think of this situation as in which a human being is learning or gaining some experience from various attributes like situation, relationships etc. Supervised, unsupervised and reinforcement learning are some ways to learn or gain experience. The experience gained by out ML model or algorithm will be used to solve the task T.

- Performance (P): An ML algorithm is supposed to perform task and gain experience with the passage of time. The measure which tells whether ML algorithm is performing as per expectation or not is its performance (P). P is basically a quantitative metric that tells how a model is performing the task, T, using its experience, E. There are many metrics that help to understand the ML performance, such as accuracy score, F1 score, confusion matrix, precision, recall, sensitivity etc

4.1.1 Machine Learing steps

Building an efficient prediction machine learning algorithm involves a number of discrete processes. The entire process entails data collection, data preprocessing, model selection, model training, model evaluation, hyperparameter adjustment, and finally predictions.

- Data collection: The process of gathering and evaluating data on particular variables is known as data collection. These data are currently in a raw format and depend on the corresponding forecasting task.

- Data Preprocessing: Preprocessing data is a crucial step that entails transforming raw data into a clear format before to usage, increasing efficiency. Cleaning up missing or noisy data, normalization, standardization and other processes could be included in this step.
- Model selection: The most pertinent model for the current issue is chosen in this step. Machine learning approaches come in a huge diversity, as was previously stated. However, performance is greatly influenced by the choice of model.
- Training Model: At this step the training model takes place. A particular subset of the available data —referred to as the training set— is used to train the model. The model adjusts to the data by changing its internal state using a variety of methods, such as utilizing the Gradient Descent methodology to minimize a cost function.
- Model Evaluation: Determining whether or not a model performs well requires doing a model evaluation. It makes use of data slicing methods like training/test/validation split. Additionally, evaluating a regression model entails investigating a variety of error metrics, including Mean Squared Error (MSE), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE).
- Hyperparameter adjustment: A set of predetermined arguments known as hyperparameters are accepted by every model. The process of carefully locating and configuring the right model hyperparameters to improve overall performance is known as hyperparameter tuning.
- Predictions: After completing all of the aforementioned processes, the model is finally prepared to be applied and provide predictions.

4.1.2 Chapter Conclusion

This chapter presents a superficial presentation of the Machine Learning theory. The proposed load balance policy will use machine learning algorithms such as Long-Short-Memory (LSTM), Support Vector Regression (SVR) and XGBoost algorithm to predict the uplink throughput of UEs to make more long-term and reliable load balance decisions.

Chapter 5

Handover policy design

5.1 Introduction

As already mentioned, the subject of my diploma thesis is to design mechanisms and policies for handover control in 5G wireless networks. Due to the rapidly rising number of mobile users and nodes, current cellular systems require manual configuration and management of networks, which is currently expensive, time-consuming, and error-prone. As a result, self-organizing capabilities for network management with minimal human participation are introduced. One of the key approach is to reduce the overall congestion across cells, by load balancing UEs among the eNBs.

My work emphasized into load-balancing reasons to trigger handovers. Three different handover policies were introduced:

- Round Robin Policy
- UE's UL Throughput Related Policy
- UE's Predicted UL Throughput Related Policy

5.2 Round Robin

The Round Robin policy is a simple algorithm to distribute UEs across a group of eNBs. This Handover policy is based on the order of arrival of UEs. More specifically, every newly connected UE will be handover from source to target eNB to balance the number of the connected UES across the two cells. This algorithm is illustrated below:

Algorithm 1 Round Robin Load Balancing Handover Policy

```
while true do
    $net\_topo \leftarrow net\_topology(Network)$
    $last\_connected\_ue \leftarrow get\_last\_connected\_ue\_remove(net\_topo)$
    $sum\_clients\_bs1 \leftarrow get\_clients\_bs1(net\_topo)$
    $sum\_clients\_bs2 \leftarrow get\_clients\_bs2(net\_topo)$
    if $abs(sum\_clients\_bs1 - sum\_clients\_bs2) > 1$ then
        $source\_bs \leftarrow attached\_bs(last\_connected\_ue)$
        $target\_bs \leftarrow x2\_connected\_bs(source\_bs)$
        $handover(last\_connected\_ue, source\_bs, target\_bs)$
    end if
end while
```

The Round Robin load balancing policy works best when UE's demands are similar to each other. Figure 5.1 shows the behavior of the RR policy for the following scenario:

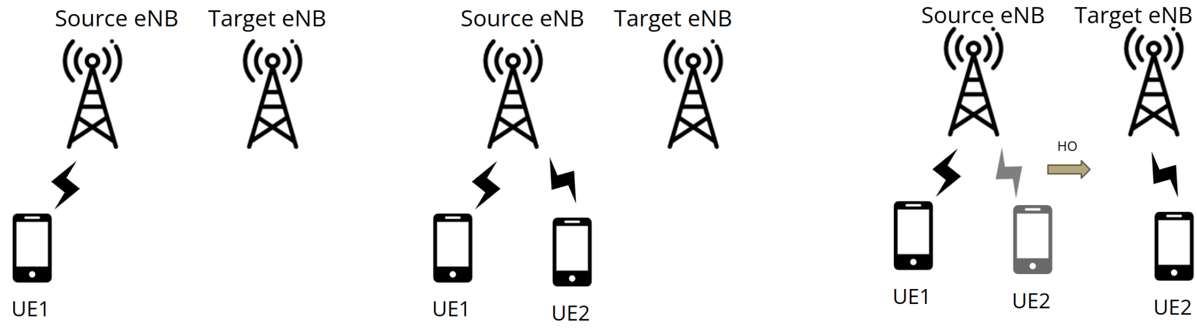


Figure 5.1: Round Robin Run Time Scenario

A UE is initially associated with base station 1, while base station 2 is currently not serving anyone. The equilibrium condition of the customers served by each eNB is breached when the second UE connects to the base station. The handover procedure then starts in order to evenly distribute the allotted UEs among the eNBs. After handover, the number of assigned UEs in both eNBs is equal, bringing our system into balance.

5.3 UE's UL Throughput Related Policy

The UE's UL Throughput Related Policy aims to spread clients over the base stations in a way that considers the sum of uplink demands made by connected UEs per Base station. The key idea is that when a new UE is connected, the UE with the lowest uplink throughput is chosen to be handovered from the source to the target eNB if the sum of the UEs' uplinks assigned to the source eNB is less than the sum of the target. This policy is illustrated as an algorithm below:

Algorithm 2 Throughput Related Load Balancing Handover Policy

```
while true do
    $net\_topo \leftarrow net\_topology(Network)$
    $predict\_thr(net\_topo)$
    $min\_thr\_ue \leftarrow get\_min\_thr\_connected\_ue\_remove(net\_topo)$
    $throughput\_bs\_s \leftarrow get\_thr\_assigned\_bs(net\_topo, min\_thr\_ue)$
    $throughput\_bs\_t, target\_bs \leftarrow min\_bs(net\_topo, source\_bs, 'throughput')$
    if $throughput\_bs\_s > min\_thr\_ue.thr + throughput\_bs\_t$ then
        $handover(min\_thr\_ue, source\_bs, target\_bs)$
    end if
end while
```

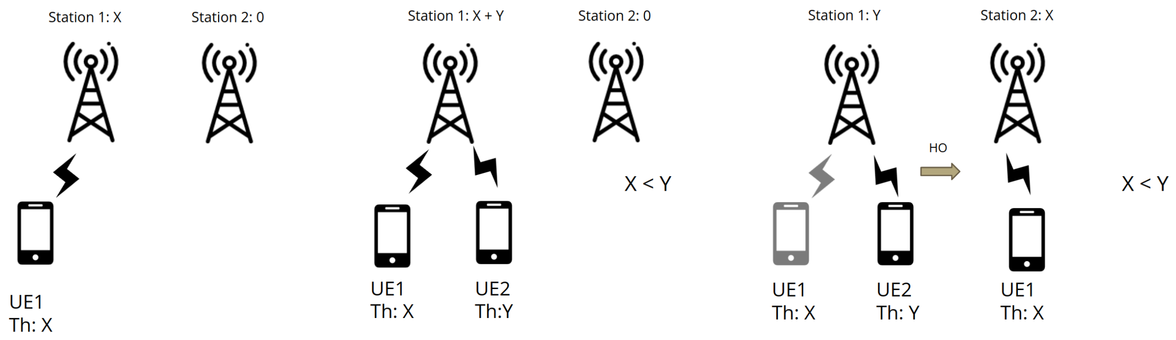


Figure 5.2: UE's UL Throughput policy Scenario

The following example, as shown in Figure 5.2 illustrates how the UE's UL Throughput policy functions. A UE is initially connected to the source eNB and generates to the uplink channel traffic of X size. When the second UE connects to the source eNB, which generates traffic of size Y, the source eNB serves UEs with X+Y traffic demands. The target eNB, in contrast, fulfills customers with no requirements. The UE with the least amount of uplink traffic is picked to be handovered according to quality of service criteria. Finally, when clients are scattered among the eNBs, network uplink traffic is distributed as well.

