UNIVERSITY OF THESSALY

SCHOOL OF ENGINEERING

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

Design and Development of a Packet Routing Mechanism in Distributed 5th Generation Base Stations with Multiple Radio Access Technologies (multi-RAT)

MSc Thesis

Dimitrios Kefalas

Supervisor: Athanasios Korakis

July 2025

UNIVERSITY OF THESSALY

SCHOOL OF ENGINEERING

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

Design and Development of a Packet Routing Mechanism in Distributed 5th Generation Base Stations with Multiple Radio Access Technologies (multi-RAT)

MSc Thesis

Dimitrios Kefalas

Supervisor: Athanasios Korakis

July 2025

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Σχεδιασμός και Ανάπτυξη μηχανισμού δρομολόγησης πακέτων σε κατανεμημένους σταθμούς 5ης γενιάς πολλαπλών τεχνολογιών ασύρματης πρόσβασης (multi-RAT)

Μεταπτυχιακή Διπλωματική Εργασία

Δημήτριος Κεφαλάς

Επιβλέπων/πουσα: Αθανάσιος Κοράκης

Ιούλιος 2025

Approved by the Examination Committee:

Supervisor **Athanasios Korakis**
Professor, Department of Electrical and Computer Engineering, University of Thessaly

Member **Parisis Flegkas**
Assistant Professor, Department of Electrical and Computer Engineering, University of Thessaly

Member **Antonios Argyriou**
Associate Professor, Department of Electrical and Computer Engineering, University of Thessaly

DISCLAIMER ON ACADEMIC ETHICS AND INTELLECTUAL PROPERTY RIGHTS

Being fully aware of the implications of copyright laws, I expressly state that this MSc 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

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Professor Athanasios Korakis, for granting me the opportunity to engage in a cutting-edge research environment and for fostering my interest in this field through his guidance and support.

I am truly fortunate to have been mentored by Dr. Nikolaos Makris, Senior Researcher and Postdoctoral, whose constant availability, insightful guidance, and unwavering support were instrumental in helping me navigate the challenges encountered throughout this work.

I would also like to extend my heartfelt thanks to Assistant Professor Parisis Flegkas and Associate Professor Antonios Argyriou from the Department of Electrical and Computer Engineering at the University of Thessaly for their participation in the examination committee of my thesis.

MSc Thesis

Design and Development of a Packet Routing Mechanism in Distributed 5th Generation Base Stations with Multiple Radio Access Technologies (multi-RAT)

Dimitrios Kefalas

Abstract

5G networks represent a major technological evolution, aiming to meet diverse service requirements through enhanced Mobile Broadband (eMBB), ultra-reliable low-latency communications (URLLC) and massive machine-type communications (mMTC). To achieve these goals, the convergence of different Radio Access Technologies (RATs) under a unified system, known as multi-RAT, has been proposed. RATs in generally can be divided into 3GPP technologies (LTE and 5G NR) and non-3GPP technologies (WiFi). Focusing on 5G networks, integrating non-3GPP access can combine the strengths of both technologies; however, core-level solutions, such as the N3IWF, lack RAN coordination and joint resource management between NR and WiFi.

In this thesis, a unified RAN-level integration of non-3GPP technology, namely WiFi, into the 5G O-RAN architecture is proposed. WiFi is integrated as a DU within the O-RAN framework, enabling packet-level traffic steering across 5G and WiFi DUs. A user-based slicing scheme for the WiFi DU is also proposed. Utilizing the Near-RT RIC, we deployed two different xApps that use LSTM-based traffic forecasting to predict UE demand: The first xApp applies a non-prioritizing policy, using the 5G DU as the main interface and dynamically offloading or splitting traffic to the WiFi DU when demand exceeds capacity. The second xApp defines two sets of priority and non-priority UEs, with both groups initially utilizing the WiFi DU as their main interface. During traffic congestion, priority UEs are able to utilize up to 70% of the WiFi capacity, while the remaining bandwidth is allocated to non-priority UEs. When WiFi resources are insufficient, both priority and non-priority UEs are offloaded to the 5G DU to meet their demand. The proposed system was implemented and evaluated on a testbed with real-world datasets. Our findings highlight that the proposed O-RAN multi-RAT system, effectively exploits their complementary capabilities to satisfy UE requirements for QoS and traffic demand.

Μεταπτυχιακή Διπλωματική Εργασία

Σχεδιασμός και Ανάπτυξη μηχανισμού δρομολόγησης πακέτων σε κατανεμημένους σταθμούς 5ης γενιάς πολλαπλών τεχνολογιών ασύρματης πρόσβασης (multi-RAT)

Δημήτριος Κεφαλάς

Περίληψη

Τα δίκτυα 5G αντιπροσωπεύουν μία σημαντική τεχνολογική εξέλιξη, με στόχο την κάλυψη ποικίλων απαιτήσεων υπηρεσιών μέσω του enhanced Mobile Broadband (eMBB), των ultra-reliable low-latency communications (URLLC) και των massive machine-type communications (mMTC). Για την επίτευξη αυτών των στόχων έχει προταθεί η σύγκλιση διαφορετικών Τεχνολογιών Ασύρματης Πρόσβασης (Radio Access Technologies, RATs) σε ένα ενοποιημένο σύστημα, γνωστό ως multi-RAT. Γενικά, τα RATs μπορούν να διακριθούν σε τεχνολογίες 3GPP (LTE και 5G NR) και σε τεχνολογίες μη 3GPP (όπως το WiFi). Εστιάζοντας στα δίκτυα 5G, η ενσωμάτωση μη 3GPP πρόσβασης μπορεί να συνδυάσει τα πλεονεκτήματα και των δύο τεχνολογιών· ωστόσο, οι λύσεις επιπέδου πυρήνα, όπως το N3IWF, στερούνται συντονισμού στο επίπεδο του RAN και κοινής διαχείρισης πόρων μεταξύ NR και WiFi.

Στην παρούσα διπλωματική εργασία προτείνεται μία ενοποιημένη ενσωμάτωση της μη 3GPP τεχνολογίας, δηλαδή του WiFi, στο 5G O-RAN αρχιτεκτονικό πλαίσιο σε επίπεδο RAN. Το WiFi ενσωματώνεται ως Distributed Unit (DU) στο πλαίσιο του O-RAN, επιτρέποντας την καθοδήγηση της κίνησης σε επίπεδο πακέτου μεταξύ των DUs 5G και WiFi. Προτείνεται επίσης ένα σχήμα slicing βασισμένο σε χρήστες για το WiFi DU. Αξιοποιώντας το Near-RT RIC, αναπτύχθηκαν δύο διαφορετικά xApps που χρησιμοποιούν πρόβλεψη ζήτησης μέσω LSTM για την πρόβλεψη της ζήτησης των UEs: Το πρώτο xApp εφαρμόζει μία πολιτική χωρίς προτεραιότητες, χρησιμοποιώντας το 5G DU ως κύρια διεπαφή και δυναμικά εκφορτώνοντας ή καταμερίζοντας την κίνηση προς το WiFi DU όταν η ζήτηση υπερβαίνει τη χωρητικότητα. Το δεύτερο xApp ορίζει δύο ομάδες, με UEs προτεραιότητας και χωρίς προτεραιότητα, όπου και οι δύο αρχικά χρησιμοποιούν το WiFi DU ως κύρια διεπαφή. Σε περιόδους συμφόρησης, τα UEs προτεραιότητας μπορούν να αξιοποιήσουν έως και 70% της χωρητικότητας του WiFi, ενώ το υπόλοιπο εύρος ζώνης κατανέμεται στα UEs χωρίς προτεραιότητα. Όταν οι πόροι του WiFi δεν επαρκούν, τόσο τα UEs προτεραιότητας όσο και τα UEs χωρίς προτεραιότητα εκφορτώνονται στο 5G DU για την κάλυψη της ζήτησής τους. Το προτεινόμενο σύστημα υλοποιήθηκε και αξιολογήθηκε σε δοκιμαστική πλατφόρμα με πραγματικά δεδομένα. Τα αποτελέσματά μας καταδεικνύουν ότι το προτεινόμενο O-RAN multi-RAT σύστημα εκμεταλλεύεται αποτελεσματικά τις

συμπληρωματικές δυνατότητες των ετερογενών RATs για την κάλυψη των απαιτήσεων των UEs σε QoS και ζήτηση κίνησης.

Table of contents

List of figures

List of tables

Abbreviations

5G	Fifth Generation
A1	Interface between Non-RT RIC and Near-RT RIC
AI	Artificial Intelligence
AMF	Access and Mobility Management Function
AP	Access Point
AUSF	Authentication Server Function
BBU	Baseband Unit
BE	Best Effort
BK	Background
CDMA	Code Division Multiple Access
CSMA/CA	Carrier Sense Multiple Access / Collision Avoidance
CU	Centralized Unit
CU-CP	Central Unit – Control Plane
CU-UP	Central Unit – User Plane
DL	Downlink
Docker	(containerization platform)
DSCP	Differentiated Services Code Point
DSSS	Direct Sequence Spread Spectrum
DU	Distributed Unit
E2	Interface between Near-RT RIC and RAN
eNB	Evolved Node B
F1AP	F1 Application Protocol
F1oIP	F1 over IP
GTP-U	GPRS Tunneling Protocol - User plane
IP	Internet Protocol

IPsec	Internet Protocol Security
LSTM	Long Short-Term Memory
MAC	Medium Access Control
MAC address	Media Access Control address
MIMO	Multiple Input Multiple Output
ML	Machine Learning
MU-MIMO	Multi-User Multiple Input Multiple Output
NAS	Non-Access Stratum
NIC	Network Interface Card
N3IWF	Non-3GPP Interworking Function
NITOS	Network Implementation Testbed using Open Source
NR	New Radio
NRF	Network Repository Function
O1	Interface between SMO and network functions
O-RAN	Open Radio Access Network
OFDM	Orthogonal Frequency-Division Multiplexing
OFDMA	Orthogonal Frequency-Division Multiple Access
OvS / OVS	Open vSwitch
PCF	Policy Control Function
PDCP	Packet Data Convergence Protocol
PHY	Physical Layer
QoS	Quality of Service
RAN	Radio Access Network
RIC	RAN Intelligent Controller
RLC	Radio Link Control
RL	Reinforcement Learning
RNN	Recurrent Neural Network
RRC	Radio Resource Control
RRU	Remote Radio Unit
RU	Radio Unit
SC-FDMA	Single Carrier Frequency Division Multiple Access
SDAP	Service Data Adaptation Protocol

SDN	Software Defined Networking
SLICES-RI	Scientific Large-scale Infrastructure for CommunicationExperimental Studies – Resear
SL	Supervised Learning
SMF	Session Management Function
SMO	Service Management and Orchestration
SSID	Service Set Identifier
STA	Station
TCP	Transmission Control Protocol
TDMA	Time Division Multiple Access
ToS	Type of Service
UDP	User Datagram Protocol
UDM	Unified Data Management
UDR	Unified Data Repository
UL	Uplink
UPF	User Plane Function
VI	Video
VO	Voice
WiFi	Wireless Fidelity
WLAN	Wireless Local Area Network
xApp	(custom application running on Near-RT RIC)
rApp	(custom application running on Non-RT RIC)

Chapter 1

Introduction

5G networks represent a major technological evolution, designed to meet the diverse and demanding requirements of modern communication systems by enabling enhanced mobile broadband (eMBB), ultra-reliable low-latency communications (URLLC), and massive machine-type communications (mMTC). One direction that the scientific community has identified in order to address these heterogeneous service requirements is the convergence of different radio access technologies within a unified system. This approach, known as multi-Radio Access Technologies (multi-RAT), enables the coordinated operation of multiple radio technologies under a unified architecture to optimize resource utilization, enhance capacity, and support flexible service delivery.

In general, Radio Access Technologies (RATs) can be classified into two main categories: 3GPP and non-3GPP technologies. 3GPP technologies refer to radio systems standardized by the 3rd Generation Partnership Project, such as LTE and 5G NR, offering native support for mobility, QoS, and advanced radio features. Non-3GPP technologies, such as WiFi, operate in unlicensed spectrum and provide cost-efficient, high-bandwidth connectivity for local-area deployments.

The integration of non-3GPP radio access technologies into the 3GPP-based systems is motivated because of the complementary advantages that these technologies offer to each other. As already mentioned, 5G NR offers QoS, mobility support, and advanced radio features, while WiFi provides cost-efficient, high-bandwidth connectivity in unlicensed spectrum, ideal for indoor use and traffic offloading. By bridging these two worlds, operators can combine the benefits of both and create a flexible, scalable and more efficient access network.

The scientific community and industry operators have performed several standardization activities to define frameworks for integrating non-3GPP technologies into 3GPP systems. In 5G systems, the Non-3GPP Interworking Function (N3IWF) was introduced to support secure and efficient connection of non-3GPP user equipments with the 5G Core Network (5GCN). This function allows mobility management, establishment of sessions and integration of the user plane using standard interfaces. However, the integration based on N3IWF suffers from operating at the core network level rather than the RAN, leading to increased signaling overhead, higher latency, lack of unified scheduling, and limited low-layer coordination between 3GPP and non-3GPP technologies.

In this thesis, an alternative approach is explored: the direct integration of non-3GPP access, and in particular WiFi, in the MAC and PHY layers of the 5G Radio Access Network (RAN). In the proposed architecture, a non-3GPP technology namely the WiFi is converged into the 5G system architecture as a Distributed Unit (DU) in a disaggregated 5G RAN, and this enables joint scheduling, traffic steering and slicing over heterogeneous interfaces, while staying compatible with the 5G protocol stack. This low-layer integration enables the packet level routing decisions, allows a tighter coordination between the RATs, and removes the need for intermediary internetworking functions.

The work of this thesis addresses not only the technical challenges for integrating WiFi as DU—like the unified identification of the user, the adaptation of protocols and the enforcement of QoS—but also proposes and evaluates traffic steering methods based on the O-RAN architecture. In particular, controllers as xApps are developed, in order to dynamically provision the resources of the 5G and WiFi DUs according to predicted demand and priorities of the users. In this way, a flexible multi-RAT system is achieved, that can adapt in real-time depending on the network conditions, the user requirements and the service-level agreements.

In summary, this thesis is contributing to the current efforts for realizing unified multi-RAT 5G networks, by showing how non-3GPP technologies can be integrated in 5G RAN, and how this can enable smarter and more efficient utilization of the wireless resources in the future heterogeneous access networks.

1.1 Thesis' scope

The primary objective of this thesis is to design and implement traffic steering mechanisms in a multi-RAT environment, with a specific focus on the collaborative operation between WiFi and 5G within the RAN. More precisely, this work explores the integration of WiFi as an additional access technology within the 5G RAN, by embedding it functionally at the MAC level of the NR protocol stack. WiFi is treated as an independent DU, similar to the 5G DU, allowing both of them to be jointly utilized by the scheduler.

Based on this architecture, two different traffic steering policies are designed and implemented, both developed as xApps and executed on top of the O-RAN architecture, using the nearRT - RAN Intelligent Controller (nearRT-RIC) as the main control point.

The first policy follows a hierarchical approach where the 5G DU is the main interface, and the WiFi DU is only activated when the predicted demand exceeds the capacity of the 5G DU. If the total traffic demand exceeds the capacity of the WiFi link, then the traffic is split and routed simultaneously to both DUs.

The second traffic steering logic follows a more priority-aware model. Priority UEs are granted access to the WiFi DU first. Based on our experiments, the WiFi DU offers lower latency and higher throughput compared to the 5G DU, making it the primary interface. As a result, non-priority UEs also utilize it when sufficient resources are available. When congestion appears or priority UEs request more resources, a 70% bandwidth threshold is applied in favor of the priority UEs, and the remaining WiFi capacity is dynamically allocated to the non-priority ones. Additionally, the policy enables simultaneous routing over both interfaces (5G and WiFi), promoting flexibility and efficient resource utilization.

The thesis is developed and evaluated over the real-world testbed NITOS. In our setup, we generate realistic traffic using UDP throughput traces from a real-world dataset, while we also use a demand forecasting method based on Long Short-Term Memory (LSTM) neural networks trained on historical traffic patterns. This forecasting is then used as input to our traffic steering logic, allowing proactive resource allocation decisions.

To conclude, this thesis goes beyond a simple convergence of WiFi into 5G at the MAC layer. It also contributes to how traffic is scheduled in a multi-RAT RAN context, making a small but meaningful step towards smarter and more flexible resource management in next-generation access networks.