5.4 UE's Predicted UL Throughput Related Policy

The main distinction between UE's Predicted UL Throughput Related Policy and the UE's UL Throughput Related Policy is that the UE's uplink values are forecasted using machine learning. The pseudocode is as follows:

Algorithm 3 Predicted Throughput Related Load Balancing Handover Policy

```
while true do
    $net\_topo \leftarrow net\_topology(Network)$
    $predict\_thr(net\_topo)$
    $min\_thr\_ue \leftarrow get\_min\_thr\_connected\_ue\_remove(net\_topo)$
    $throughput\_bs\_s \leftarrow get\_thr\_assigned\_bs(net\_topo, min\_thr\_ue)$
    $throughput\_bs\_t, target\_bs \leftarrow min\_bs(net\_topo, source\_bs,' throughput')$
    if $throughput\_bs\_s > min\_thr\_ue.thr + throughput\_bs\_t$ then
        $handover(min\_thr\_ue, source\_bs, target\_bs)$
    end if
end while
```

5.5 Chapter Conclusion

This chapter describes the three policies that were created for my thesis's objectives and tested in a real-world environment in order to conduct experiments, observe their behavior, and evaluate their efficiency.

Chapter 6

Experimental Tools

6.1 Introduction

In this section, we outline the key instruments we employ to assess the experimental implementation. These tools are the following:

1. NITOS, the wireless testbed used to perform experiments in real environments, located in University of Thessaly, Greece.
2. NextEPC, the Open Source implementation of the 4G/5G 3GPP used to deploy 4G core network.
3. the OpenAirInterface platform, the software component used to create the 4G RAN cellular network with the usage of Commercial off-the-shelf hardware.
4. FlexRAN, a Flexible and Programmable Platform for Software-Defined Radio Access Networks, used to collect metrics and perform different network operations.

Each component is described below.

6.2 NITOS testbed

Since the year 2007, the University of Thessaly's NITLAB [9] team has created and maintained the Network Implementation Testbed utilizing Open Source platforms (NITOS) [10]. It has evolved into a feasible solution for evaluating cutting-edge ideas and technologies in networking-related research. The NITOS testbed, one of the most extensive open experimental facilities in Europe, offers highly programmable, remotely accessible technologies to users worldwide via its architecture shown in Figure 6.1 and via the cOntrol and Management Framework (OMF) opensource software. NITOS is a key element of larger resource federations, such as OneLab and Fed4FIRE; the testbed offers the possibility for testing with a wider variety of resources.

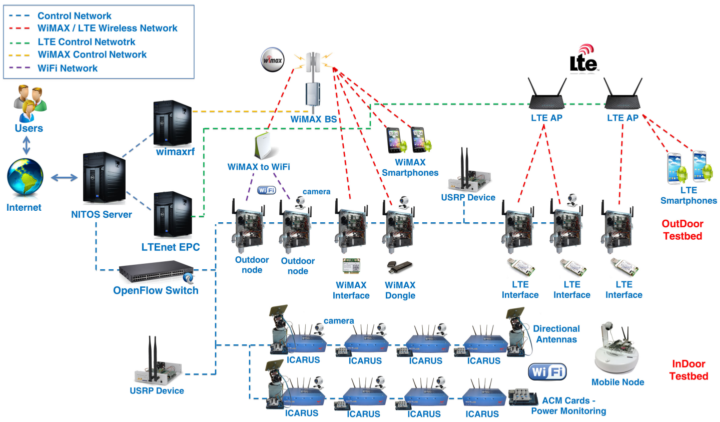


Figure 6.1: Nitos Architecture

The NITOS testbed consists of three individual testbeds of different scales: the Indoor RF Isolated Testbed, the Outdoor Testbed, and the Office Testbed. All the nodes are high-end and powerful machines, with quad-core Intel Core i5 and Core i7 processors, 4 or 8 GBs of RAM, and SSD disks running UNIX-based operating systems. The nodes are connected to Commercial off-the-shelf (COTS) hardware such as multiple User Equipment (UE), including USB dongles and Android smartphones, providing a highly configurable LTE macrocell. Some nodes are also equipped with directional antennas and other cutting-edge prototypes and technologies and connected with 10 USRPs N210, 12 USRPs B210, 4 USRPs X310, and 4 ExMIMO2 FPGA boards making up a Software Defined Radio (SDR) 5G testbed.

6.2.1 Indoor RF Isolated Testbed

The NITOS RF Isolated Indoor Deployment is set up inside one of the University of Thessaly's campus buildings (see Figure 6.2). It consists of 50 Icarus nodes placed symmetrically around the isolated environment forming a grid topology with Wi-Fi, LTE, and 5G wireless interfaces, each of which uses open-source drivers. Experimenters can deploy many scenarios, such as running and assessing power-demanding processing algorithms and protocols.

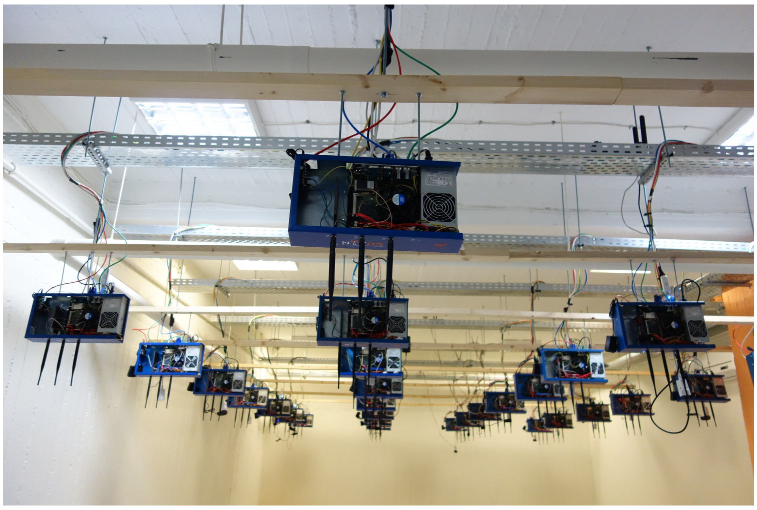


Figure 6.2: Nitos Indoor testbed

6.2.2 Outdoor Testbed

The NITOS Outdoor deployment comprises nodes with multiple wireless interfaces that enable the testing of heterogeneous (Wi-Fi, LTE and 5G) wireless technologies. It has been installed in an outside non-isolated environment at the University of Thessaly (UTH) campus building, as seen in the Figure 6.3, and it has 50 nodes, 25 of them in a grid topology placed all over the roof of the building, and 25 of them spread across the building's floors.

Figure 6.3: Nitos Outdoor testbed

6.2.3 Office Indoor Testbed

The Office Indoor Testbed is made up of ten ICARUS nodes distributed accross a floor (see Figure 6.4). The nodes encompass a variety of heterogeneous technologies, including Wi-Fi, LTE, and 5G, and enable experimenters to create and carry out real-world scenarios in a deterministic office enviroment.