1.1.1 Contribution

The contribution of this thesis can be summarized as follows:

- Designed and implemented a multi-RAT O-RAN system with WiFi DU integration
- Designed and implemented a CU-scheduler for packet-level traffic steering among the Multi-RAT DUs
- Proposed a user-based slicing mechanism at the WiFi DU.
- Designed and implemented a non-prioritizing xApp for dynamic traffic steering using LSTM-based UE traffic demand prediction.
- Designed and implemented a prioritizing xApp that allocates WiFi DU capacity preferentially to priority UEs while offloading non-priority UEs to the 5G DU during congestion.
- Experimentally validated the system in a real testbed (NITOS/SLICES-RI) using real-world datasets to evaluate throughput, interface utilization, and latency.

1.2 Thesis Structure

This thesis is organized into six chapters. Chapter 1 introduces the general context of the work, starting with a brief historical overview of the evolution of cellular networks and WiFi technology. It also discusses the motivation behind the convergence of heterogeneous wireless access technologies, highlighting the need for unified operation within the RAN.

Chapter 2 provides the theoretical foundation of the work, presenting the essential concepts underlying 5G architecture, the O-RAN model, and the functionality of the RAN Intelligent Controller (RIC).

Additionally, Chapter 3 discusses related research works from the literature are reviewed, particularly those focusing on multi-RAT operation, traffic steering, and resource management in radio access networks.

In Chapter 4, the system architecture is described, where WiFi is integrated as an additional Distributed Unit (DU) at the MAC layer of the 5G NR protocol stack. The operation logic of the proposed system is analyzed in detail. It also presents the design and implementation of the two traffic steering algorithms, both developed as xApps operating over the

RIC. The first one follows a non-prioritizing hierarchical decision-making strategy, while the second introduces UE priority awareness and dynamic bandwidth allocation between users.

Chapter 5 presents the experimental platform and testbed environment employed for the development and validation of the proposed system. It details the experimental evaluation of both xApps, conducted on the real-world NITOS testbed. The performance analysis focuses on interface utilization, per-UE throughput distribution, and fairness within the dual-RAT setup.

Finally, Chapter 6 concludes the thesis with a summary of the findings and proposes directions for future work, such as the integration of slicing mechanisms and the exploration of reinforcement learning techniques for more intelligent and adaptive scheduling policies.

Chapter 2

Background

This section will analyze some basic features of the technologies used in the proposed system. Initially, the basic architecture of cellular networks will be presented, giving emphasis to the 5G, followed by a discussion regarding the WiFi technology and its integration aspects.

2.1 Cellular Networks - From Analog to Broadband: The Path to 4G

Cellular networks have experienced rapid technological advancements in recent decades. In particular, the traditional, monolithic architectures of earlier generations of cellular networks (from 1G to 4G) have evolved into highly flexible, disaggregated, virtualized, and cloud-based systems (5G and beyond).

1G

To understand the overall evolution of cellular networks, it is worthwhile to present some brief details of each generation. Starting with the 1G cellular network, the first generation of cellular networks, which supported only Analog voice transmission, had small data transmission capabilities, and operated in the 800MHz band. It was characterized by enormous energy consumption and a lack of security systems.

The system architecture of the 1G system as depicted in Figure 2.1 presents significant differences from the architectures of later cellular networks. It was the first one to introduce the innovation of organizing an area into cells, where each cell contains a base station (BS) responsible for wireless communication with user equipment (UE). The BS and the Mobile

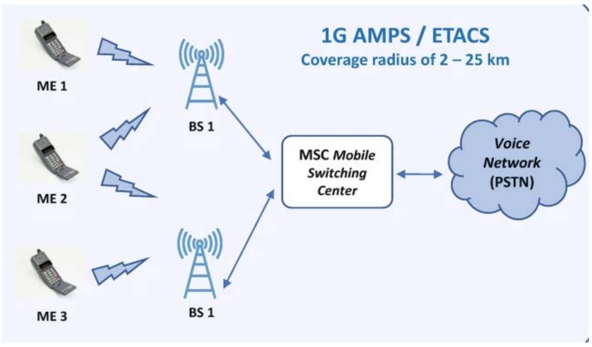

Figure 2.1: 1G system architecture

Switching Center (MSC) communicated with wired lines. An MSC was primarily responsible for basic connectivity such as connecting base stations, routing voice calls and others. Each MSC was responsible to communicate with the Public Switched Telephone Network (PSTN) in order to enable external voice communication.

2G

The second generation of cellular networks, 2G, was the first cellular network capable to support digital transmission. It introduced the digital voice communication through GSM and enabled the basic data services such as SMS and MMS. 2G was able to operate in multiple bands such as 850 MHz, 900 MHz, 1800 MHz, 1900 MHz, achieving better spectral efficiency through the proposed access techniques such as the Time Division Multiple Access (TDMA) and the Code Division Multiple Access (CDMA), offering better call quality.

2G inherited its system architecture from the 1G, with one key difference: the existence of the Base Station Controller (BSC), which is located between the BTSs and MSC and is responsible for handling Radio channel allocation, Handovers, and Power level control, as well as Frequency hopping. Although 2G inherits the same concept of cells and centralized switching from 1G, its architecture is more layered and functionally advanced marking a clear evolution rather than a reuse of the 1G architecture. The system architecture of the 2G cellular networks is depicted in Figure 2.2.

3G

The third generation of cellular networks, commercial named 3G, it introduced significant improvements in mobile data transmission rates and multimedia capabilities over its predescessors, paving the way for the upcoming mobile internet era. As its primary radio

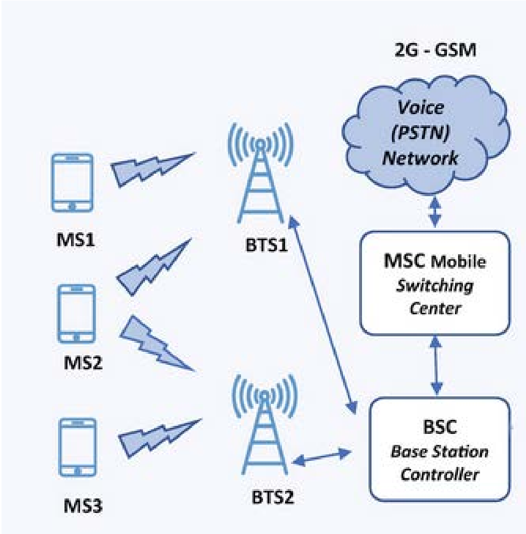

Figure 2.2: 2G system architecture

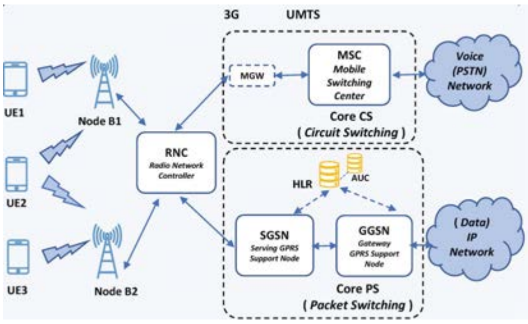

Figure 2.3: 3G system architecture

access technology, 3G utilized the Wideband Code Division Multiple Access (W-CDMA), having peak data rates of 2 Mbps for stationary users and around 384 kbps for the mobile ones, operating under various frequency bands such as 850 MHz, 900 MHz, 1700 MHz, 1900 MHz, and 2100 MHz. 3G was the first cellular technology that introduced a packet-switched data oriented architecture, making feasible the mobile internet connectivity and building the foundation for creating IP-based mobile networks. The aforementioned network transformation from traditional circuit-switched voice services to packet-switched oriented networks is one of the most significant evolutions in the cellular networks.

Focusing more to the 3G architecture, we can easily understand that it builds upon the foundation of 2G by maintaining the existing voice-centric MSC and introduced two new key components: the Universal Terrestrial Radio Access Network (UTRAN) and the evolved Core Network, which had as responsibility to support internet connectivity (see Figure 2.3).

As already mention, with the integration of these new entities, 3G managed to achieve a seamless transition from traditional voice communication to a more data-oriented mobile experience.

4G

The next cellular generation of the 3G, namely the 4G, played a crucial role in the evolution of the system architecture of cellular networks since it redesigned an entirely all-IP based system architecture, which significantly increased data rates and enabled seamless support for real-time applications. This all-IP based architecture eliminated all the circuit-switched components and supported both voice and data transmissions over a unified, packet-switched IP network. Significant evolution was not only limited to the system architecture but also took place at the RAN with the adoption of OFDMA (Orthogonal Frequency-Division Multiple Access) in the downlink and SC-FDMA in the uplink, which offer high spectral efficiency, low latency, and improved data transmission speeds. These transmission speeds could reach up to 100 Mbps under ideal conditions. This high transmission rate capability occurred due to the adaptation of the MIMO (Multiple Input Multiple Output) antenna technology, which further enhanced throughput and network reliability of the 4G RAN. Depending on the region in which the 4G network is deployed, 4G supports a wide range frequency operating bands, offering greater flexibility and capacity than 3G.

The 4G system architecture can be divided into two main domains: the Evolved UMTS Terrestrial Radio Access Network (E-UTRAN) and the Evolved Packet Core (EPC), as depicted in Figure 2.4. Both of these two components which constitute the Evolved Packet System (EPS). Focusing more on the RAN part, the E-UTRAN, is consisted of two primary components: the User Equipment (UE) and the eNodeB (Evolved Node B). The eNodeB which actually is the base station of the 4G era, it is responsible for control and user plane operations such as managing the radio resources, performing handovers, maintaining the Quality of Service (QoS), and forwarding user data to the EPC. As already mentioned, it handles both user plane and control plane operations and interfaces directly with EPC components, specifically the Mobility Management Entity (MME) and the Serving Gateway (S-GW).

The EPC is consisted of four key entities: the Mobility Management Entity (MME), the Serving Gateway (S-GW), the Packet Data Network Gateway (P-GW), and the Home Subscriber Server (HSS). The MME is responsible for control plane functions, such as control-signaling, user authentication, session establishment, and bearer management. The EPC in-

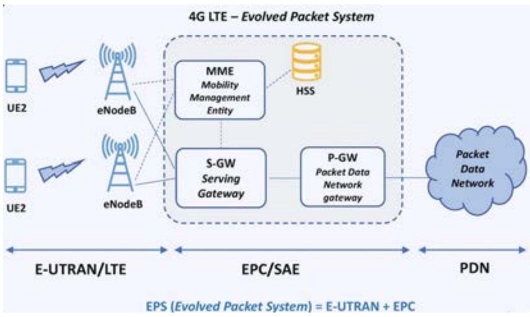

Figure 2.4: 4G system architecture

teracts with the HSS to get the subscriber profiles and authentication credentials. The S-GW is responsible for forwarding user plane traffic between the eNodeB and the P-GW and for maintaining seamless data flow as the user moves between different cells.

The P-GW connects the LTE network to the external Packet Data Networks (PDNs), such as the global internet, and also is responsible for IP addressing, maintaining QoS policies, and applying traffic management rules. Finally, the HSS serves as subscriber database responsible for storing authentication information, mobility data, and service authorizations.

To sum up, 4G was a turning point, marking the shift from centric systems to a distributed, all-IP architecture capable of supporting application with high-speed data transmission and real-time communication demands. This evolution expanded mobile network capabilities but also raised user demands for greater transmission speed, lower latency and reliability. As 4G was not designed to meet the enormous requirements of today UEs which want the support of emerging use cases like massive IoT and ultra-low-latency control, it paved the way for 5G, which introduces a cloud-native, service-based architecture to address these new challenges.

2.2 5G: A New Era in Mobile Network Architecture

5G is the direct successor to 4G in the evolution of cellular technologies and was developed to overcome the limitations of earlier generations. Its primary objective is to enable more flexible and scalable network architectures, offering significantly improved performance in terms of latency, reliability, and data throughput. These enhancements aim to meet the growing demands of users and support a wide range of emerging applications, including real-time services, and large-scale IoT deployments. Furthermore, 5G aims to transform cellular networks into cloud-native environments through the adoption of a service-based architecture, converting all 5G entities into virtualized network functions designed to support scalability and dynamic resource orchestration.

5G also introduces three primary service categories namely Ultra-Reliable Low-Latency Communication (URLLC), enhanced Mobile Broadband (eMBB), and massive Machine-Type Communication (mMTC), aiming to enable the support for a broad range of demanding and diverse applications.

- Ultra-Reliable Low-Latency Communication (URLLC): URLLC aims to achieve both ultra-high reliability and ultra-low latency, two performance metrics that have traditionally been difficult to achieve simultaneously. This service focuses on meeting the needs of mission-critical applications that require minimal delays and data loss.

- Enhanced Mobile Broadband (eMBB): eMBB aims to support data-intensive applications by addressing the demand for significantly higher data rates and increased network capacity. It focuses to provide seamless high-speed connectivity by exploiting capabilities of previous mobile generations, offering improved spectral efficiency, wider bandwidth, and advanced antenna technologies such as massive MIMO.

- Massive Machine-Type Communication (mMTC): mMTC is designed especially to support the connectivity needs resulting from the Internet of Things (IoT) devices. The network behaviour these devices have is to generate small, infrequent data transmissions. mMTC focuses on enabling energy-efficient and cost-effective communication for low-power devices, often deployed in dense or remote environments.

To support the aforementioned services, 5G utilized multiple technologies in order to deal with the corresponding demands. The following examples are indicative:

- Millimeter Wave (mmWave): With the usage of mmWave, 5G can expand the operational band frequencies above 24 GHz, aiming to achieve very high data rate transmission speed. This technology is especially suitable for short-range communication.

- Massive Multiple Input Multiple Output (MIMO): Massive MIMO is the evolution of the MIMO technology used in LTE It can utilize up to hundreds of antennas at the BS, enabling multiplexing and serving multiple users simultaneously, improving the spectral efficiency and the network capacity.

- Beamforming: This technology aims to direct the radio signals transmitted from the BS toward specific connected devices, reducing the interference while simultaneously increasing the signal strength.
- Network Slicing: Network slicing is responsible for creating multiple virtual networks on top of the same physical infrastructure. Each slice serves specific service requirements, and isolation between slices is ensured to improve resource utilization.
- Virtualization: 5G utilizes virtualization to decouple hardware from software functions. This adaptable architecture streamlines network management, enhances resource allocation, and speeds up the deployment of new services.

2.2.1 5G System Architecture

5G has two main architectures for deploying it in real networks: Non-Standalone Architecture (NSA) and Standalone Architecture (SA).

In the NSA, the 5G network is deployed with the support of a CN derived from 4G, supporting some functions from 5G. This architecture connects eNBs and gNBs to the same CN while they can directly communicate via the Xn interface. Thus, NSA networks can support higher data rates, lower latency, and increased capacity compared to 4G, as 5G innovations at the RAN level have been incorporated into the system. However, since the CN is limited to 4G functionalities, NSA lacks features such as network slicing and other capabilities introduced in the 5G Core.

As mentioned above, 5G, to meet the increasing demands of customers in high-volume data transmission and low-latency demands of services, had to adopt more flexible and distributed architectures. 5G utilizing the Network Function Virtualization (NFV) technology creates a service-based architecture (SBA) at both CN and RAN levels to solve the problems of monolithic architectures of previous generations. Thus, the 5G CN is transformed into a cloud-native deployment through the decomposition of the functions of the 4G CN entities. More specifically, the basic 5G CN now consists of the following entities:

- Authentication Server Function (AUSF): The AUSF handles user authentication in 5G, replacing the Home Subscriber Server (HSS) authentication functions in LTE.
- Unified Data Management (UDM): The UDM manages subscriber data and policies, serving a role similar to the HSS and parts of the PCRF in LTE.

- User Data Repository (UDR): The UDR acts as the central data storage for network functions, equivalent to the database portion of the HSS in LTE.
- Network Repository Function (NRF): The NRF, which allows dynamic discovery of network functions, has no direct equivalent in LTE, as EPC functions are statically configured.
- Access and Mobility Management Function (AMF): The AMF manages user access and mobility, replacing the Mobility Management Entity (MME) in LTE.
- Session Management Function (SMF): The SMF handles session setup and traffic path selection, splitting the role of the LTE Serving Gateway (SGW) and Packet Data Network Gateway (PGW).
- User Plane Function (UPF): The UPF is the data forwarding plane in 5G, taking over the user plane responsibilities of the SGW and PGW in LTE.

In SA 5G, the network is deployed using both the SBA-based 5G CN and various 5G RAN implementations. A key characteristic of the SBA in the 5G CN is the separation of network functions into Control Plane (CP) and User Plane (UP) entities, named Control User Plane Separation (CUPS). Entities belonging to the UP are responsible for transmitting data packets, while those belonging to the CP are responsible for creating control and signaling messages. This design decision is reflected in the structure of the 5G protocol stack, which clearly separates control and user plane protocols to align with the functional split of network entities.

5G NR Protocol Stack

This paragraph will present all protocols constituting the 5G NR User and Control plane Protocol stack. The protocols will be presented in a bottom-up approach, starting from the layers that are common to both the control and user planes:

- PHY (Physical Layer): Modulates and transmits data over the air, handles MIMO, coding, beamforming.
- MAC (Medium Access Control): Responsible for scheduling, multiplexing, HARQ, and prioritization of data.

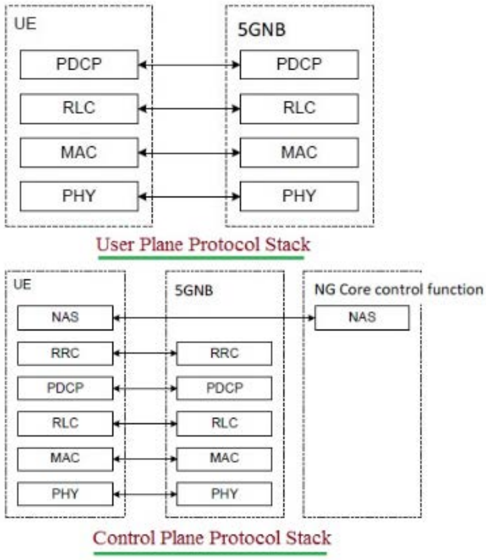

Figure 2.5: 5G NR Protocol Stack Control & User Plane

- RLC (Radio Link Control): Manages segmentation/reassembly, ARQ-based retransmissions, and in-sequence delivery.

These three layers form the foundation of both data and signaling traffic processing between the UE and the gNB.

The User Plane Specific Layers:

- PDCP (Packet Data Convergence Protocol): Handles user data encryption, header compression, and in-sequence delivery. Used specifically for transferring user traffic.

The protocols that implement the control plane Protocol stack are the following:

- PDCP (Control Plane use): Similar to user plane PDCP, but handles only control signaling data (e.g., RRC messages).

- RRC (Radio Resource Control): Manages connection setup/release, mobility, handovers, security config. Present only in the control plane.

- NAS (Non-Access Stratum): It is located above RRC, connects UE directly with the 5GC (AMF), handling authentication, registration, and session control.

Figure 2.5 shows all the protocols previously presented in both the control and user plane protocol stacks.

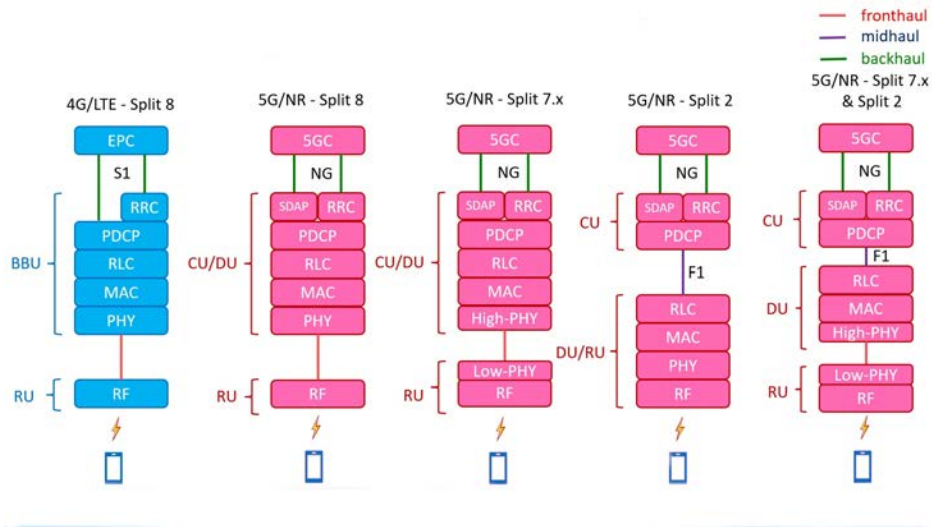

Figure 2.6: RAN Functional Splits

2.2.2 5G RAN Functional splits

The primary architecture of the RAN in LTE relies on the D-RAN architecture. The CN is connected via the backhaul interface to the Baseband Unit (BBU), which is responsible for digital processing and various baseband functions. Then the BBU is connected to the Remote Radio Unit (RRU) via the Fronthaul interface, which is responsible for transmitting and receiving wireless signals. In this monolithic architecture, the BBU and RRU are typically located together at the cell site.

Moving to the 5G era, the concept of dividing the computational processing of the BBU to multiple nodes to support the growing demands and address scalability challenges dominated. 3GPP has divided the BBU entity into two components: the Centralized Unit (CU), responsible for network layer functionality, and the Distributed Unit (DU), which hosts the Data Link Layer. Different functional splits were proposed based on the protocol stack, as depicted in Figure 2.6.

In this thesis, the functional split that will be used is the 3GPP Option 7.X, also referred to as the PDCP/RLC split. In this configuration, the Central Unit (CU) is responsible for the SDAP, RRC, and PDCP protocols, while the DU handles the RLC, MAC, PHY, and RF functions.

In modern 5G disaggregated deployments, the DU is further split into the DU and the Radio Unit (RU). In this architecture, the RU is responsible for the lower physical layer (Low PHY) and the radio frequency (RF) functions, while the DU handles the upper PHY, MAC, and RLC layers. This separation allows for greater flexibility in deploying RAN components, improved scalability, and more efficient utilization of edge and centralized resources.

The communication between the CU and the DUs has been standardized through the F1 interface, also known as the midhaul. The F1 Application Protocol (F1AP) enables interaction between the CU and multiple DUs, supporting various wireless technologies such as 5G, LTE, and WiFi from the RAN perspective. While each DU is connected to a single CU (one-to-one), a CU can serve multiple DUs (one-to-many), enabling flexible and scalable deployments. This architecture also supports multiple cells, as each DU can manage one or more, facilitating the implementation of micro-, pico-, and femto-cell networks.

2.2.3 O-RAN Architecture and Principles

Although the disaggregation and virtualization of the 5G RAN has solved many challenges that traditional monolithic approaches faced due to their limited flexibility and scalability, it did not fully resolve critical issues such as optimizing radio resource management and efficient spectrum utilization through real-time adaptation. To solve these limitations, the O-RAN architecture was introduced to enable the reconfigurability of the RAN via open and interoperable interfaces. O-RAN has three core principles: intelligence, openness, and interoperability.

Intelligence

O-RAN introduces intelligence into the RAN through two specialized controllers: the Near-Real-Time RIC and the Non-Real-Time RIC. The Near-RT RIC manages latency-sensitive functions (e.g., handovers, interference control) via modular xApps, while the Non-RT RIC handles higher-level tasks like policy management and ML model training. Another key component in RAN management is the Service Management and Orchestration (SMO) framework, responsible for overseeing configuration, monitoring, and lifecycle management across the RAN.

Openness

O-RAN standardizes three key open interfaces E2, A1, and O1 that enable seamless interaction between components from different vendors, promoting interoperability and vendor diversity. The E2 interface connects the Near-RT RIC to the disaggregated components of

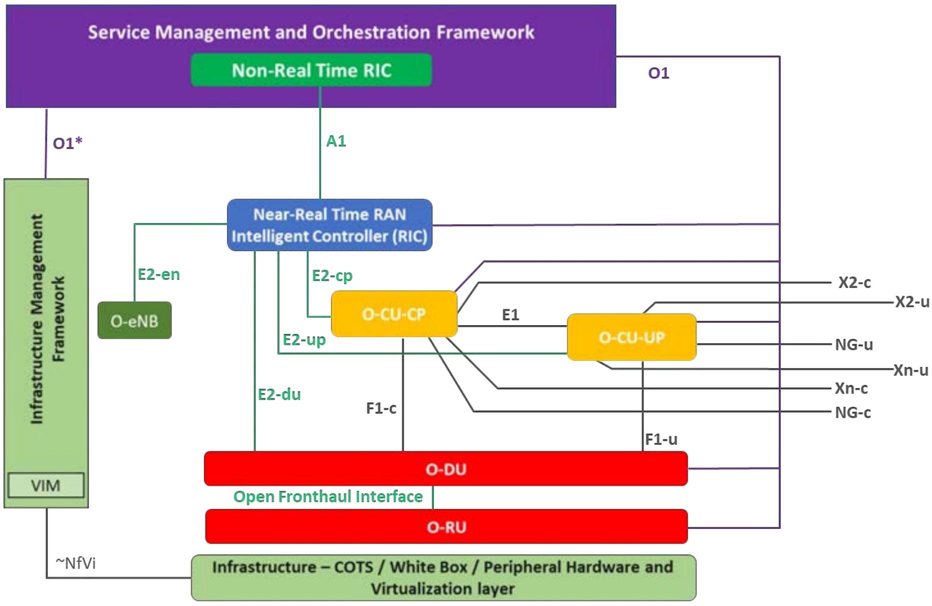

Figure 2.7: O-RAN architecture

the gNB, such as the CU and DU, each of which is considered an E2 node. The A1 interface allows the Non-RT RIC to provide policy guidance and ML model updates to the Near-RT RIC, while the O1 interface enables the SMO to manage and monitor all network functions. This open interface design ensures flexible integration and facilitates innovation across the RAN ecosystem. The overall architecture including these open interfaces is illustrated in Figure 2.7.

Interoperability

O-RAN integrates the Interoperability in the RAN with the usage of xApps and rApps—software applications that run on the Near-RT RIC and Non-RT RIC, respectively. A key enabler of this modularity is the use of xApps and rApps—software applications that run on the Near-RT RIC and Non-RT RIC, respectively. xApps and rApps aims to be vendor-neutral, allowing operators to deploy, update, or replace RAN functions and policies on the air such as traffic prediction and management, or fault detection and others.

2.3 WiFi

WiFi protocols are currently the most widespread wireless data transfer technologies, without need for cables. The first WiFi protocol was released in 1997, and since then, multiple

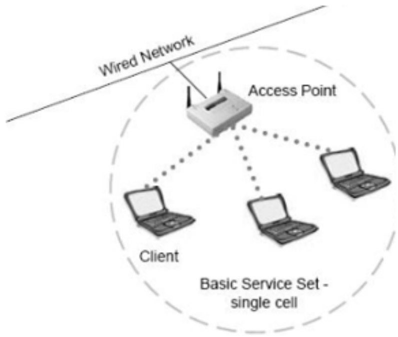

Figure 2.8: WLAN architecture

upgrades and optimizations have been developed in lacking generations, making it the leading player in Wireless Local Access Networks (WLAN).

A WLAN typically follows a star topology, in which all wireless devices, known as stations (STAs), communicate through a central node called the Access Point (AP). The AP is the bridge between the wireless client and the chosen wired network. This topology is depicted in Figure 2.8.

The WiFi protocol stack, based on IEEE 802.11, follows the OSI model, and mainly focus on the PHY and MAC layers. The PHY layer handles modulation, coding, and transmission over the radio channel using techniques like DSSS, OFDM, or MIMO depending on the standard. The MAC sublayer manages medium access using CSMA/CA, ensuring collision avoidance and reliability through acknowledgements and retransmissions. Higher layers, such as network and transport, remain protocol-agnostic, allowing seamless integration with TCP/IP networks.

The first WiFi standard, named IEEE 802.11, was introduced in 1997 and operated in the 2.4 GHz ISM band, supporting a maximum theoretical data rate of 2 Mbps. It employed Direct Sequence Spread Spectrum (DSSS) and Frequency Hopping Spread Spectrum (FHSS) at the physical layer and used the Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA) protocol for medium access control. Due to its low data rate and limited range, it was primarily suitable for basic wireless communication tasks and served as the foundation for later, more advanced WiFi standards.

The second WiFi standard, IEEE 802.11b, was introduced in 1999 and also operated in the 2.4 GHz frequency band. It significantly improved the maximum data rate, reaching up to 11 Mbps, while maintaining compatibility with the original 802.11 standard. Like its predecessor, it employed DSSS at the physical layer and utilized CSMA/CA as its medium access control protocol. One of its key advantages was its backward compatibility, allowing seamless coexistence with 802.11-compliant devices, which facilitated broader adoption in consumer and enterprise environments.

The IEEE 802.11a standard was released alongside 802.11b in 1999 but adopted a different approach by operating in the 5 GHz frequency band instead of 2.4 GHz. It introduced OFDM at the physical layer, enabling significantly higher data rates of up to 54 Mbps. Like previous standards, it relied on the CSMA/CA protocol for medium access control. Despite offering better spectral efficiency and reduced interference due to the less crowded 5 GHz band, its adoption was initially slower because of higher hardware costs and shorter range compared to 802.11b. Moreover, it was not backward compatible with 802.11b due to the different frequency bands.

IEEE 802.11g was introduced in 2003 as a hybrid solution combining the best features of 802.11a and 802.11b. It operated in the 2.4 GHz band—ensuring backward compatibility with 802.11b—while adopting OFDM as its modulation scheme, like 802.11a, to achieve data rates up to 54 Mbps. The use of CSMA/CA remained consistent as the medium access protocol. By offering higher speeds without sacrificing compatibility with existing 2.4 GHz devices, 802.11g became widely adopted and played a key role in expanding WiFi deployment in both home and business environments.

IEEE 802.11n, released in 2009, marked a significant advancement in WiFi technology by introducing support for both the 2.4 GHz and 5 GHz frequency bands (dual-band operation). It increased theoretical data rates up to 600 Mbps through the introduction of Multiple Input Multiple Output (MIMO) technology, channel bonding (up to 40 MHz), and frame aggregation techniques. The standard maintained CSMA/CA for medium access but greatly improved spectral efficiency and throughput. Backward compatibility with 802.11a/b/g ensured smooth integration into existing networks, making 802.11n one of the most influential and widely adopted WiFi standards at the time.

IEEE 802.11ac, also known as WiFi 5, was introduced in 2013 and operates at the 5 GHz frequency band. WiFi 5 is able to support wider channels of 80 and 160 MHz, having higher-order modulation (up to 256 QAM), and Multi-User MIMO (MU-MIMO) in the DL. As its maximum theoretical data rate, it has 1 Gbps, making it the best candidate for high-bandwidth applications such as HD video streaming and real-time communication. Like its predecessors, it utilizes CSMA/CA for medium access control.

The integration of WiFi protocols into the 5G architecture unleashes significant potential to create more flexible, scalable, and cost-efficient wireless networks. 5G can benefit from the exploitation of unlicensed spectrum in the 2.4 GHz and 5 GHz bands. Operators can offload traffic intelligently, extend indoor coverage, and reduce deployment costs due to the significantly lower price of WiFi access points. The synergy between WiFi and 5G paves the way for advanced applications in smart cities, industrial automation, and high-density enterprise environments.

2.4 Machine Learning - LSTM

AI is one of the most innovative technologies that has been adopted in a wide range of applications in today's world. AI is concerned with creating systems that are capable of performing tasks that require intelligence such as making decisions and predicting future values. It has already been integrated into several areas in computer science such as networks, operating systems, digital signal processing, and many others. 5G networks have already integrated the use of AI through the O-RAN ecosystem providing suitable components for the design, training and use of AI systems as we saw in the O-RAN subsection 2.2.3. The applications of AI in 5G address various challenges, such as predicting network data requirements, forecasting UE mobility, making decisions for tasks like network slicing, and many more.

One of the most widespread AI systems is Machine Learning, in which systems can be trained from data by recognizing patterns and behaviors of the data set and improve their performance without being explicitly programmed. The use of Machine Learning is particularly adopted in applications such as recommendation systems, image recognition, and traffic prediction.

ML can be classified into the following categories:

- Supervised Learning (SL) – learns from labeled data (e.g., classification, regression).
- Unsupervised Learning (UL) – finds structure in unlabeled data (e.g., clustering, dimensionality reduction).
- Reinforcement Learning (RL) – learns optimal actions through interaction with an environment using rewards and penalties.

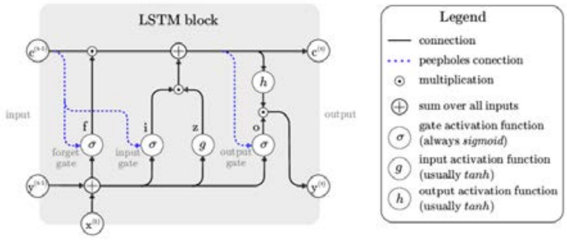

Figure 2.9: LSTM architecture

Among the aforementioned ML categories, SL is particularly relevant for time-series prediction tasks, where historical input-output pairs are used to train models. A widely used SL method for sequential data is the Recurrent Neural Network (RNN), which can capture temporal patterns by retaining information from previous time steps. However, standard RNNs often struggle with learning long-term dependencies due to the vanishing gradient problem. To address this, Long Short-Term Memory (LSTM) networks were introduced. LSTMs enhance traditional RNNs by introducing memory cells and gating mechanisms that help preserve relevant information over longer sequences, making them particularly effective for time-series prediction tasks.