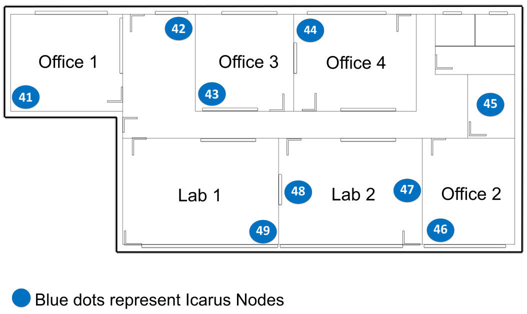


Figure 6.4: Office indoor testbed

6.3 NextEPC 4G Core Network

For the experimental evaluation, the core Network that was deployed was a 3rd party EPC. NextEPC [11] is a C-language Open Source implementation of the 3GPP Evolved Packet Core, more specifically, the core network of a LTE network. It is compatible with various Linux distributions, such as Debian, Ubuntu, and others. Unlike previous EPCs, NextEpc offers voice and text message transmission over the LTE network instead of conventional circuit switching networks and is used with various 4G and 5G software stacks and physical radio equipment. It provides multiple LTE core network entities such as Mobility Management Entity (MME), Serving Gateway (SGW), Packet Data Network Gateway (PGW), Home Subscriber Server (HSS), and Policy and Charging Rules Functions (PCRF). The S1 interfaces to the eNodeBs, the S11 interface to the SGW, and the S6a interface to the HSS are all provided by NextEPC MME. The NextEPC SGW implements the S5 interface to the PGW and the S11 interface to the MME. In IP networks, NextEPC PGW serves as an edge router. It has S5 and SGi interfaces for the Internet and S7 and PCRF interfaces. The user subscription database is called NextEPC HSS. Using the DIAMETER protocol, it implements the S6a interface for MME. NextEPC PCRF controls the norms and guidelines for QoS of LTE users and bearers. It offers PGW's Gx interface. All the entities components alongside the name of the interconnected interfaces are shown in Figure 6.5.

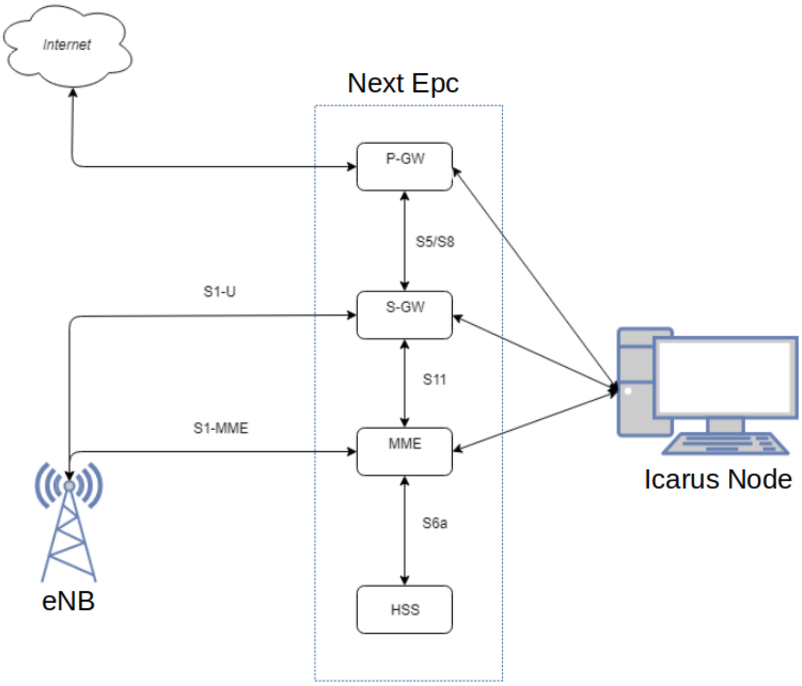


Figure 6.5: Next-Epc Core Network diagram

6.4 The OpenAirInterface Platform

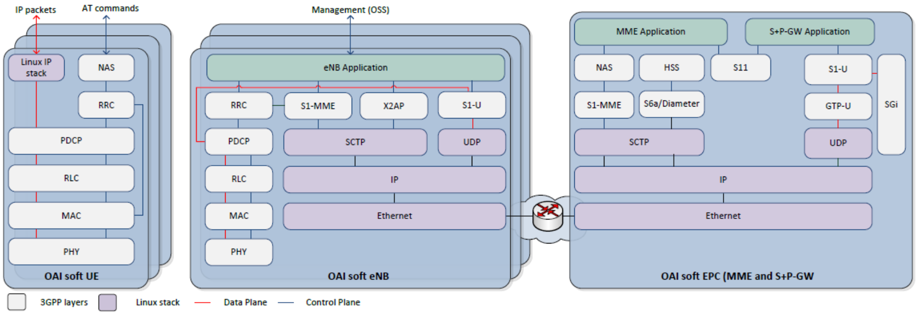


Figure 6.6: OAI - protocol stack

The first open-source software-based implementation of the LTE system covering the entire protocol stack of 3GPP standards is the OpenAirInterface (OAI) [12] wireless technology platform. It is also the first fully open-source SDR solution x86-based and supports UE, eNB, and core-network functions as shown in Figure 6.6. On a host machine, it can be used to design and modify an LTE base station and core network, connect commercial user equipment (UEs), test various network configurations, and keep track of the network and mobile devices in real-time. The foundation of OAI is a radio frontend software architecture hosted on a PC. OAI uses a software radio front end connected to a host computer for processing to implement the transceiver capability. OAI offers a robust development environment with several built-in tools, including highly realistic emulation modes, soft monitoring and debugging tools, a protocol analyzer, a performance profiler, and a configurable logging system for all layers and channels.

6.5 FlexRAN controller

Another software tool used in this thesis is the FlexRAN controller [13]. FlexRAN is the First Open-source Implementation of a Flexible and Programmable Platform for Software-Defined Radio Access Networks. The FlexRAN Service, the Control Plane, and the FlexRAN Application plane synthesize the FlexRAN architecture. The Real-time Controller (RTC), connected to various underlying RAN runtimes, one for each RAN module, makes up the FlexRAN service and control plane, which has a hierarchical design (see Figure 6.7). A monolithic eNB, for instance, has one RTC assigned to it, but a disaggregated eNB or gNB might have several RTCs. The RAN runtime environment, which serves as an abstraction layer with the RAN module on one side and the RTC and control apps on the other, separates the control and data planes. The real-time controller and RAN agent installed in the runtime environment can communicate, thanks to the FlexRAN protocol. Applications to control the RAN infrastructure can be created using both the RAN runtime and the RTC SDK. These applications enable coordination, control, and monitoring of RAN infrastructure. All RAN data and APIs created are available for use by third parties.

The OAI architecture incorporates the FlexRAN software. This integration breaks down the fundamental FlexRAN architecture into FlexRAN agent and FlexRAN controller. According to FlexRAN's architecture, the OAI eNB source code was refactored to separate the eNodeB data plane from the control logic, and the necessary function calls were introduced to the FlexRAN Agent API to facilitate this separation. For instance, function calls to access data plane MAC/RLC information, such as the sizes of UEs' transmission queues, as well as calls to implement scheduling decisions, were introduced to the API. Another important functionality that the integration of FlexRAN in OAI offers is the handover procedure functionality.