In the Figure 2.9 below the gates of the lstm are illustrated:

- The forget gate decides which parts of the previous cell state to discard.
- The input gate determines which new information will be added to the cell state, combining the sigmoid output and the candidate values produced by the tanh layer.
- The updated cell state is computed by combining the retained past memory and the new information.
- Finally, the output gate regulates what part of the updated cell state is passed to the next hidden state, again using tanh and sigmoid activations.

2.5 Conclusion

In summary, this chapter provided a detailed overview of the technological foundations relevant to the design and development of the proposed multi-RAT 5G system. It presented the evolution of cellular networks from the analog 1G all the way through the virtualized and disaggregated 5G, highlighting key innovations such as functional splits, O-RAN principles, and the integration of WiFi as a complementary access technology. Furthermore, it examined the important role of ML and particularly the LSTM networks in enabling intelligent, data-driven decision-making within the network. This section built the foundation for understanding the system-level contributions and mechanisms proposed in the following chapters.

Chapter 3

Related Work

In the effort to integrate non-3GPP Distributed Units (DUs) into the disaggregated 5G Radio Access Network (RAN) architecture, it is essential to follow the frameworks and guidelines defined for untrusted non-3GPP access in 5G systems. From the early stages of cellular evolution, particularly in the 3G era, the integration of WLAN technologies had already been identified as a key objective. In [1], WLANs were proposed as independent components interfacing with the core network (PLMN) via proxy servers, rather than being integrated into the UTRAN access layer.

With the advent of 4G, 3GPP Release 12 introduced two methods for WLAN integration: LTE-WLAN Aggregation (LWA) and LTE-WLAN Radio Level Integration with IPsec Tunnel (LWIP). LWA operates at the PDCP layer [2], while LWIP uses IP-layer tunneling through IPsec connections [3].

The formalization of non-3GPP integration matured in 5G Release 15, which introduced the Non-3GPP Interworking Function (N3IWF) [4], a core network function that interfaces non-3GPP UEs with the 5G Core Network (5GCN). N3IWF acts as a gateway, communicating with the AMF, SMF, and UPF to manage mobility, session establishment, and user-plane tunnels via GTP-U. In [5], the integration of WiFi and satellite systems through the N3IWF component is examined, showing that the introduced overhead is minimal when processing standard WiFi traffic. Likewise, the work in [6] explore the authentication and authorization mechanisms involved in integrating WiFi access into 5G via the N3IWF, and provides useful insight regarding convergence and security design aspects.

At the same time, 3GPP 5G-NR Rel. 16 [7] has standardized the notion of Access Traffic Steering, Switching and Splitting (ATSSS). This newly introduced functionality allows traffic steering across multiple access technologies, allowing high granularity beyond a single PDU session [8]. A Multi-Access PDU session is introduced for each UE, allowing data to be served over one or more simultaneous connections (3GPP access, non-3GPP access, untrusted non-3GPP access). ATSSS resides on the core network side, and allows end-to-end PDU sessions to be established over multiple RATs, by making use of a Multipath-TCP (MPTCP) proxy established with the UPF function.

In parallel, 5G introduced the concept of network slicing, enabling multiple logical networks to coexist over shared infrastructure. Extending the slicing concept into the RAN, known also as RAN slicing, brings several challenges especially because of the unstable radio environment and constraints in available spectrum. The work in [9] deals with the problem of interference between slices by proposing a heuristic algorithm for resource block allocation, aiming to achieve more efficient and also fair resource distribution in networks with multiple slices.

Although WiFi does not implement slicing in the same way as 5G, recent standards such as IEEE 802.11ax and 802.11be have incorporated mechanisms like Multi-SSID (Multi-SS) to support user separation and QoS differentiation [10]. However, these mechanisms primarily address bandwidth efficiency and user isolation rather than fine-grained, traffic-aware resource allocation.

Several research efforts have explored WiFi slicing from an architectural and algorithmic perspective. In [11], FlowVisor and OpenFlow are used to implement isolated slices by managing flows and hardware resources such as CPU and forwarding tables. In [12], slicing is treated as a stochastic optimization problem with the goal of ensuring minimum bitrates and bounded queue delays through packet scheduling. At the MAC level, [13] proposes a time-division approach based on pausing and resuming mac80211 queues, enabling fine-grained, runtime slice control via TDMA-like mechanisms.

Some recent works have further explored the performance implications of hybrid 5G/WiFi deployments in practical settings. For instance, [14] proposes a dual-zone architecture in which WiFi-6 serves the inner zone of a cell and 5G the outer zone, leveraging spatial distribution to balance traffic load. While this approach highlights the benefits of multi-RAT coexistence, it relies on static assumptions and cannot dynamically adapt to user mobility or real-time demand. Likewise, in [15], WiFi mesh links are examined in the context of Integrated Access and Backhaul (IAB). Even though they are cost-efficient, such links brings variability in both latency and throughput, more noticeable in multi-hop topologies. The study points out the importance of using traffic steering mechanisms and choosing dynamic paths, in order to reduce the performance drops that may happen in these kind of hybrid backhaul configurations. Unlike the approach in [16], which relies on CQI prediction and a centralized traffic steering logic for selecting between LTE and WiFi DUs using pre-defined thresholds, this thesis propose a MAC-level integration of WiFi as a Distributed Unit (DU) inside the 5G RAN. Furthermore, it introduces two xApp-based traffic steering schemes that runs on top of the O-RAN Near-RT RIC and supports per-UE routing with prioritization and flexible bandwidth allocation in real time.

Despite the significant progress in integrating non-3GPP technologies like WiFi into 5G networks and in exploring slicing mechanisms in isolated environments, most existing works consider either high-layer interworking solutions (e.g., N3IWF) or MAC-level slicing in standalone WiFi systems. These approaches often lack real-time traffic adaptability and unified control under a common 5G RAN architecture. Moreover, prior research generally treats cellular and non-cellular access as loosely coupled components, without a shared, packet-level steering logic. In contrast, this thesis introduces a unified MAC-layer integration of WiFi as a fully DU within the disaggregated 5G RAN, enabling joint scheduling across heterogeneous interfaces. It further proposes and evaluates two xApp-based traffic steering policies executed on the Near-RT RIC, showcasing how demand-aware, prioritization-sensitive scheduling can be achieved within a hybrid multi-RAT setup.

Chapter 4

System Architecture

4.1 Introduction

This chapter describes the system architecture that integrates the necessary components and modifications for a 5G system to support multi-RAT integration, with a primary focus on WiFi. The role and usage of each component in our system architecture are subsequently explained, along with a discussion of the potential limitations of our implementation. As mentioned earlier, the reference 5G stack that will be utilized in order to perform the integration is the 3GPP Option 2 split described in [17] and demonstrated in [18], [19], in which CU includes the protocols of the PDCP layer and above, and DU includes the RLC layer and below.

4.2 Enabling Multi-RAT Support in 5G-RAN

5G and WiFi protocol stacks present significant differences, since these wireless technologies were designed with different objectives. 5G aims to support connectivity in wide-area mobile networks, guaranteeing QoS and mobility control, in contrast to WiFi, which is optimized for local-area access, being simpler and flexible. 5G has been specifically designed to support a full core network stack for session management, mobility, and user authentication, whereas WiFi mainly acts as a wireless last-hop AP, bridging end-user devices to the broader IP network without offering native core network functionalities.

Although the lower protocol stacks share the same layers (MAC and PHY), 5G and WiFi have a totally different point of view regarding the functionality of these layers. The MAC layer in 5G is scheduled and fully controlled by the gNB or the DU to coordinate transmissions. In contrast, the MAC layer in WiFi is contention-based, relying on CSMA/CA, where devices compete for access, resulting in unpredictable delays. The PHY layer in 5G is designed for high performance and flexibility across various spectrum bands offering features like scalable OFDM numerology, massive MIMO and tight synchronization for mobility. On the other hand, WiFi's PHY is optimized for simplicity and short-range operation in unlicensed bands with less emphasis on synchronization and mobility support.

Despite these differences, interaction between the MAC and PHY layers remains essential, as this is the point where RAT interface directly with the rest of the RAN. By performing the WiFi integration in the 5G-RAN at the MAC/PHY layers, it becomes feasible to incorporate non-3GPP technologies like WiFi into the 5G architecture without compromising the integrity or functionality of the core 5G protocol stack.

By integrating non-3GPP technologies such as WiFi into the 5G architecture at the MAC and PHY layers, we preserve the core functionalities of the 5G protocol stack without the need to introduce additional components like the N3IWF. However, approaches that rely on higher-layer interworking mechanisms tend to introduce complexity and latency when they are integrated in sensitive networks such as the 5G one.

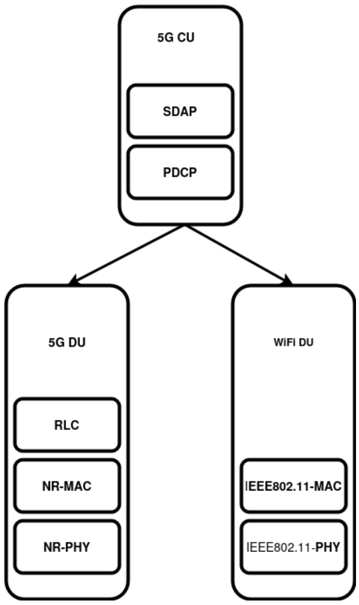

Figure 4.1: Hybrid 5G RAN architecture with support for WiFi-based DU

In this implementation, the goal is to replace the 5G MAC and PHY layers at the DU side with their WiFi counterparts, effectively incorporating a WiFi-based DU into the 5G RAN. This approach aims to enhance the performance of the RAN by leveraging the advantages of

WiFi—such as low-cost deployment, operation in unlicensed spectrum, and efficient traffic offloading—while still preserving the upper-layer functionalities of the 5G stack at the CU. By maintaining compatibility with the 5G core and control architecture, this design enables seamless multi-RAT integration within a unified framework.

The network stack concept of our proposed integration is illustrated in Figure 4.1. As is depicted, the 5G CU keeps the standard 5G protocol stack at the upper layers, the SDAP and the PDCP, which are responsible for QoS mapping, encryption, and in-sequence delivery. The interesting part takes place at the DU level, where two parallel implementations are supported: a typical 5G DU, which includes RLC, NR-MAC and NR-PHY and the WiFi-based DU, which replaces the lower layes of the typical 5G protocol stack with EEE 802.11 MAC and PHY. This hybrid configuration enables the CU to interface with either a traditional 5G radio stack or a WiFi-based radio access, allowing for flexible and efficient multi-RAT support while maintaining compatibility with the 5G CN.

Since this work aims to support hybrid DU implementations, it is necessary to discuss the required operations over the F1 protocol related to DU registration, capability exchange, and data transfer processes. The communication between the 5G CU and a standard 5G DU follows the typical F1 Application Protocol (F1AP) as defined by 3GPP, operating over the traditional GTP-U tunnel for user-plane data forwarding. However, to support the integration of WiFi-based DUs, we utilized a custom protocol referred to as F1 over IP (F1oIP) based on TCP. The TCP usage was preferred over the UDP one to ensure reliable, in-order delivery of packets, which is essential for maintaining the integrity of the F1 protocol in WiFi-based DU integration.

To integrate WiFi DUs into the 5G architecture, we define three key operational processes within the custom F1oIP communication protocol:

- **Registration Process:** Enables initial connectivity and topology awareness between the CU and WiFi DU, establishing their identities, roles, and available capabilities.

- **Capability Exchange Process:** Allows the CU and WiFi DU to communicate supported functionalities, ensuring compatibility and appropriate resource management.

- **Data Transfer Process:** Manages the user-plane transmission of PDCP SDUs between CU and WiFi DU, encapsulating data within a unified and technology-agnostic packet format.

Registration

In our proposed architecture, the registration and discovery process for the WiFi DU does not follow the traditional approach defined by 3GPP. In our case the CU is the one responsible to perform a registration request to the WiFi DU, not the opposite way as happens in typical 3GPP systems. The decision for this reversal implementation was motivated by the role of the WiFi DU integration, which is implemented as a specific feature of the modified CU. This approach enhances security, as the CU maintains explicit control over the WiFi DU association process, ensuring only authorized and pre-configured WiFi DUs are accepted.

Capability

The proposed system is designed to support multiple non-3GPP wireless technologies, such as mmWave, WiFi, and others. Therefore, after completing the registration process, the WiFi DU advertises its DU Type ID to the 5G CU, indicating the specific technology supported by the DU. Knowing the DU type enables the CU to efficiently manage network resources, steer traffic appropriately, and balance load across multiple RATs. Additionally, it ensures correct packet handling, protocol compatibility, and awareness of each DU's supported capabilities and features.

Data Transfer Process

The F1oIP protocol defines how data is transferred between the CU and WiFi DU over the midhaul interface. For downlink (DL) traffic, the CU sends data packets encapsulated in F1oIP headers, including necessary identifiers such as DU type and packet information. On the uplink (UL), the WiFi DU encapsulates user payloads received from UEs with PDCP headers and F1oIP metadata before transmitting them back to the CU. This allows both ends to handle data consistently, regardless of the underlying wireless technology.

Putting everything together, F1oIP is a lightweight protocol that enables seamless communication between the 5G CU and non-3GPP DUs, such as the WiFi DU used in this implementation. It ensures unified data and control exchange across heterogeneous access technologies while preserving compatibility with the 5G core.

4.3 WiFi-DU Functionality

In a heterogeneous RAN environment where multiple DUs—such as 5G NR and WiFi-based units—coexist, a fundamental challenge arises in consistently identifying UEs across different access technologies. Each RAT uses its own identification scheme: 5G NR relies on RNTIs (Radio Network Temporary Identifiers), while WiFi operates based on MAC addresses. Since these identifiers are not shared or synchronized across DUs, the CU cannot directly associate incoming or outgoing packets with the correct UE unless a translation mechanism is in place. This mismatch complicates unified scheduling, traffic steering, and session management. To address this issue, a mapping layer is introduced at the WiFi DU to translate between RNTIs and MAC addresses. In the downlink, the DU maps the RNTI provided by the CU to the corresponding MAC address of the target UE. In the uplink, the DU extracts the MAC address from incoming frames and resolves it to the correct RNTI for encapsulation before forwarding to the CU.

Different procedures are applied depending on whether data is routed back to the CU through the F1 interface or directly transmitted to UEs via the WiFi NIC on the DU. This distinction is necessary due to the dual role of the WiFi DU, which functions both as part of the 5G disaggregated RAN and as an IEEE 802.11 AP. As an AP, the WiFi DU handles standard WiFi procedures such as association, authentication, and link management, requiring UEs to establish a connection using the 802.11 protocol stack.

In the UL direction, data packets are received from connected UEs through the WiFi stack and processed by the DU to extract the payload, typically at the IP layer or above. The WiFi DU receives data frames from UEs over the IEEE 802.11 stack. Upon reception, it identifies the source MAC address and performs a MAC-to-RNTI lookup to assign the correct RNTI. Before forwarding these packets to the CU, the DU encapsulates the payload in a PDCP header to ensure compatibility with the PDCP instance at the CU. This step is crucial for maintaining uniformity in user-plane data handling across both 5G and WiFi DUs.

Conversely, in the DL direction, data from the CU arrives at the DU already encapsulated in a PDCP PDU. The WiFi DU first removes the PDCP header and extracts the raw payload. Before transmitting the packet, the DU is responsible for adding the appropriate IEEE 802.11 MAC header and mapping the RNTI to the corresponding MAC address of the target UE. This mapping ensures that each packet is correctly addressed at the link layer. The finalized frame is then passed to the WiFi NIC driver for wireless transmission over the 802.11 interface. The above-described procedure is depicted in Figure 4.2.