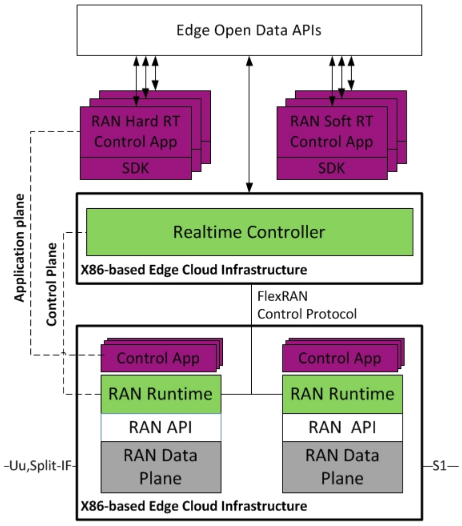


Figure 6.7: FlexRAN architecture

6.6 Chapter Conclusion

This chapter presents the infrastructure and all the software tools used to deploy our experimental scenarios. The LTE system was deployed using different Icarus nodes of the Nitos Indoor RF Isolated Testbed to install the LTE entities. NextEPC was used to install the LTE Core Network, OAI and USRP b210 devices were used to deploy the RAN, and the Flexran Controller was used to track handovers and RAN metrics. UE devices were utilized with the Huawei E3272 LTE dongles.

Chapter 7

Experimental System Architecture and Handover Analysis

7.1 Experimental System Architecture

To conduct the experiments, we built an LTE system suited for performing X2 handovers, using the Nitos Indoor RF Isolated Testbed's nodes. This topology includes the entities listed below and is seen in Figure 7.1:

- EPC node
- Interconnnected Source and Target eNB
- UE
- Flexran Controller
- Flexran Scrapper

As mentioned above, the machines used to create the LTE network were the Nitos Indoor RF Isolated Testbed nodes. Each node has two ethernet interfaces, where one of them is used for remote control of the machine, while the other is for research purposes. For the communication of the various LTE entities, the second interface was used, with which we managed to create the corresponding S1 and X2 links between eNB-MME and SeNB(Source eNB)-TeNB(Target-enB) respectively. All the nodes have the same image using as operating system the Ubuntu 16.04 LTS (Xenial Xerus).

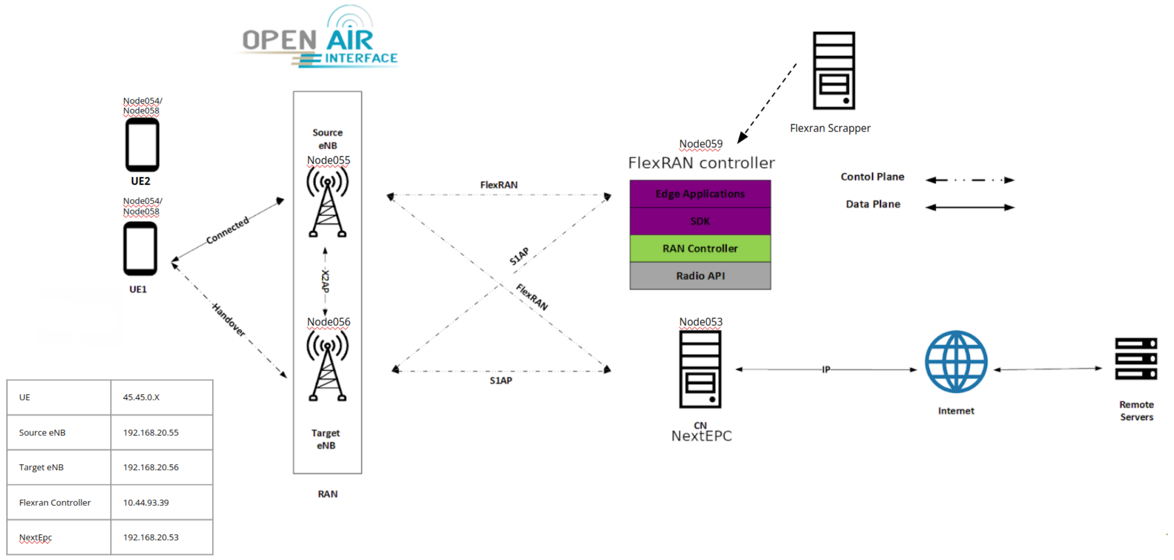


Figure 7.1: X2 handover topology

EPC node

On the node used for EPC, the nextEPC software was installed and configured so that core network entities MME, SGW, PGW, HSS, and PCRF will be hosted on the same machine while exploiting MME's S1 interface via the host's second interface. Using NextEPC's Web UI, we added a subscriber updating HSS's database with the IMSI, security context(K, OPc, AMF), and APN of the subscriber.

eNB nodes

The nodes that will be used to implement two monolithic LTE eNBs should be connected to some USRP devices in order to deploy the radio part of a 4G network using SDR. The USRP B210 was selected as the USRP device for this purpose because it offers a fully integrated single board general software radio peripheral platform with continuous frequency coverage from 70 MHz to 6 GHz. It is made for low-cost experiments and works with the AD9361 RFIC Direct Frequency Conversion Transceiver to provide up to 56MHz Real-Time Bandwidth and a handy bus power supply with fast USB 3.0 data connectivity. Additionally, we configured both eNB's USRPs to have a bandwidth of 5 MHz while operating on different center frequencies, one at 2685000000 Hz and the other at 2665000000L Hz. We also used the second interface to interconnect the node with the neighboring eNB (X2 interface), the MME (S1 interface) and the FlexRAN controller (FlexRAN interface).

UE node

The minicom application was installed on the UE node to connect the LTE dongle with the eNB and create the RAN (Uu) radio link. After the connection is established, UE's lte dongle will be assigned with an IP addressed by the Core Network.

FlexRAN-Controller node

The FlexRAN node, on which the FlexRAN controller was installed, comes in last but certainly not least. The FlexRAN controller communicates with the FlexRAN agent (placed on the eNB) using host's interface designed for research purposes. It is worth mentioning that since the FlexRAN controller offers access to all RAN data and APIs to use by third parties, the IP assigned to this interface must be global instead of private, as used in all the intra-LTE interfaces.

FlexRAN-Scrapper

Alongside the Flex-RAN-Controller node, a custom software developed for my thesis's purposes operates as a third-party software to EPC. During the experiments, this software was installed on my personal computer, and I had remote access to the infrastructure, specifically the FlexRAN controller. Its functionality is separated into three parts. The first part is to get information about the topology and network conditions of the RAN. Then FlexRAN-scrapper processed the collected data to make the handover decision and sent the handover command to the Flex-RAN controller to inform the RAN to begin the handover procedure.

More specifically, every second the FlexRAN-Scrapper retrieves statistical information from the FlexRAN-Controller via a http request. The FlexRAN-Controller's http response was a jason file that contained information for the whole RAN network such as:

- eNB Configuration:
  - Cell Id
  - UL and DL bandwidth
  - UL and DL frequency
- UE Configuration:
  - RNTI

  - IMSI
- UE stats:
  - RLC Report
  - UL and DL Cqi Report
  - RRC Measurements
  - pdcpStats

Parsing the response Jason file, this software creates the topology of the RAN network by attaching active UEs to the corresponding eNBs. Due to radio issues, FlexRAN could not decode UE's IMSI value. This malfunction of Flexran led us to use the RNTI as a unique identifier for each user. RNTI stands for Radio Network Temporary Identifier, and as the name implies, it is a kind of Identification number that can change during the connection. So one UE could be assigned with multiple RNTIs by the time it is attached to the eNB. Therefore we had to find a way to distinguish active UEs from non-active ones. FlexRAN-Scrapper identifies active UEs using the metric pdcp_metric.pktRxBytes, which increases every time a new packet is transmitted from the UE to the eNB. This metric also allows us to calculate each UE's uplink throughput, so we can store its value in a database to perform better inference results in the prediction-based load balance policy. Each newly connected UE is tracked. Last but not least, FlexRAN-Scrapper takes the handover decision according to different approaches:

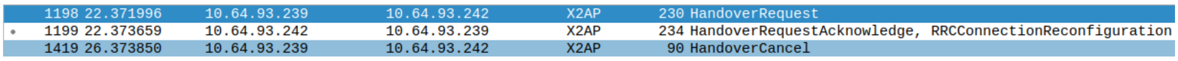
- Round Robin Policy
- UL Throughput Related Policy
- Predicted UL Throughput Related Policy

and sends the handover command to the FlexRAN controller.