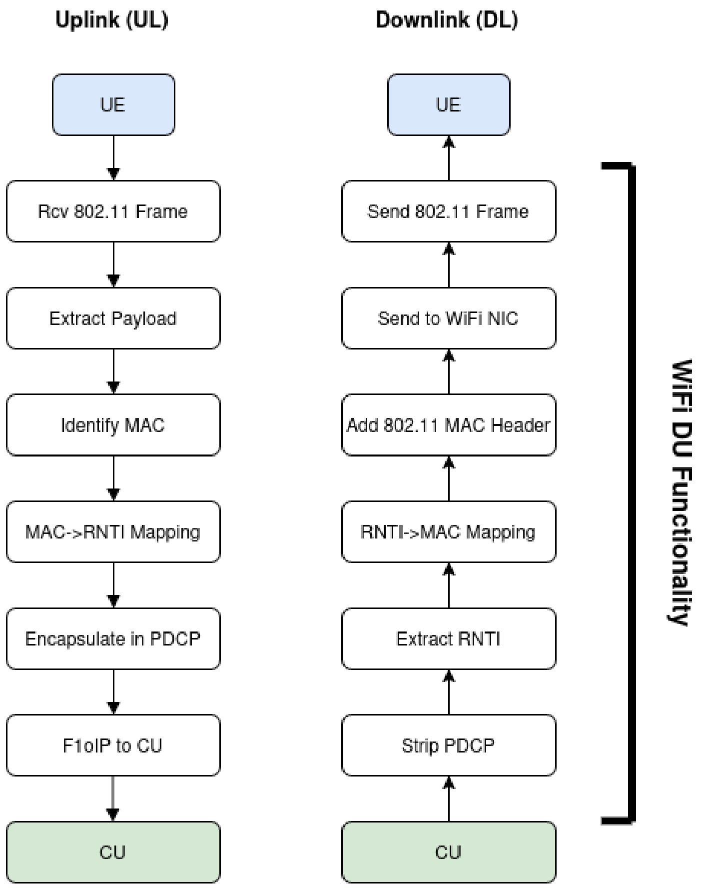

Figure 4.2: WiFi DU Functional Workflow for UL and DL Data Paths

4.3.1 Slicing in WiFi-DU

One of the primary objectives in the integration of WiFi as a non-3GPP DU in 5G RAN architectures is to extend slicing capabilities of a typical 5G system to the newly added WiFi DU entity. The proposed slicing mechanism for the WiFi DU exploits the Quality of Service (QoS) features of the IEEE 802.11 standard to enforce user-based slicing at the data plane using software-defined networking (SDN) components.

The proposed WiFi DU slicing mechanism utilizes the Type of Service (ToS) field in the IP packet headers, which corresponds to the Differentiated Services Code Point (DSCP). This field is used to classify and prioritize traffic flows, mapping each user or service to a specific Access Category (AC) in the MAC layer. The mac80211 driver, which is used in Linux-based WiFi implementations, provides four hardware-supported priority queues: Background (BK), Best Effort (BE), Video (VI), and Voice (VO). Each queue is associated with different medium access parameters, allowing higher-priority traffic (e.g., Voice) to gain preferential access to the wireless channel as shown in Table 4.1.

Table 4.1: Mapping of UP, ACs and MQs

User Priority	Access categories (ACs)	mac80211 queue (MQs)
1	Background (BK)	3
2		
0	Best Effort (BE)	2
3		
4	Video (VI)	1
5		
6	Voice (VO)	0
7		

User Priority (UP) value acts as an essential mapping layer between the ToS/DSCP field and the MAC-layer ACs. Specifically, each ToS value is translated into a corresponding UP, as defined by the IEEE 802.11 standard, and each UP is then mapped to a particular AC. This process ensures that packets are enqueued into the appropriate MAC queue (MQ) based on their QoS class.

Table 4.2: ToS-UP mapping

ToS val	User Priority
0-31	0
32-63	1
64-95	2
96-127	3
128-159	4
160-191	5
192-223	6
224-255	7

Slicing at the WiFi DU is happening by dynamically adjusting the ToS field of outgoing IP packets , in accordance with the slicing policies advertised from a controller. For each DL flow, the WiFi DU's agent intercepts the traffic and assigns an appropriate ToS/DSCP value, reflecting the intended slice requirements. This ToS value is mapped to the corresponding User Priority (UP), which determines the Access Category (AC)—namely Background (BK), Best Effort (BE), Video (VI), or Voice (VO)—within the MAC layer, as supported by the mac80211 driver. Through this process, packets are steered into distinct hardware-backed MAC queues, enabling the WiFi DU to enforce prioritization among users or services and effectively allocate wireless resources in line with the network slicing framework.

4.3.2 Anatomy of WiFi DU

Figure 4.3: WiFi DU Functional Workflow for UL and DL Data Paths

The proposed architecture of the WiFi DU is presented in Figure 4.3. As described in the previous section, we implement user-based slicing at the WiFi level by assigning different ToS values to IP packets depending on their destination UE, based on earlier scheduling decisions. To keep the configuration delay as low as possible, we avoided changing the ToS values at the application layer using standard libraries. Instead, we forward the packets to an Open vSwitch (OvS) instance, which is programmed with flow rules that handle the necessary ToS modifications. As shown in Figure 4.3, to supportslices with different priorities, we connected as many virtual Ethernet (veth) interfaces as the ACs along with the WiFi NIC to the OvS. Since veth devices are created in pairs, we added extra virtual interfaces to complete the setup—one end of each pair connects to OvS, while the other receives packets from our agent. On the agent side, in the downlink data path, we updated the application so that packets are routed through the appropriate veth interface depending on their assigned priority. OvS changes the ToS value of each packet based on the interface it enters through and forwards it to the WiFi NIC. Packet routing between the high- and low-priority paths is adjusted every few packets, depending on the client or destination IP, to match the desired traffic distribution ratio.

4.4 5G CU

For the 5G CU to be able to support a seamless integration of the WiFi-DU, it needs to adopt a centralized PDCP aggregation mechanism. In the proposed system architecture, PDCP sequence numbers play a pivotal role in preserving packet order and ensuring reliable data delivery across multiple DUs. In case each DU was responsible for maintaining its own independent PDCP instance and sequence number management, significant problems would arise related to sequence number consistency, resulting in potential issues such as packet re-ordering and loss at the UE side. By centralizing the PDCP sequence number management at the CU, all the 5G data traffic in the uplink (UL)—independently from the selected DU—is assigned sequence numbers from a unified PDCP context per UE. This centralized mechanism is particularly necessary in the UL direction, as a multi-homed UE can generate uplink traffic through both the 5G and WiFi interfaces, making it essential for the PDCP sequence number to be assigned at the CU to maintain ordering and integrity across heterogeneous paths. This centralized approach guarantees that packets can be correctly reassembled and delivered in order, even when traffic is split dynamically between 5G and WiFi DUs as part of the proposed load balancing strategy.

A specific scheduler must be developed at the CU level in order to support traffic distribution among the heterogeneous DUs from the CU. In a multi-RAT environment, traffic conditions, user demands, and radio resources are highly dynamic and may vary per user or service slice. Static rules or conventional round-robin algorithms do not support the flexibility to adapt to rapid fluctuations in channel quality, network congestion, or evolving QoS requirements. Contrary to these, the introduction of a dedicated scheduler that adapts forwarding decisions per UE enables the enforcement of advanced slicing policies, efficient resource utilization, and dynamic service differentiation. In the proposed system, the scheduler is integrated at the CU between the PDCP and RLC layers in order to enable dynamic forwarding decisions per packet to the 5G DU or WiFi DU, based on xAPP's decision. Although this approach introduces additional processing latency, the overall latency is minimized due to the performance impact of delay-sensitive services.

The presented scheduler implements a per-UE, proportional traffic allocation policy between the WiFi DU and the 5G DU. At the beginning of each scheduling interval, the scheduler receives, for each UE, the data traffic and the target WiFi traffic ratio. Using these parameters, it computes the total number of packets to be transmitted and the number of packets to be forwarded to the WiFi DU and to the 5G DU, respectively.

As packets arrive during the interval, the scheduler extracts the RNTI from each packet to identify the corresponding UE. For each UE, it maintains counters tracking the number of packets sent to each interface. The scheduler then checks whether the number of packets already sent to the WiFi DU is less than or equal to the computed quota. If so, it forwards the incoming packet to the WiFi DU and increments the corresponding counter. Otherwise, the packet is forwarded to the 5G DU.

This approach ensures that, over each interval, the traffic for every UE is distributed between the WiFi and 5G DUs according to the forecasted volume and the specified allocation ratio, thereby enabling fine-grained control over load balancing and resource usage across heterogeneous access technologies.

The proposed scheduler implements the algorithm depicted in 1. In this algorithm the traffic of each UE is distributed proportionally between the WiFi DU and the 5G DU in each scheduling interval. Each predefined interval, the scheduler receives from the xApp the total forecasted data volume and the target WiFi allocation ratio for each UE. To calculate the number of packets for each UE that must be sent to each DU, the scheduler first determines the total number of packets based on the forecasted data volume and the predefined packet size. Then, the ratio provided by the xApp determines the number of packets that should be forwarded to the WiFi DU.

When a packet arrives, the scheduler extracts the RNTI from each packet and identifies the UE destination. For each UE counters are maintained to track the packets forwarded to the WiFi DU and the 5G DU within the interval. If the counter for WiFi DU is less than the computed quota, the packet is forwarded to the WiFi DU, and the counter is incremented. Otherwise, the packet is sent to the 5G DU.

Algorithm 1 Scheduler Algorithm

Require: For each UE u: ratio_wifi[u], bw_forecast[u], rnti[u]; constant pkt_len

for each scheduling interval **do**
  for each UE u **do**
    num_packets_total[u] $\leftarrow \left\lfloor \frac{\text{bw\_forecast}[u]}{\text{pkt\_len}} \right\rfloor$
    num_packets_wifi[u] $\leftarrow$ $\lfloor$num_packets_total[u] $\times$ ratio_wifi[u]$\rfloor$
    num_packets_5g[u] $\leftarrow$ num_packets_total[u] $-$ num_packets_wifi[u]
    count_wifi[u] $\leftarrow 0$
    count_5g[u] $\leftarrow 0$
  end for
  for each incoming packet **do**
    Extract RNTI from packet and identify UE u
    if count_wifi[u] $<$ num_packets_wifi[u] **then**
      Forward packet to WiFi DU
      count_wifi[u] $\leftarrow$ count_wifi[u] $+ 1$
    else
      Forward packet to 5G DU
      count_5g[u] $\leftarrow$ count_5g[u] $+ 1$
    end if
  end for
end for

4.5 Experiment Assumptions

Before presenting the scheduling algorithms and the criteria we followed, we should first discuss the policy with which the tested 5G DU shares its resources among the UEs. When a UE connects to the 5G system, the DU assigns radio resources, ensuring fair allocation among the active UEs, in case the UEs belong to the same slice. The available DL Resource Blocks (RBs) are evenly divided among the connected UEs. Subsequently, if N UEs are connected to the same 5G DU and have similar channel conditions, then each of them is assigned with $total\_RBs/N$ RBs per transmission interval, resulting in nearly equal maximum DL throughput.

Compared to 5G, WiFi follows an entirely different strategy. Regarding the DL, a WiFi AP does not perform fair allocation in advance but in a more dynamic way. WiFi does not evenly share the medium among all connected STAs (stations), but only among those actively transmitting or receiving at a given moment. As long as the capacity of the wifi channel is not filled, active STAs can transmit or receive as many data packets as they require. Only when the demands of the active clients reach the capacity of the WiFi channel does the contention-based access mechanism of the WiFi result in almost fair resource sharing among them.

4.6 Scheduling xApps

For the proposed system to forward packets effectively across the multi-RAT DUs, the aforementioned scheduling algorithm needs to be managed from a dedicated xAPP deployed at the near-RT RIC. In this section, two different xAPPs will be proposed, the first one applies a fair-share, non-prioritizing scheduling policy. In contrast, the second one is based on a prioritization scheme using the WiFi DU slicing described in Section 2, aiming to enhance the QoS by prioritizing latency-sensitive UEs in the 5G multi-rat environment.

4.6.1 Non-Prioritizing Multi-RAT Scheduling with Demand-Aware WiFi Allocation

The first xAPP implements a Non-Prioritizing Multi-RAT Scheduling with Demand-Aware WiFi Allocation scheme. This algorithm aims to distribute traffic across the 5G and WiFi DU without performing UE prioritization. This xAPP can be divided into three phases. Initially, the xApp connects to the Near-RT RIC, retrieves a pre-trained machine learning model from the Service Management and Orchestration (SMO) framework, and subscribes to the E2 Agent's Radio Config (RC) Key Performance Metrics (KPM), to obtain the KPI DRB.UEThpDl from the CU node, which is the only E2 node that registers with the near-RT RIC.

In the first phase (Phase 1), the xAPP continuously monitors the reported DL throughputs of the UEs. In the second phase (Phase 2), at every predefined interval T, the xApp uses the pre-trained LSTM model to forecast the UE throughput for the next two time steps, based on a sliding window of the past 10 measurements. In the last phase (Phase 3), xAPP performs the routing decision for the packets. The core idea is to assign each UE to a DU that can independently satisfy its forecasted demand. In case the users' requirements exceed the capacity of the WiFi DU and the 5G DU, then a converging policy arises that forwards the packet to both DUs simultaneously. Based on the discussion in Section 4.5, 5G performs fair resource sharing, as its capacity is evenly distributed among all connected UEs, whereas WiFi utilizes a contention-based access scheme. Therefore, in the following algorithm, the 5G capacity is evenly shared among all connected UEs. In contrast, the WiFi capacity is evenly distributed only among the active UEs—and only when their aggregate demand reaches the total channel capacity. Given the presence of a single Access Point (AP), this distribution does not rely on CSMA/CA contention mechanisms, but is instead explicitly enforced by our fair-share algorithm, which deterministically allocates resources based on current UE demand. The WiFi model introduced in this xApp is not entirely accurate, but it is a high-level abstraction that effectively approximates WiFi performance. This simplification enables fast estimation of the dynamic channel conditions, essential for maintaining near real-time responsiveness within the xApp operating on the Near-RT RIC. All the aforementioned resulted to following: The scheduler follows a hierarchical decision process. First, all traffic is routed to the 5G DU, which is treated as the main interface. If the predicted UE demand is higher than the capacity of the 5G DU, then all traffic gets forwarded to the WiFi DU instead. In case the total demand is even higher than what the WiFi DU can handle alone, then the traffic is splitted between both 5G and WiFi DUs at the same time, filling first the 5G DU and send the remaining to the WiFI DU. After completing the routing decision, the xApp informs the CU scheduler the forecasted demand of each UE, along with the corresponding packet ratio to be routed through the WiFi DU. The described algorithm can be seen at Algorithm 2.

4.6.2 Prioritizing Multi-RAT Scheduling with Demand-Aware WiFi Allocation

The first two phases of the proposed algorithm that implements the Prioritizing Multi-RAT Scheduling with Demand-Aware WiFi Allocation are the same as those in the Prioritizing Multi-RAT Scheduling with Demand-Aware WiFi Allocation. In the third phase (Phase 3), the xApp performs the core routing decision based on the predicted demands. The key innovation of this scheduling policy is that it prioritizes certain UEs, explicitly defined in a priority set, granting them preferential access to the WiFi DU with the ability to utilize up to 70% of the channel's capacity. If the total demand from these priority UEs exceeds the available WiFi capacity, the WiFi bandwidth is equally divided among them. Only after this allocation the algorithm evaluates whether any WiFi bandwidth remains to be offered to the non-priority UEs.