7.2 Experimental Handover Analysis

Using the network, the topology of which is shown in Figure 7.1, a UE was handed off from the source eNB to the target eNB, as was previously mentioned. Our first approach was to perform intra-frequency handovers, where the source and target eNB operate at the same center frequency, which is already integrated into OAI [14]. Intra-frequency handovers in OAI are happening according to UE's RSRP (Reference Signal Received Power) reported metric. The change of the UE's RSRP depends on the mobility of the UE. Since we use static nodes to attach the LTE dongle and do not create any motion with an attenuator, UE's RSRP value tends to be the same through the experiment execution. Intra-frequency handover has malicious behavior in a static UE environment. When a UE wants to connect to the cellular network, one metric to choose the best base station is RSRP. Originally, the UE is assigned to the eNB with the best RSRP value when connecting to the network; when we trigger the handover, the UE will be hand-offed to the target eNB with a lower RSRP value. Then the OAI will handover the UE again to the source eNB. So triggering an intra-frequency LTE handover without changing the position of the UE has a ping pong effect that overloads the network.

A solution to this problem was to trigger inter-frequency handovers where the source and target eNBs operate at different center frequencies. As mentioned in the previous section, we trigger handovers with RNTI usage. Because during an inter-frequency handover, the UE must first disconnect from the source eNB and then tune to the target eNB, which operates in different center frequencies, the UE's RNTI value changes during this procedure. This leads to an X2 protocol cancellation since the target eNB waits for a UE to connect with the original RNTI value as shown in Figure 7.2.



```
   1198 22.371996    10.64.93.239     10.64.93.242     X2AP      230 HandoverRequest
•  1199 22.373659    10.64.93.242     10.64.93.239     X2AP      234 HandoverRequestAcknowledge, RRCConnectionReconfiguration
   1419 26.373850    10.64.93.239     10.64.93.242     X2AP       90 HandoverCancel
```

Figure 7.2: X2 handover failure

Although the failure in the protocol level, the UE manages to connect to the target eNB as a new UE at the MAC level. Something worth mentioning is that UE does not lose connection with the Core Network during the X2 handover protocol level failure. but the total time to perform handover increases during an inter-frequency handover over an intra-frequency one by approximately to 10 seconds from 7 seconds, as shown in Figure 7.3.

During the handover procedure, as it is reasonable the round trip time of some packets increases since the UE disconnects from the source eNB and then connects to the target eNB. As shown in Figure 7.4, the round trip time has a peak with the value of 10 seconds which is approximately the same value that the handover procedure takes.

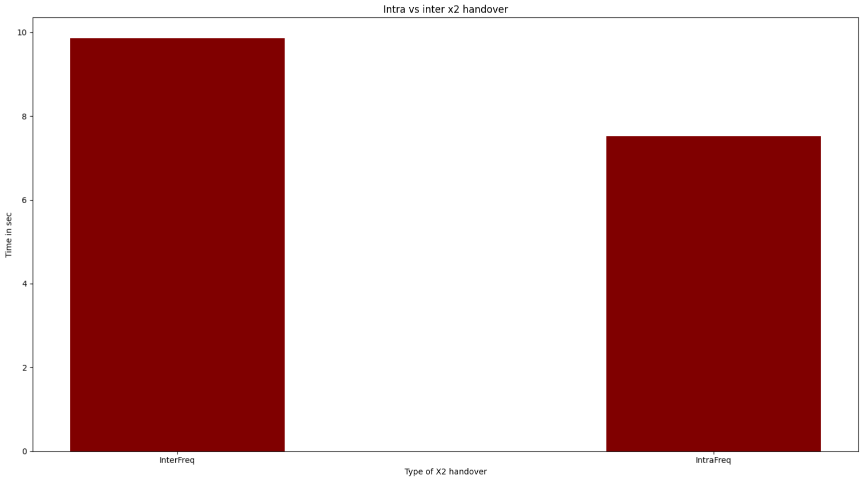


Figure 7.3: X2-type handover time duration

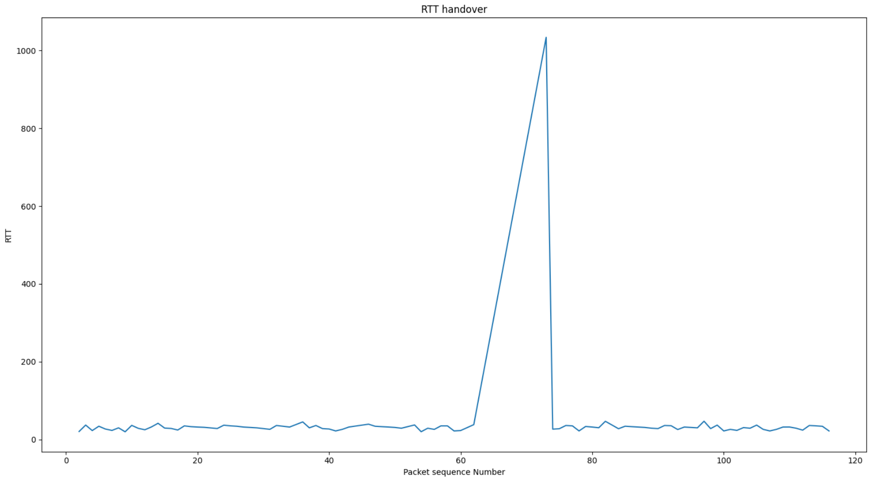


Figure 7.4: Inter-frequency handover Round Trip Time

After the experimental study of x2-handover, the next step was to implement and evaluate the load balancing policies designed in chapter 5.

7.2.1 Experimental results - Load balance policy Round Robin

The first algorithm we implemented was the round-robin one. In this scenario, the UE1 is already connected to the source eNB and begins to ping the Core Network. After 12 seconds, a new UE, the UE2, links to the same source, eNB, and starts to ping the core network. At this point, let's remind that the round-robin tries to distribute UEs across a group of eNBs. Because the target eNB is not associated with any client, Flexran-Scrapper sends a handover command to hand off the last connected UE, UE2, to the target eNB to decongest the source eNB. The UE2 finally connected to the target eNB 10 seconds after the handover was triggered due to the inter-frequency handover delay described in the previous section.

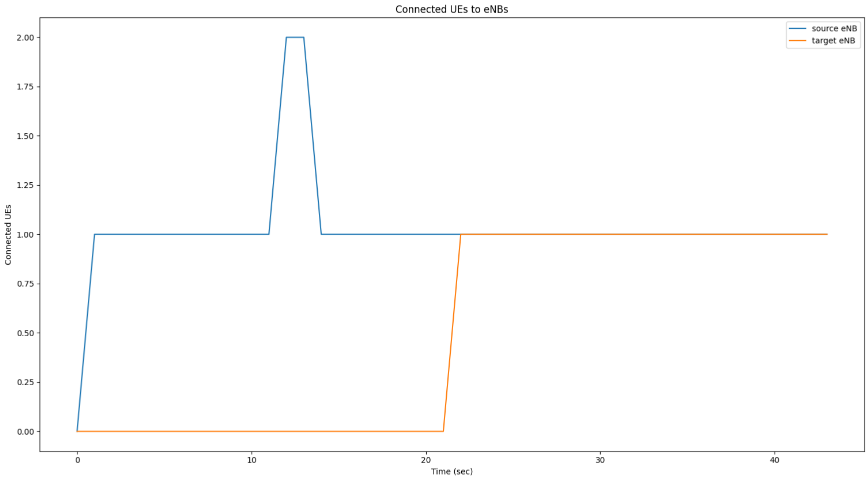


Figure 7.5: Round Robin Policy - Connected Ues to eNBS

As seen in Figure 7.5, the number of associated UEs to eNBs becomes equal to two when the second UE is connected to the source eNB, and immediately (after 1 second) when the handover is triggered, the source eNB is decongested. Finally, when the second UE manages to connect to the target eNB, the number of assigned UEs in both eNBs are equal, bringing our system into balance.