Algorithm 2 xApp-Non-Prioritizing Multi-RAT Scheduling with Demand-Aware WiFi Allocation

```
Require: constants TOTAL_5G_CAPACITY, TOTAL_WIFI_CAPACITY
Require: Pretrained LSTM Model Phase 1 – Monitoring:
Receive current thrDl[u] from E2 interface
Phase 2 – Prediction:
prediction[u] ← LSTM_forecasting(curr_thr_Dl[u])
FAIR_5G_SHARE ← TOTAL_5G_CAPACITY / #UEs
Phase 3 – Routing Decision:
for i = 0 to prediction_step step s do
    wifiCandidates ← {}
    for each UE u do
        if prediction[u][i] > FAIR_5G_SHARE then
            ADD u TO wifiCandidates
        end if
    end for
    totalWifiDemand ← Σ_{u∈wifiCandidates} prediction[u]
    if totalWifiDemand ≤ TOTAL_WIFI_CAPACITY then
        for each u IN wifiCandidates do
            wifiShare[u] ← prediction[u]
        end for
    else
        share ← TOTAL_WIFI_CAPACITY / |wifiCandidates|
        for each u IN wifiCandidates do
            wifiShare[u] ← share
        end for
    end if
    for each UE u do
        if prediction[u + i] ≤ FAIR_5G_SHARE then
            route[u] ← 5G_DU_only
            wifiRatio[u] ← 0
        else if u ∈ wifiCandidates and prediction[u + 1] ≤ wifiShare[u] then
            route[u] ← WiFi_DU_only
            wifiRatio[u] ← 1
        else
            route[u] ← Split_5G+WiFi
            wifiRatio[u] ← (prediction[u] − FAIR_5G_SHARE) / prediction[u]
        end if
    end for
    SEND routing decisions, wifiRatio TO CU via E2 control
    SEND prioritized UEs to WiFi-DU via custom E2 interface
    SEND forecasted Throughput per UE TO CU via E2 control
end for
```

Algorithm 3 xApp-Prioritizing Multi-RAT Scheduling with WiFi Allocation to Priority UEs

```
Require: Constants TOTAL_5G_CAPACITY, TOTAL_WIFI_CAPACITY
Require: Pretrained LSTM model, Priority UE set P
Phase 1 – Monitoring:
Receive current thrDl[u] from E2 interface for all UEs
Phase 2 – Prediction:
prediction[u] ← LSTM_forecasting(thrDl[u])
FAIR_5G_SHARE ← TOTAL_5G_CAPACITY / #UEs
Phase 3 – Routing Decision:
for i = 0 to prediction_horizon step s do
    priorityWiFiCap ← 0.7 * TOTAL_WIFI_CAPACITY
    nonPriorityWiFiCap ← 0.3 * TOTAL_WIFI_CAPACITY
    priorityUEs ← {u ∈ P | prediction[u][i] > 0}
    nonPriorityUEs ← {u ∉ P | prediction[u][i] > 0}
    {– Allocate WiFi to Priority UEs –}
    totalPriorityDemand ← Σ_{u ∈ priorityUEs} prediction[u][i]
    if totalPriorityDemand ≤ priorityWiFiCap then
        for each u ∈ priorityUEs do
            wifiShare[u] ← prediction[u][i]
        end for
        priorityWiFiCap ← priorityWiFiCap - totalPriorityDemand
    else
        share ← priorityWiFiCap / |priorityUEs|
        for each u ∈ priorityUEs do
            wifiShare[u] ← share
        end for
        priorityWiFiCap ← 0
    end if
    {– Allocate WiFi to Non-Priority UEs –}
    totalNonPriorityDemand ← Σ_{u ∈ nonPriorityUEs} prediction[u][i]
    if totalNonPriorityDemand ≤ nonPriorityWiFiCap then
        for each u ∈ nonPriorityUEs do
            wifiShare[u] ← prediction[u][i]
        end for
        nonPriorityWiFiCap ← nonPriorityWiFiCap - totalNonPriorityDemand
    else
        share ← nonPriorityWiFiCap / |nonPriorityUEs|
        for each u ∈ nonPriorityUEs do
            wifiShare[u] ← share
        end for
        nonPriorityWiFiCap ← 0
    end if
    {– Final Routing and Ratio Computation –}
    for each UE u do
        demand ← prediction[u][i]
        wifiUsed ← wifiShare.get(u, 0)
        if wifiUsed ≥ demand then
            route[u] ← WiFi_DU_only
            wifiRatio[u] ← 1
        else
            g5Alloc ← min(FAIR_5G_SHARE, demand - wifiUsed)
            served ← wifiUsed + g5Alloc
            if wifiUsed > 0 then
                route[u] ← Split_5G+WiFi
            else
                route[u] ← 5G_DU_only
            end if
            wifiRatio[u] ← wifiUsed / demand
        end if
    end for
    SEND routing decisions, wifiRatio TO CU via E2 control
    SEND forecasted throughput per UE TO CU via E2 control
end for
```

4.7 Final System Architecture

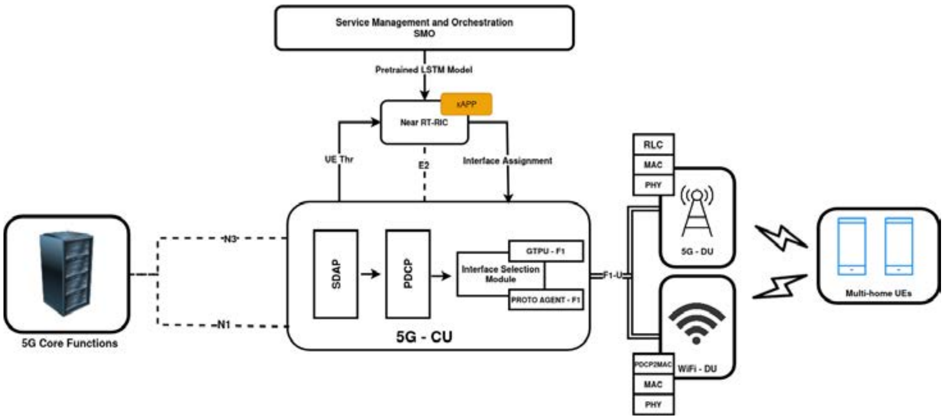

Figure 4.4: 5G Multi-RAT system architecture

Putting everything together, the final system architecture is depicted in Figure 4.4, which proposes a 5G multi-RAT architecture integrating 5G RAN and WiFi DUs under a unified 5G CU. The CU includes SDAP, PDCP, and an interface selection module that routes traffic either to the 5G DU via GTP-U or to the WiFi DU via F1oIP over TCP. A Near-RT RIC hosts two xApps: a non-prioritizing xApp implementing fair-share scheduling, and a prioritizing xApp that applies slice-aware decisions using DSCP-based classification at the WiFi DU. Both xApps utilize a pretrained LSTM model from the SMO to predict UE Thr and send interface assignment instructions over E2. The WiFi DU replaces RLC/MAC/PHY with PDCP2MAC, 802.11 MAC, and PHY, enabling packet translation and slicing via UP-to-AC mapping. Multi-homed UEs receive dynamically steered traffic per xApp policy, while centralized PDCP sequence number management at the CU ensures sequence consistency across RATs.

In this chapter, the system architecture of a 5G RAN network that supports a WiFi-based DU was presented. It introduces the architectural and functional differences between 5G and WiFi stacks. At the same time, it proposes a hybrid integration of the WiFi DU into the 5G RAN at the MAC/PHY layer using the 3GPP Option 2 split. Following, the custom data plane protocol (F1oIP over TCP) for the communication between the 5G CU and the WiFi-based DU was presented. Moreover, the internal mechanisms of the WiFi DU were analyzed, including user identification, uplink/downlink packet translation, and MAC-layer slicing through DSCP-to-AC mapping. The role of the 5G CU was also discussed, highlighting the centralized PDCP mechanism and the integration of a packet scheduler capable of distributing traffic dynamically across heterogeneous DUs. Finally, the chapter introduced two xApps hosted at the Near-RT RIC: one implementing a non-prioritizing scheduling policy and another applying prioritization through slice-aware WiFi allocation. These xApps rely on an LSTM-based prediction model and provide the CU with per-UE traffic steering decisions, enabling intelligent, adaptive multi-RAT scheduling under realistic QoS constraints.

Chapter 5

Experimental Architecture & Results

5.1 Introduction

This chapter introduces the software and hardware tools, system setup, and experimental infrastructure used to implement and evaluate the proposed 5G multi-RAT system. It also presents and analyzes the experimental results of the proposed 5G system, focusing on metrics to evaluate the effectiveness of the WiFi DU integration and the proposed traffic steering mechanism.

5.2 Software Tools

5.2.1 Docker & Docker-Compose

As mentioned earlier, the 5G CN is based on NFV, where each NF is deployed as a containerized service. Containers implement a virtualization method that provides isolated environments for NFs. Some of their advantages over other traditional virtualization methods, such as Virtual Machines, are that they are more lightweight, offer faster startup times and better resource efficiency, making them ideal for deploying microservices applications, such as the 5G CN. The most widely used container engine, and the one chosen for our implementation, is the Docker Container Engine. Some of their advantages over other traditional virtualization methods, such as Virtual Machines, are that they are more lightweight and offer faster startup times and better resource efficiency, making them ideal for deploying microservices applications, such as the 5GCN. Since 5GCN involves multiple NFs (such as

AMF, SMF, and UPF), in this implementation, the Docker Compose tool was used. Docker Compose is a tool that allows you to define and deploy multi-container Docker applications using a single YAML file. In the context of the 5G CN, Docker Compose enables the orchestration of all necessary NFs by specifying their images, environment variables, networking configuration and dependencies.

5.2.2 OpenAirInterface

To deploy all 5G components, we used OpenAirInterface (OAI), an open-source software-based implementation of the full 5G protocol stack aligned with 3GPP standards. It supports fully open-source SDR solution built on x86 architecture, supporting UE, gNB, and CN functionalities. It is one of the most wide-used 5G implementations that supports end-to-end 5G system testing, all the way from radio access through core network functionalities. Its usage is especially highlighted in experimental setups, algorithm prototyping, and validating custom network configurations in research environments.

OAI - 5G CN

OPENAIR-CN-5G is an open-source implementation of the 3GPP specifications for the 5GCN. It includes all major NFs required for full 5G core functionality: AMF, AUSF, LMF, NEF, NRF, NSSF, NWDAF, PCF, SMF, UDM, UDR, and UDSF. The UPF component is provided in two variants: a standard UPF with optional eBPF support, and a high-performance VPP-based version (UPF-VPP). This modular architecture enables flexible deployment, supports network slicing and analytics, and provides a complete SBA-compliant 5G core suitable for research and prototyping.

OAI - 5G RAN

OAI-5G RAN framework supports multiple gNB deployment architectures, ranging from monolithic implementations—where all RAN functions are co-located in a single entity—to fully disaggregated setups that follow the 3GPP-defined split into CU-CP, CU-UP, and DU components. This flexibility allows researchers to explore and deploy different levels of RAN functional separation. In our work, we adopted the F1-based architecture, where the gNB is split into a CU and one or more DUs, and communication between them is governed by the 3GPP-compliant F1 interface, consisting of the F1-C (control plane) and F1-U (user plane) protocols. This configuration enabled us to integrate and manage multi RAT DUs, including our proposed WiFi-based DU, under a common CU control framework.

OAI-RF Simulator

The OAI RF simulator is a software-based radio interface emulator tool that allow testing and validation of gNB and UE without needing physical RF hardware like USRPs. It simulate the RF link in digital domain between the gNB and the UE. The generated I/Q samples are exchanged through shared memory or UDP sockets. This makes it suitable for functional testing, debugging and protocol development in a more controlled and repeatable environment. In this thesis, the RF simulator is used to simulate the physical RF hardware between 5G DU and the connected UEs.

OAI-FlexRIC

OAI-FlexRIC is an open-source, lightweight RAN Intelligent Controller (RIC) platform designed to support the development and deployment of custom xApps for near-real-time control in Open RAN environments. It adheres to the O-RAN architecture and provides a modular and programmable framework for controlling various RAN components through the standardized E2 interface. FlexRIC includes an E2 termination, a flexible agent framework, and a runtime environment for executing xApps that can subscribe to RAN KPIs and apply control actions. Its architecture enables fine-grained control over radio resource management functions, such as scheduling, handovers, and slice enforcement, making it ideal for experimental research and prototyping of RIC functionalities. In our implementation, FlexRIC served as the near-RT RIC hosting the xApps responsible for multi-RAT traffic steering.

5.2.3 Open vSwitch

Open vSwitch (OVS) is an open-source virtual switch that allows programmable control of network traffic in virtualized and software-defined environments. In our system, OVS was used inside the WiFi DU to apply user-based traffic slicing. Using custom flow rules, OVS changes the ToS field of IP packets depending on the slice to which each user belongs. Packets with specific ToS value are then forwarded to the coresponding AC in the 802.11 MAC layer, enabling packet prioritization. This enables user-based slicing in the WiFi without requiring changes to the application logic or kernel, providing a lightweight solution for slice-aware traffic handling.

5.2.4 Ath10k

Ath10k is an open-source Linux driver for Atheros 802.11ac WiFi chips. Qualcomm Atheros first developed it and then released to the Linux community. It supports chipsets like QCA988x and QCA6174 and can run in both AP and STA mode. Ath10k works with the mac80211 layer in the Linux kernel and supports features like MU-MIMO, beamforming, and packet injection. It also allows for configuring channels and regulatory domains to comply with local wireless rules. One important thing is that it supports firmware offloading, so not everything runs on the CPU—things like encryption can be handled directly by the chip, which helps the system perform more efficiently. Ath10k is part of the Linux kernel and is still being updated. Due to these factors, it is widely used in various wireless devices, including routers, access points, and network cards. In our implementation, the ath10k driver was used in the WiFi DU to enable communication with the 802.11ac wireless interface acting as an AP and support slicing through the mac80211 subsystem.

5.2.5 Hostapd

Hostapd is a user-space daemon used to enable AP functionality on Linux-based systems by interfacing with wireless drivers and the mac80211 framework. It allows a system with a compatible wireless card to operate as a full-featured WiFi access point, supporting features such as WPA/WPA2 authentication, IEEE 802.1ac, MAC address filtering, and multiple SSIDs. In our implementation, hostapd was used in the WiFi DU to manage the AP mode of the wireless interface, allowing UEs to associate and communicate through the WiFi link.

5.3 Hardware Components

5.3.1 WiFi NIC - Atheros AR9380

As WiFi NIC we utilized the Atheros AR9380, which support the IEEE 802.11ac protocol suite, offering high-speed wireless communication over the 5 GHz band. In our setup, we used the AR9380 both as the AP and STA side, establishing the WiFi connection between the

WiFi DU and the UE. This card is compatible with the ath10k driver and performs well under 802.11a/n/ac modes, making it a good choice for experimental setups. It supports MIMO capabilities and high data rates, which helped to achieve reliable connection for our traffic steering and slicing tests. The AR9380 allowed us to focus on the software part of the implementation without worrying for the stability of the wireless link, which was very important for the evaluation of the multi-RAT system.

5.4 Experimental Infrastructure

Slices-RI

Not being limited to simulation results, the proposed system was tested under real-world conditions in a testbed. SLICES-RI is a European Research Infrastructure that aims to support large-scale experimentation in future network technologies like 5G, 6G, IoT and cloud-native systems. It provides access to remotely controllable and programmable testbeds located across various European countries. These testbeds are equipped with advanced hardware and tools that help researchers conduct experiments in both real and emulated environments. One of these national nodes is NITOS in Greece, which we used in this thesis to deploy and test the proposed system.

Nitos

Since 2007, the NITLAB team of the University of Thessaly has developed and maintained the NITOS testbed, an open-source experimental platform for networking research. As already mentioned, NITOS is the Greek node of the SLICES-RI infrastructure [**?**], hosted at the University of Thessaly, Greece. It offers a powerful and remotely accessible environment for testing next-generation wireless systems under real-world conditions. To evaluate our proposed multi-RAT system, we utilized the NITOS testbed, as it provides a controlled yet realistic setup for deploying and testing 5G and Wi-Fi networks. It consists of three parts: the Indoor RF Isolated Testbed, the Outdoor Testbed, and the Office Testbed. All nodes are high-end machines equipped with Intel Core i5/i7 CPUs, 8 GB RAM, SSDs, and UNIX-based systems. The nodes are also combined with commercial hardware and various types of UEs like USB dongles and Android phones. Some nodes include SDR platforms, such as 10 N210 USRPs, 12 B210 USRPs, 4 X310 USRPs, and 4 ExMIMO2 boards. In this thesis we utilized the Office Testbed to perform the experiments.