7.2.2 Experimental results - Load balance policy Throughput related

The second policy has chosen to be evaluated the Throughput related. In this approach, the FlexRAN Scrapper aims to spread connected UEs over the eNBs in a way that considers the sum of uplink demands made by connected UEs per eNB. In this scenario, we used the i-perf command to generate UDP traffic in the LTE network. In the beginning, UE1 is already connected to source eNB and generates UDP traffic of 3 Mbps to the Core Network, loading its uplink channel. After 12 seconds, the second UE also connects to the source eNB and generates UDP traffic of 1 Mbps to the Core Network, loading its uplink channel. The FlexRAN Scrapper realizes the congestion in the source eNB since a total of 4 Mbps are transmitted through the source eNB while the target eNB does not serve any UE. FlexRAN

Scrapper selects the UE with the minimum uplink demands to handover it over to the target eNB. In this case, the UE2 is the one that will be handover.

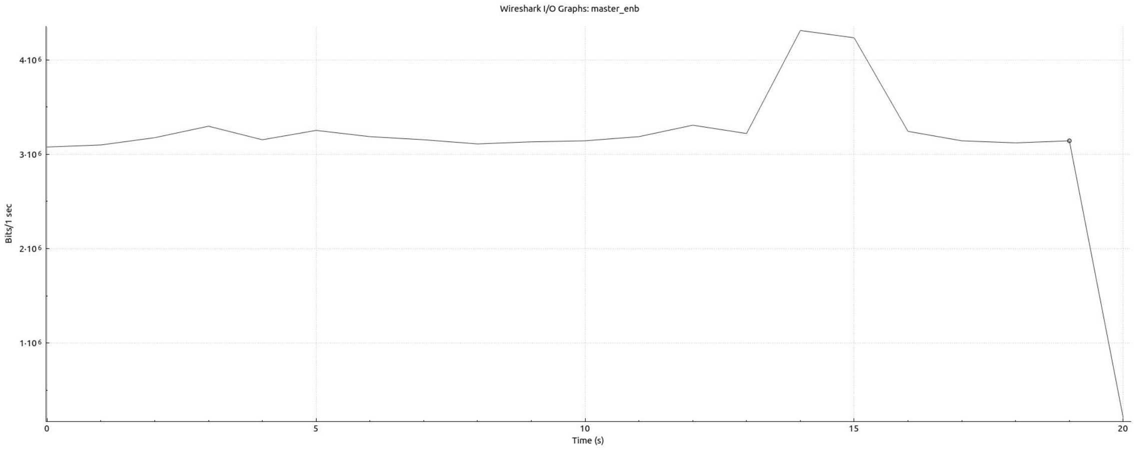


Figure 7.6: Master eNB SPGW-U interface

In order to evaluate and compute this load balance policy, we sniffed the packets at the SPGW-U interfaces in both the source and target eNBs. As seen in Figure 7.6, during the traffic co-existence of the two UEs in the source eNB, the source eNB serves more traffic than previously, which can quickly escalate into a congestion problem. After the handover procedure begins, we quickly see that the source eNB becomes decongested; since UEs are distributed among the eNBs, the network uplink traffic is also distributed. Figure 7.7 shows the target eNB's SPGW-U interface, which serves 1 Mbps, the uplink traffic of the second UE, after the handover procedure.

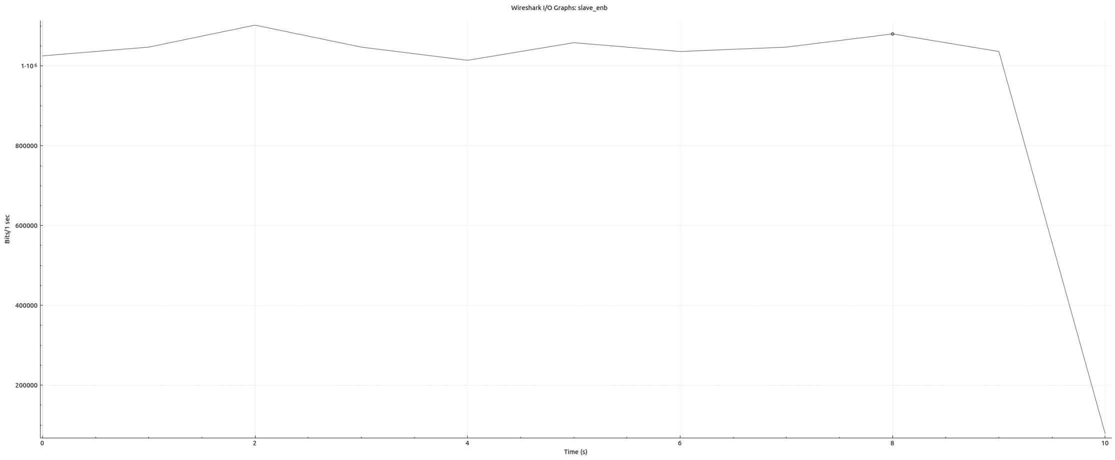


Figure 7.7: Target eNB SPGW-U interface

7.2.3 Experimental results - Load balance policy Predicted Throughput Related

The second and the third policies are almost identical, having the same functionality, with the noticeable difference between them being that the last one is prediction based.

More specifically, we tried to predict the uplink throughput of each UE having as a reference real-environment behaviors and situations of mobile subscribers. The dataset used as a reference contains client-side cellular key performance indicators (KPIs) gathered from two major Irish mobile operators for the 4G trace dataset across various mobility patterns (static, pedestrian, car, bus, and train). With a viewable throughput range of 0 to 173 Mbit/s at a granularity of one sample per second, the 4G trace collection consists of 135 traces with an average duration of fifteen minutes per trace. In our scenario, we made the assumption that UEs are static, so signal strength and motion-related values tend to be the same through the experiment execution, and as a result of that, columns such as RSRQ, RSRP, RSSI, SNR, CQI, Velocity, and Longitude and Latitude were not used. This thesis mainly focused on Uplink throughput prediction degenerating the problem to a time-series forecasting one.

Pre-processing steps

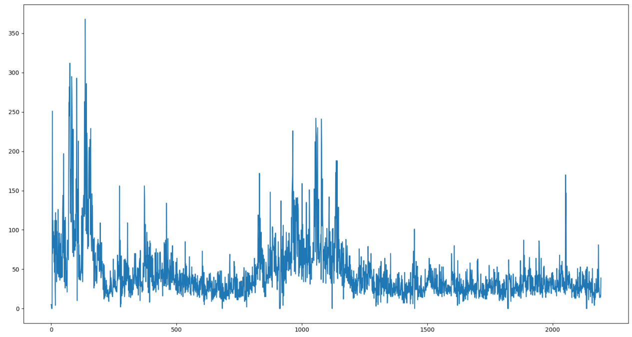


Figure 7.8: Uplink Throughput

As seen in Figure 7.8, the dataset shows sharp fluctuations in UL throughput values, a phenomenon that makes it difficult to train machine learning models, since it causes overfitting or large prediction deviation. A technique named Sequence Flattening was used to reduce sharp fluctuations, in which four consecutive values of the dataset correspond to one, normalizing the abrupt changes as shown in Figure 7.9. Also the sliding window technique was used, having window size of 4.

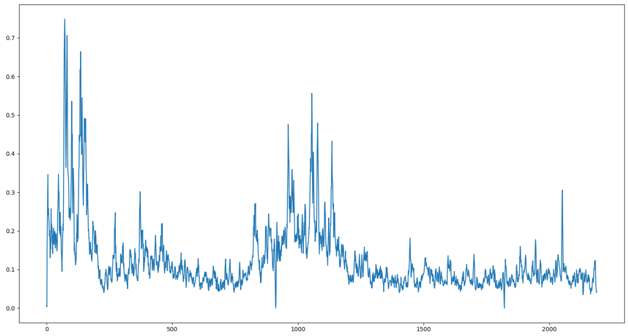


Figure 7.9: Flattened Uplink Throughput

Machine Learning Models evaluation

Three machine learning models were tested and evaluated using the tensorflow [15] python library.