5.5 Experimental Architecture

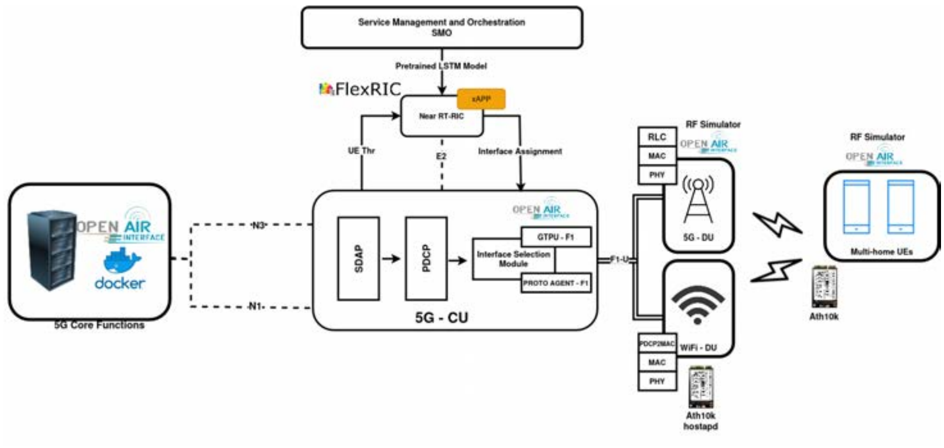

Figure 5.1: 5G Multi-RAT Experimental Architecture

Putting everything together, below is the list of all components used in the implementation along with the corresponding software and hardware tools as illustrated in **??**:

- **5GCN:** Docker, OAI
- **Near-RT RIC:** OAI FlexRIC
- **5G CU:** OAI
- **5G DU:** OAI (RF Simulator)
- **WiFi DU:** Hostapd + Ath10k driver + Open vSwitch (OVS) + Atheros AR9380
- **Multi-homed UE (5G side):** OAI UE (RF Simulator)
- **Multi-homed UE (WiFi side):** Ath10k driver + Atheros AR9380 (ath10k)

5.6 Dataset

The publicly available dataset in [20] contains 5G traces collected from a major Irish mobile operator and includes two mobility scenarios (car and static) and two application types (file download and video streaming). From all possible combinations, we selected the static mobility pattern with a downloading application because it has the highest bitrate demand with less variation, which makes it more suitable for our system evaluation. Another reason is that the static scenario contains more entries compared to the mobility one, which also helps in training the ML model better. Because our proposed method is based on forecasting the UE throughput, we kept only the DL bitrate column and removed the rest 25 features (such as Longitude, Latitude, PINGAVG, PINGMIN, PINGMAX, etc.) as they were either not helpful or redundant. For training the LSTM model, we used data from three out of the five available days, and the other two were used to simulate the traffic demand of two different UEs. During the preprocessing stage, we applied min-max normalization in the range of 0–100 Mbps, and used spline interpolation to smooth the data and reduce unnecessary noise, making it more suitable for model training.

5.7 LSTM

To estimate the upcoming throughput demands of the UEs, we replay a realistic dataset as described in Section 5.6, generating traffic over the 5G network components. The xApp collects relevant KPI values, which are used to predict near-future network demand. We treat this as a time-series forecasting task and utilize a Long Short-Term Memory (LSTM) neural network, a commonly used deep learning model for demand prediction in modern cellular networks.

The model comprises two LSTM layers with 224 and 64 neurons respectively, followed by a dense output layer with 2 units. The architecture and hyperparameters, including the number of neurons per layer, learning rate, and dropout configuration, were proposed by the Keras Tuner tool based on the training dataset and the target prediction horizon. The model takes the last 20 time steps as input and produces predictions for the next 2 steps. The LSTM model was pre-trained offline at the SMO layer, using 10 epochs and the Adam optimizer with a learning rate of 0.0034. It achieved a MAE of 1.384 on the test set, providing a good balance between accuracy and generalization.

As the focus of this thesis is not on the optimization of the ML model itself but on its application within a real-time RAN control framework, no comparative study with other algorithms was performed. The trained model is deployed as a forecasting module within the xApp, enabling real-time throughput prediction using the most recent 20 UE Throughput values as input.

Table 5.1: LSTM Model Parameters

Parameter	**Value**
Architecture	2 LSTM layers + Dense output layer
1st LSTM Layer	224 neurons
2nd LSTM Layer	64 neurons
Dense Output Layer	2 units
Input	Last 20 time steps
Output	Prediction for next 2 time steps
Optimizer	Adam
Learning Rate	0.0034
Epochs	10
Tuning	Keras Tuner (hyperparameter search)
Performance (MAE)	1.384 on test set

5.8 Experimental Results

Based on the experiments conducted on the NITOS testbed, we observed the following behaviors. In our experimental setup, as aforedescribed the Wi-Fi access point was implemented using the Atheros AR9380 chipset with the ath10k_pci. The NIC was configured for operation with a single spatial stream (1×1 MIMO) and a 40 MHz channel width in the 5 GHz band, enabling 802.11ac functionality. Under these conditions, the system operated at MCS 7 achieving a theoretical maximum PHY rate of 150 Mbps. Through extensive experiments using UDP traffic, the setup consistently delivered real-world throughput of approximately 120 Mbps, corresponding to about 80% of the theoretical PHY rate. This value aligns with expectations, considering the MAC-layer overhead, interframe spacing, CSMA/CA contention, and protocol headers inherent in Wi-Fi operation.

The maximum downlink throughput through the 5G DU using the RF simulator was approximately 60 Mbps, while for the WiFi DU, UDP traffic managed to succeed almost the same with 120 Mbps, closely matching the performance of the baremetal WiFi access point and demonstrating that the integration of the WiFi DU within the 5G RAN architecture does not impose significant overhead or degradation in achievable throughput.

Particularly interesting are the uplink measurements, where the WiFi DU outperformed the 5G DU, as the 5G DU was limited to around 10 Mbps, whereas the WiFi DU, due to its symmetric nature, achieved similar uplink transmission speeds as in the downlink. A noteworthy observation is the latency difference between the two DU technologies: the WiFi DU, due to its simpler protocol stack, achieved an average RTT of approximately 2.509 ms, while the 5G DU achieved around 11.752 ms.

5.8.1 xApp-Non-Prioritizing Multi-RAT Scheduling with Demand-Aware WiFi Allocation

As described in Section 4.6.1, the non Prioritizing Multi-RAT Scheduling with Demand-Aware WiFi Allocation scheme implements a scheduler that follows a hierarchical traffic decision process. At first, all the traffic is routed through the 5G DU, which is considered as the primary interface. If the estimated UE demand exceeds the capacity of the 5G DU, then the traffic is entirely forwarded to the WiFi DU. In case the demand is even higher than what WiFi DU alone can handle, the traffic is split and transmitted over both 5G and WiFi DUs simultaneously.

Figure 5.2 illustrates the throughput per UE under the Non-Prioritizing Multi-RAT Scheduling with Demand-Aware WiFi Allocation algorithm. The top subplot represents UE1, while the bottom one shows UE2. The color coding indicate the selected routing path for each session: blue for 5G DU, green for WiFi DU, and red when both DUs are used together.

The scheduler does not prioritize any specific UE, but instead distributes the traffic based on the predicted throughput demands and available capacity on each DU. As it shown, both UEs start by using the 5G DU, but when the demand increases, traffic is routed through the WiFi DU. In sessions where the predicted demand is higher than the capacity of any single DU, the traffic is splitted and transmitted over both DUs at once.

This behavior proves that the algorithm adapts to dynamic traffic patterns and performs interface allocation in a fair and responsive manner. No user is given priority, and the interface usage is guided mainly by demand.

Figure 5.3, illustrates the interface utilization for both UEs for each DU. At session $approx$50, the demands of the UE1 increase, resulting in the steering of the UE1's traffic

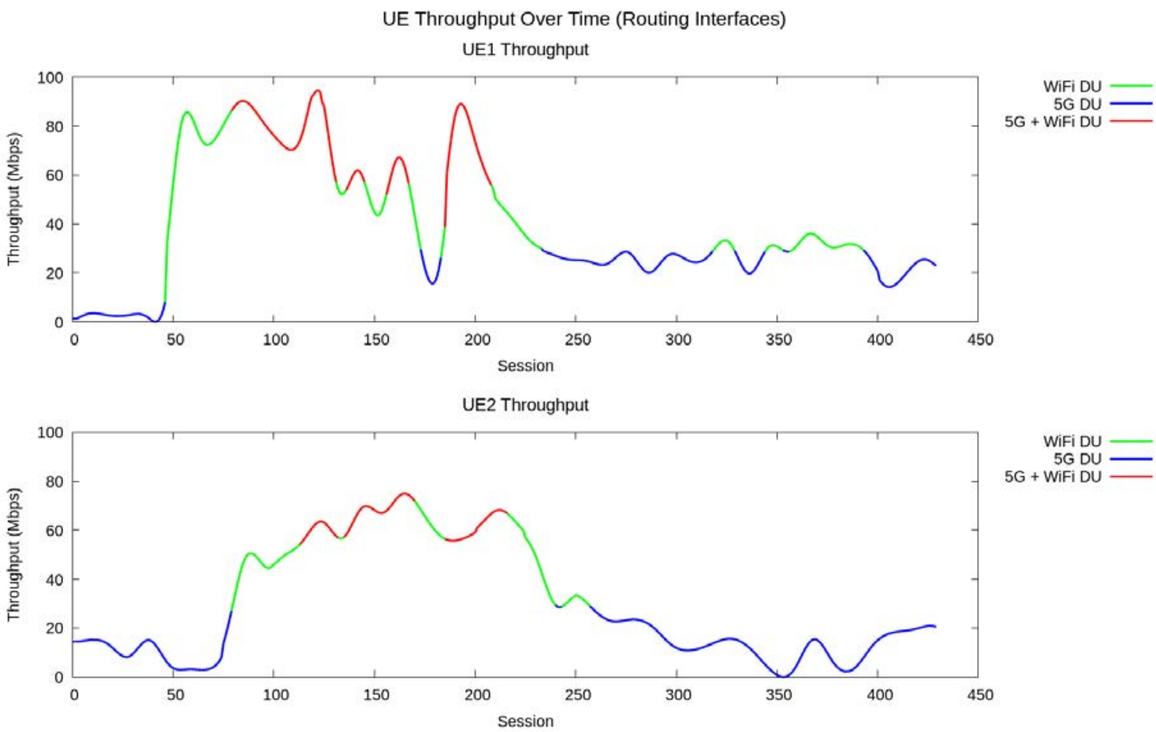

Figure 5.2: Per-UE throughput and interface selection over time

via the WiFi DU interface. Around session ≈75, a significant increase in UE2's traffic demand renders the 5G DU capacity inadequate to serve it fully. As a result, the xApp initiates traffic offloading for UE2 through the WiFi DU, leading to a noticeable increase in WiFi DU utilization. This creates a contention scenario, prompting UE1 to distribute its traffic load across both interfaces to mitigate potential WiFi congestion. Due to the non-prioritizing nature of the xApp, both UEs are treated equally, allowing them to utilize a comparable share of the available DU resources. Consequently, UE1 begins dynamically balancing its traffic between the 5G and WiFi DUs. From that point onward, UE1 maintains a dual-interface transmission strategy, driven by its consistently higher predicted demand relative to UE2. This behavior is clearly illustrated in the interface utilization plots, where UE1 exhibits more sustained usage across both DUs, while UE2, having lower demand, eventually returns to relying solely on a single interface. Following session ≈200, the total demand from both UEs declines. This trend is reflected in the reduction of throughput levels and the corresponding decrease in utilization across both interfaces.

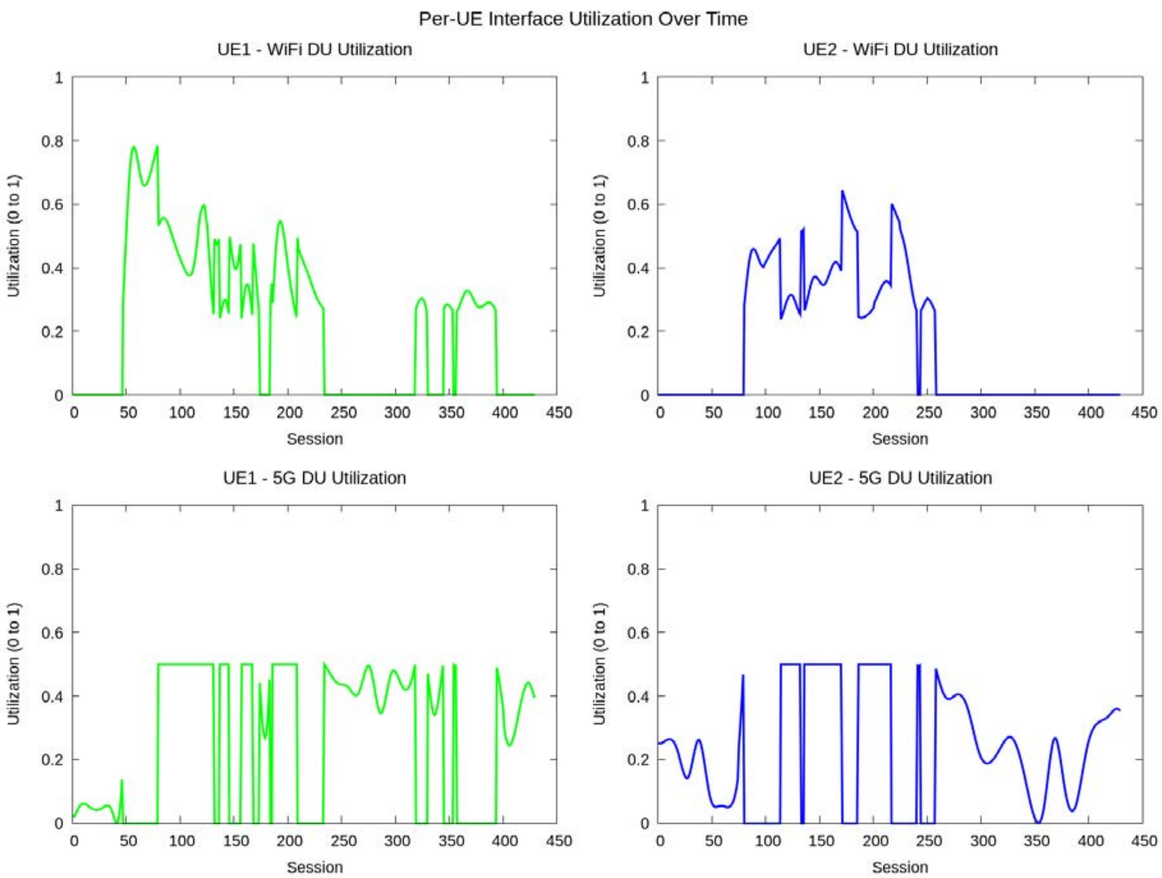

Figure 5.3: Per-UE throughput and interface selection over time

5.8.2 Prioritizing Multi-RAT Scheduling with Demand-Aware WiFi Allocation

As described in Section 4.6.2, the proposed algorithm follows a dynamic policy that differentiate between priority and non-priority UEs. Based on the experiments that was performed in our tested environment, WiFi appears to offer lower latency and better bandwidth, that's why it is selected as the priority interface. The idea behind the algorithm is simple — both priority and non-priority UEs are allowed to use the link with better quality, which is the WiFi DU in our case, while there is no contention. Once the priority UEs start to require more bandwidth, the non-priority ones are getting redirected also to the 5G DU, even though it is not as efficient in latency performance. In scenarios where the demand of both categories exceeds the total WiFi capacity, the algorithm applies prioritization, reserving 70% of the WiFi DU capacity for the priority UEs, since it offers lower latency and higher bandwidth. This allocation ensures that priority users maintain a higher quality of service, while non-priority users offload to the highest latency DU accordingly.

In our tested environment, we assume ideal conditions where no external interference affects the WiFi channels. This ensures that the contention-based nature of the WiFi DU is not impacted by surrounding devices or background traffic. Under such a controlled setup, the available bandwidth of the WiFi interface is fully dedicated to the connected UEs, enabling a precise evaluation of the traffic steering logic without unexpected variability. In real-world deployments, such interference could be mitigated by employing dynamic frequency selection mechanisms on the access point, allowing it to switch to less congested or free channels, thereby maintaining stable performance and minimizing contention.