- Support Vector Regression: An approach for supervised learning called support vector regression is used to forecast discrete values. The SVMs and Support Vector Regression both operate on the same theory. Finding the optimum fit line is the fundamental tenet of SVR. The hyperplane with the most points is the best-fitting line in SVR. The SVR seeks to match the best line within a threshold value, in contrast to other Regression models that aim to reduce the error between the real and projected value. The distance between the boundary line and the hyperplane is the threshold value. The MAE that our SVR model produced is 19.95 when the dataset is not flattened and 8.99 when it is flattened.

- Long Short Term Memory: Long Short Term Memory Network is an advanced RNN (Recurrent Neural Network), a sequential network, that allows information to persist (Memory). It is capable of handling the vanishing gradient problem faced by RNN. In our case we used 3 layers of neural networks having 128, 64 and 32 neurons in this order. The MAE that our LSTM model produced is 13.89 when the dataset is not flattened and 4.625 when it is flattened.

- XGBoost: In XGBoost the trees can have a varying number of terminal nodes and left weights of the trees that are calculated with less evidence are shrunk more heavily.

Newton Boosting uses Newton-Raphson method of approximations which provides a direct route to the minima than gradient descent. The extra randomisation parameter can be used to reduce the correlation between the trees, the lesser the correlation among classifiers, the better our ensemble of classifiers will turn out. The MAE that our XGBoost model produced is 16.88 when the dataset is not flattened and 5.002 when it is flattened.

Figures 7.10 and 7.11 provide a comparison graph with the MAE values for the relevant models for the unflattened and flattened datasets, respectively.

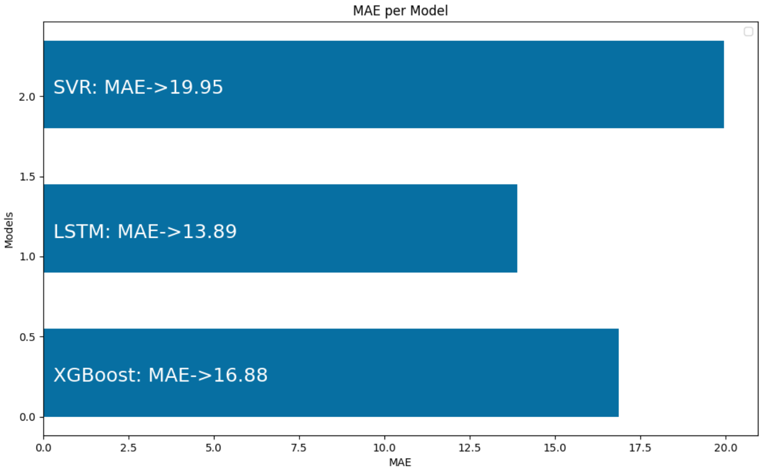


Figure 7.10: Models’ MAE using unflattened dataset

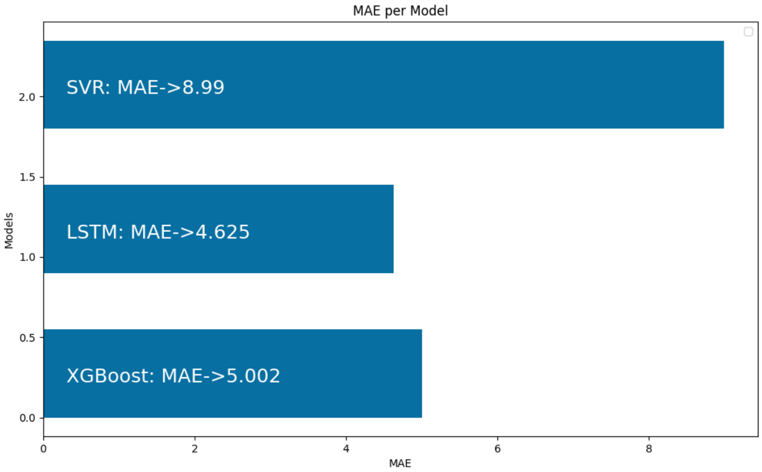


Figure 7.11: Models’ MAE using flattened dataset

The LSTM was chosen to be pre-trained and integrated into the Flexran Scrapper to perform the UL bandwidth prediction of the UEs based on the outcomes of analyzing the predictions of the machine learning models.

In this case, we created UDP traffic in the LTE network using the i-perf command. When UE1 first connects to the source eNB, it generates 7 Mbps of UDP traffic to the Core Network, filling up its uplink channel. The second UE establishes a connection to the source eNB after 16 seconds and generates 1 Mbps of UDP traffic to the Core Network, also utilizing its uplink channel. Since a total of 8 Mbps are transmitted through the source eNB while the target eNB does not service any UEs, the Flexran Scrapper recognizes the congestion in the source eNB. The Flexran Scrapper uses the pre-trained model and predicts the uplink throughput demands for each connected UE. It is important to note that the window size was set to 4, as the Flexran Scrapper needs to wait at least 4 seconds to gather data on a newly connected UE to make assured predictions. After the prediction step, the next step is to select the UE with the minimum predicted throughput to be handover. This load balance policy was evaluated using the same method as previously, which involves sniffing packets at SPGW-U interfaces on both source and target eNBs.

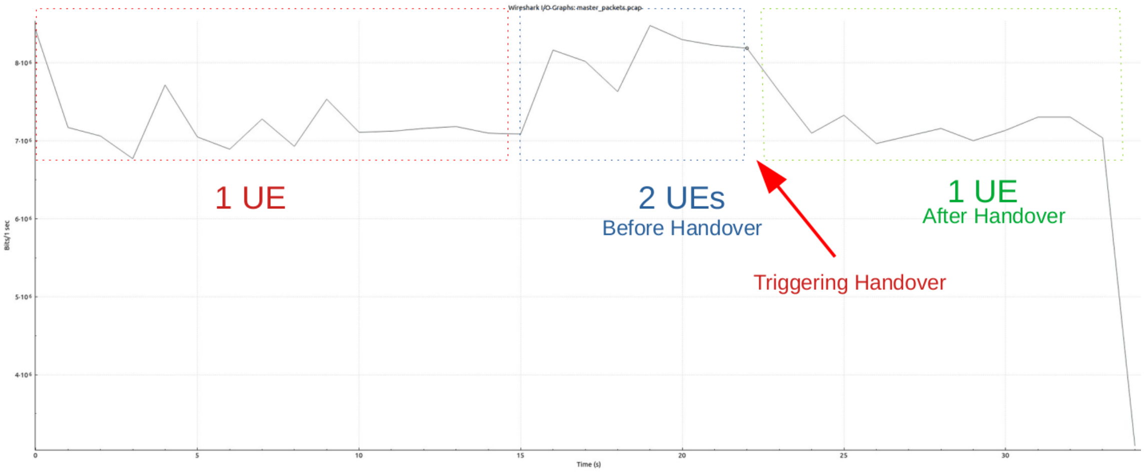


Figure 7.12: Flattened Uplink Throughput

As seen in Figure 7.12, the source eNB serves more traffic than it did prior to the coexistence of the two UEs assigned with, which might soon turn into a congestion issue. The prediction stage causes a brief delay in the handover decision while the data required to make the forecast is gathered. The predicted values for the UE1 were 8.3333 Mbps and the UE2 were 1.23333 Mbps, demonstrating the model's predicted trend of increasing predicted value. The source eNB immediately becomes load balanced after the handover procedure starts; since

UEs are scattered throughout the eNBs, the network uplink traffic is similarly distributed.

Chapter 8

Conclusions

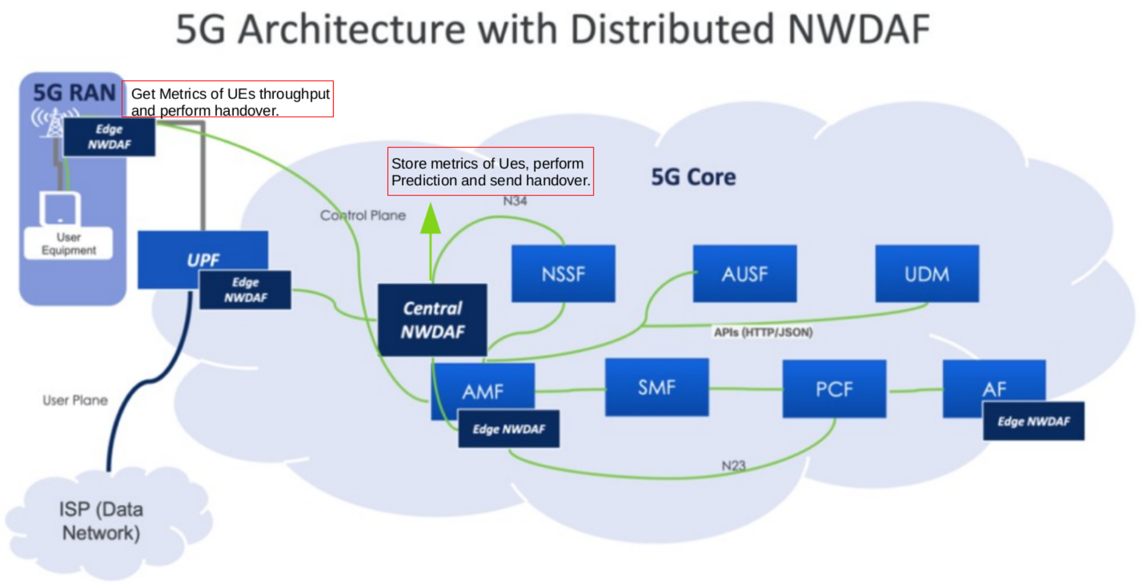
8.1 Summary and Conclusions

We conducted a brief historical analysis of the development of cellular networks in this thesis, understanding the variations and the evolution in network architecture from 1G to 5G. The various handover designs, architectures and protocols utilized in LTE systems were presented. Then, in NITOS-testbed, we installed a fully operational monolithic LTE network in a real-world environment, using the nextEPC as CN and OAI-eNB as RAN alongside with the FlexRAN controller and the third party software named FlexRAN scrapper, that was exclusively developed for this thesis. Using avialable datasets, we emulated mobility in a static UE scenario, and examined the intra-frequency handover performance, compared to the inter-frequency one. An evaluation of the performance of LSTM, XGBoost, SVR in predicting the future UL throughput demand of UEs was performed. Using the LTE infrastructure created for this thesis, we examined different handover policies and mechanisms in the Load Balancing concept that were designed, implemented and evaluated in a real-world environment.

8.2 Future Work

The Network Data Analytics Function (NWDAF) is a unique analytics function that has been described by the 3GPP in the Rel. 15 and 16 specifications for 5G. Using ML/AI algorithms, NWDAF derives the analytics and distributes them to other network functions (NFs). In the meantime, data from many sources (i.e., NFs) is needed for the analytics to be derived [16] [17]. For instance, data from the application function (AF), user plane function (UPF), session management function (SMF), and access and mobility management function (AMF) are needed for the user equipment (UE) communication pattern prediction. However, gathering a lot of input data might use up a lot of network resources and lead to issues with data security. The addition of the NWDAF component in 5G Core simplifies the interworking challenges for the providers of analytics solutions. The 3GPP also offers standard use cases that may be used by any operator to manage subscriber quality of experience (QoE) in a network that is constantly expanding and becoming more sophisticated.

In the distributed version of NWDAF, we integrate our architecture as shown in Figure 8.1:


5G Architecture with Distributed NWDAF

Get Metrics of UEs throughput and perform handover.

Store metrics of Ues, perform Prediction and send handover.

Figure 8.1: NWDAF integration

Multiple Edge NWDAF functions will be placed in the 5G RAN and be responsible for collecting all the necessary data for the UEs and forward them to the Central NWDAF. The Central NWDAF will store these metrics in a database, to retrain the model or to perform forecasting and also will choose the best candidate UE and target eNB. Then the handover command will be sent to the source eNB to perform the handover process.

Bibliography

[1] Wasan Kadhim Saad, Ibraheem Shayea, Bashar J. Hamza, Hafizal Mohamad, Yousef Ibrahim Daradkeh, and Waheb A. Jabbar. Handover Parameters Optimisation Techniques in 5G Networks. *Sensors*, 21(15), 2021.

[2] Marcelo Sampaio De Alencar and Djalma De Melo Carvalho Filho. *2 Cellular Communications Models*. 2017.

[3] SeungJune Yi, SungDuck Chun, YoungDae Lee, SungJun Park, and SungHoon Jung. *Overview of LTE and LTE□Advanced New Features*, pages 151–158. 2013.

[4] Mountassir Jamal, Balta Horia, Kovaci Maria, and Isar Alexandru. Study of multiple access schemes in 3GPP LTE OFDMA vs. SC-FDMA. In *2011 International Conference on Applied Electronics*, pages 1–4, 2011.

[5] Wang Anyi and Liu Zhaoyang. The study and analysis of MIMO technology based on LTE physical layer. In *2010 The 2nd Conference on Environmental Science and Information Application Technology*, volume 2, pages 306–308, 2010.

[6] Rupayali Swaroop and Arun Kumar. A brief study and analysis of NOMA techniques for 5G. In *2020 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE)*, pages 13–16, 2020.

[7] Sangchul Oh, Byunghan Ryu, and Yeonseung Shin. EPC signaling load impact over S1 and X2 handover on LTE-Advanced system. In *2013 Third World Congress on Information and Communication Technologies (WICT 2013)*, pages 183–188, 2013.

[8] Konstantinos Alexandris, Navid Nikaein, Raymond Knopp, and Christian Bonnet. Analyzing X2 handover in LTE/LTE-A. pages 1–7, 2016.

[9] NITLAB. `https://nitlab.inf.uth.gr/NITlab/`. Accessed: 16/08/2022.

[10] NITOS. `http://nitlab.inf.uth.gr/doc/index.html`. Accessed: 16/08/2022.

[11] NextEPC. `https://nextepc.com/open-source`. Accessed: 16/08/2022.

[12] Cleverson Nahum, José Soares, Pedro Batista, and Aldebaro Klautau. Emulation of 4G/5G Network Using OpenAirInterface. 01 2017.

[13] Xenofon Foukas, Navid Nikaein, Mohamed M. Kassem, Mahesh K. Marina, and Kimon Kontovasilis. FlexRAN: A Flexible and Programmable Platform for Software-Defined Radio Access Networks. In *Proceedings of the 12th International on Conference on Emerging Networking EXperiments and Technologies*, CoNEXT '16, page 427–441, New York, NY, USA, 2016. Association for Computing Machinery.

[14] Konstantinos Alexandris, Navid Nikaein, Raymond Knopp, and Christian Bonnet. Analyzing X2 handover in LTE/LTE-A. *2016 14th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*, pages 1–7, 2016.

[15] Tensorflow. `https://www.tensorflow.org/`. Accessed: 16/09/2022.

[16] Ali Chouman, Dimitrios Michael Manias, and Abdallah Shami. Towards Supporting Intelligence in 5G/6G Core Networks: NWDAF Implementation and Initial Analysis. In *2022 International Wireless Communications and Mobile Computing (IWCMC)*, pages 324–329, 2022.

[17] Salih Sevgican, Meriç Turan, Kerim Gökarslan, H. Birkan Yilmaz, and Tuna Tugcu. Intelligent network data analytics function in 5G cellular networks using machine learning. *Journal of Communications and Networks*, 22(3):269–280, 2020.