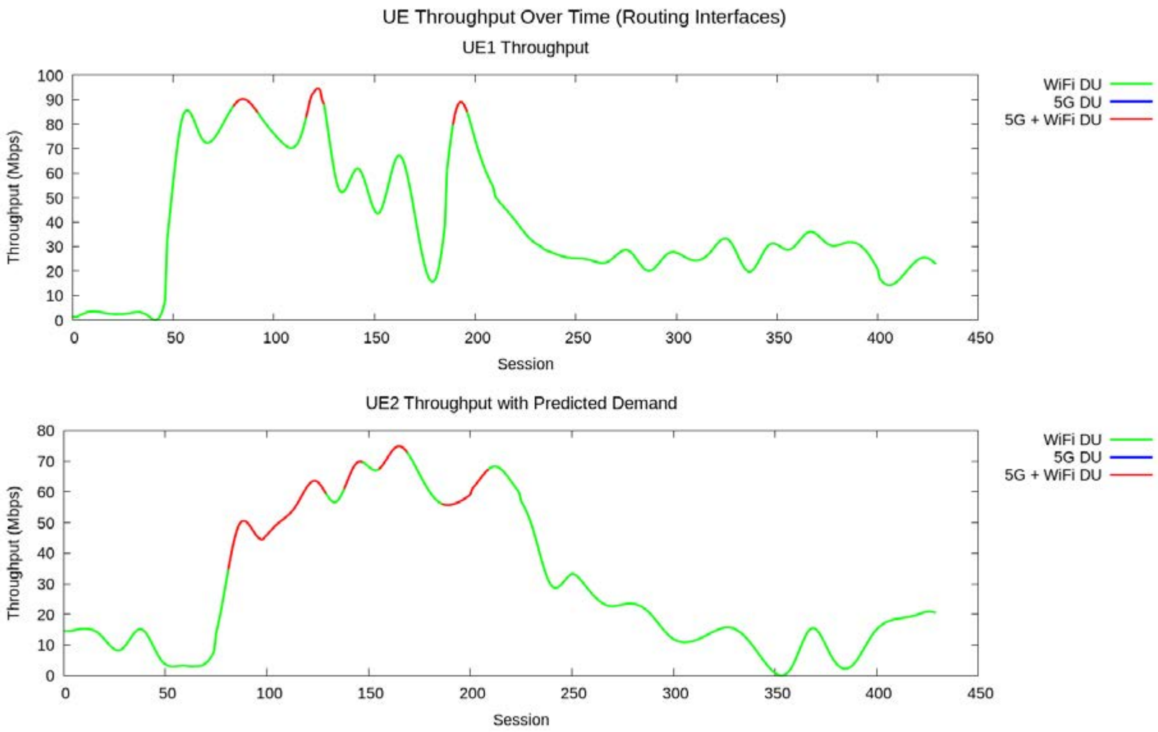

Figure 5.4: Per-UE throughput and interface selection over time

Figure 5.4 presents the received throughput for two UEs with coloring that represent the selected routing interface. The upper plot corresponds to UE1 which is the priority UE, while the bottom one shows the throughput for UE2 which is a non-priority UE. The green color indicates that the WiFi DU was selected, blue indicates the 5G DU, and red color means that both interfaces are used in parallel to serve the demand. As seen, UE1 is routed mainly through the WiFi DU, which is expected since the scheduler always prioritize this interface for the priority UEs. This choice is based on the fact that the WiFi DU provide lower latency and higher bandwidth, making it more suitable for serving priority traffic. In this setup, priority UEs are allocating 70% of the WiFi DU capacity, ensuring preferential access to this interface.

In contrast, UE2 uses the 5G DU interface alongside the WiFi DU for much longer periods than UE1, as it is a non-priority UE. This means that when the WiFi DU bandwidth portion allocated for non-priority UEs becomes insufficient to serve its demand, the scheduler leverages the 5G DU to compensate and attempt to fulfill the requested throughput. This behavior is clearly observed in the intervals where the red color appears in the plot for UE2, indicating periods where both interfaces are engaged concurrently.

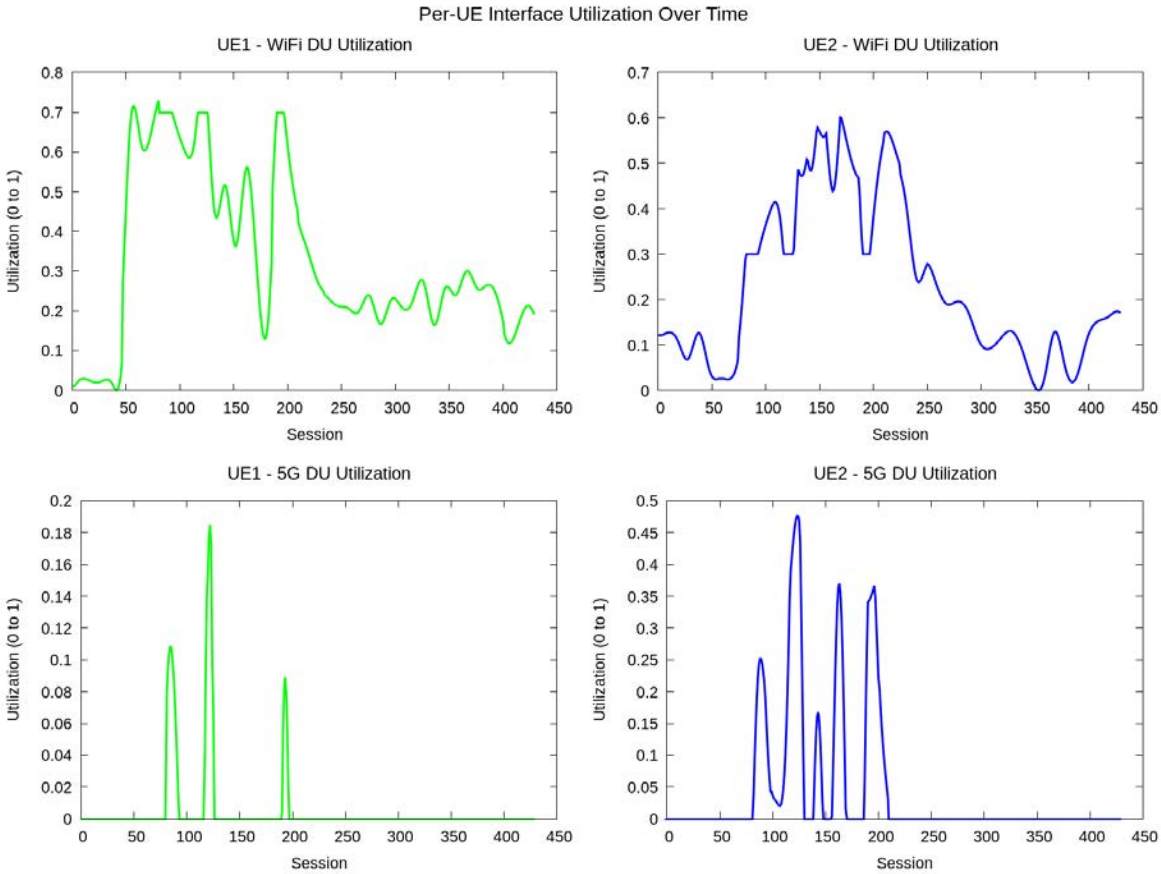

Figure 5.5: Per-UE throughput and interface selection over time

Figure 5.5 presents the per-UE interface utilization over time, illustrating how the scheduler distributes traffic across the available DUs for both UE1 and UE2. The upper plots show the WiFi DU utilization for each UE, while the lower plots reflect their corresponding 5G DU usage. It can be observed that UE1, as the priority UE, maintains consistently high utilization of the WiFi DU, often reaching 70% usage, especially during the early sessions. This confirms the intended behavior of the algorithm, which favors WiFi for serving the traffic of priority UEs. The 5G DU utilization for UE1 remains close to zero for most of the time, indicating that WiFi alone is sufficient to meet its demand, except for a few spikes where fallback to 5G was necessary since the requested demand of the priority UE exceed the WiFi DU capacity. On the other hand, UE2 exhibits more dynamic behavior: it initially shares the WiFi DU, but as UE1's (priority UE) demand increases, UE2 is progressively pushed towards the 5G DU, as seen from the fluctuations in its 5G utilization. This offloading pattern is a direct result of the 70/30 resource split mechanism applied when WiFi becomes over-utilized.

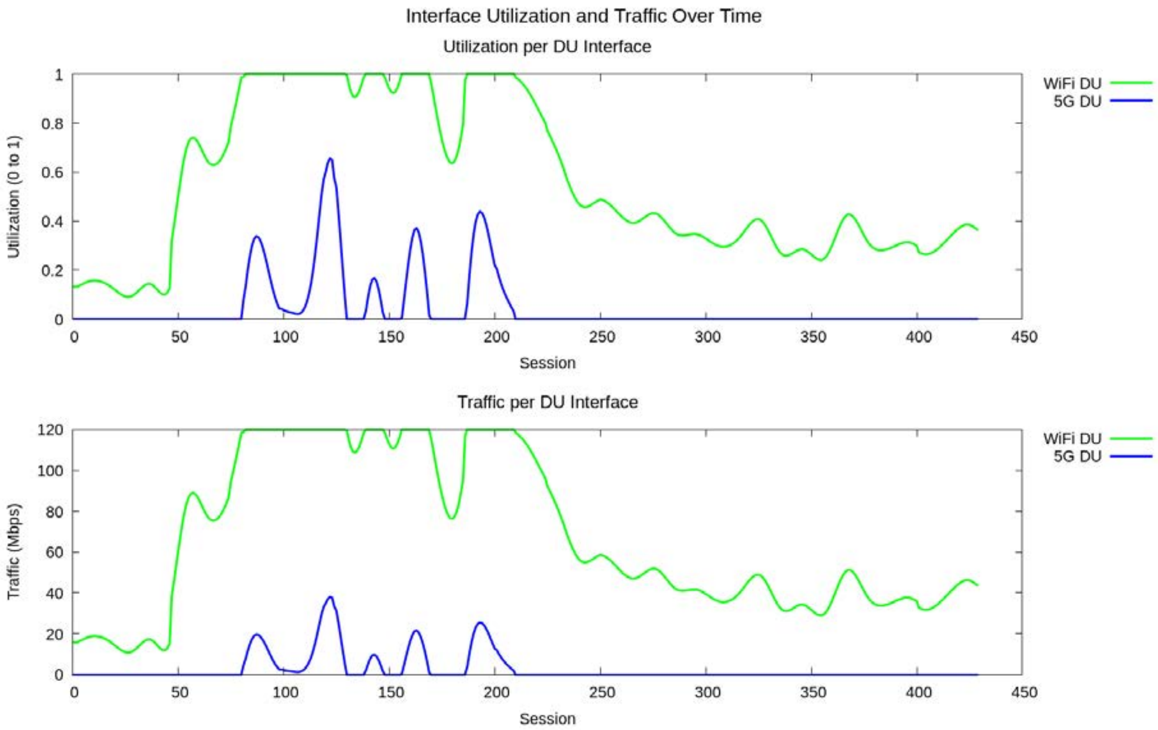

Figure 5.6: Per DU Utilization and transmitted Traffic

Figure 5.6 illustrates the overall interface utilization and traffic per DU over sessions, aggregating the behavior of both UEs. The upper subplot shows the utilization of the WiFi and 5G DUs as a fraction of their total capacity, while the lower subplot reports the actual traffic transmitted through each interface in Mbps. It can be seen that the WiFi DU reaches full utilization (close to 1.0) during the interval between session 60 and 200, which correspond to the period where both UEs are actively demanding higher throughput. This heavy usage of WiFi DU is expected, since it operates as the low latency and high bandwidth link in this setup, and the scheduler naturally prefers it as the primary interface. During the same interval, the 5G DU is also utilized in parallel, with its traffic and utilization increasing especially when WiFi becomes saturated. This reflects the behavior of the algorithm where non-priority UEs are offloaded to 5G once the priority users need more capacity on WiFi. After session 200, demand starts to drop, and the WiFi alone is enough to handle the load, so 5G DU goes back to zero.

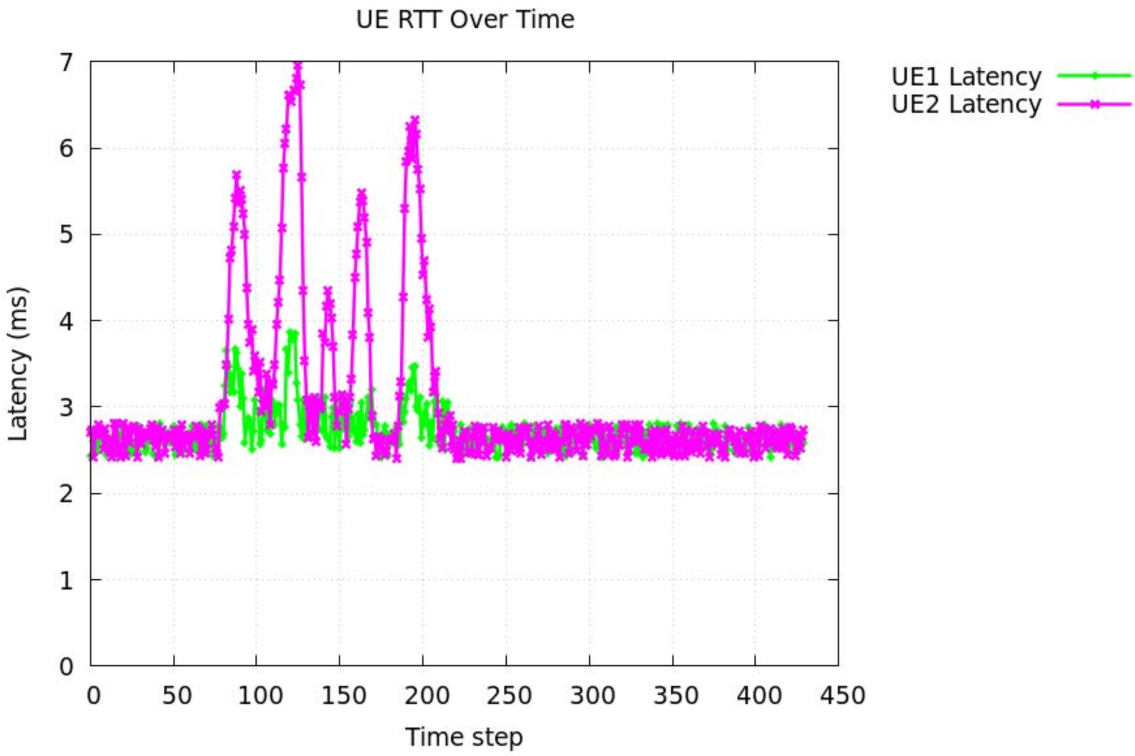

Figure 5.7: Per UE RTT

The RTT behavior observed in Figure 5.7 is closely linked to the per-UE interface utilization patterns shown in Figure 5.6. As the utilization plots illustrate, UE1 primarily uses the WiFi DU, with utilization levels reaching up to 0.7 during high-demand periods, while its use of the 5G DU is minimal — limited to just three distinct intervals, where utilization peaks around 0.18. This explains why UE1's RTT remains consistently low, typically between 2.5 ms and 3.5 ms, with only small increases during those brief 5G DU activations.

In contrast, UE2 relies much more heavily on the 5G DU interface. The UE2 5G DU utilization plot shows several periods, particularly between sessions 100 and 200, where utilization frequently exceeds 0.3 and even reaches 0.45. This greater reliance on the 5G DU, which inherently introduces higher round-trip delays compared to the WiFi DU, results in the notable RTT spikes seen for UE2 — with RTT values rising as high as 6–7 ms. These spikes align with intervals where the WiFi DU share available to non-priority UEs is insufficient, and the scheduler compensates by offloading significant portions of UE2's traffic to the 5G DU. Outside of these high-load periods, when UE2 can be served primarily by the WiFi DU, its RTT aligns closely with that of UE1, confirming that the scheduler reverts to low-RTT paths whenever possible.

5.9 Conclusion

Putting everything together, the experimental results presented in this section demonstrate the practical value of the proposed traffic steering strategy within a multi-RAT environment, where WiFi and 5G DU operate concurrently. The algorithm effectively distinguishes the priority and non-priority UEs, favoring the WiFi DU for priority users due to its lower latency and higher bandwidth properties. When the WiFi DU becomes congested, non-priority UEs are dynamically redirected to the 5G DU, allowing for more balanced load distribution and maintaining acceptable service levels. The results indicate that the proposed mechanism improves overall interface utilization while still respecting QoS needs across varying traffic conditions. This behavior provides a solid basis for more scalable multi-RAT orchestration in future wireless systems.

Chapter 6

Conclusion

6.1 Summary

In this thesis, we presented how the cellular networks have evolved over the years, exploring the changes that occurred in the architecture from the early days of 1G up to the modern 5G systems. We continued with a brief historical discussion on the WiFi evolution, and then we discussed the potential benefits of the convergence of these two technologies. Following that, we presented our Wi-Fi integration into the 5G System architecture by integrating at the MAC level of the RAN as a Wi-Fi DU. We then implemented two distinct traffic steering algorithms, each reflecting a different strategy for handling dual-RAT connectivity in the RAN. The first approach is a non-prioritizing scheduler that uses the 5G DU as the default interface, offloading traffic to WiFi only when 5G capacity is exceeded, and splitting it if necessary. The second algorithm adopts a priority-aware policy, assigning priority UEs to the WiFi DU due to its better latency and bandwidth, while non-priority UEs use the 5G DU unless excess WiFi capacity becomes available. Both mechanisms were deployed and evaluated on the NITOS testbed — the Greek node of the SLICES-RI infrastructure — allowing us to study their performance under controlled but realistic conditions. The results offer useful insights into how traffic can be dynamically routed in multi-RAT environments, highlighting the trade-offs between fairness, prioritization, and resource efficiency.

6.2 Future Work

While the proposed traffic steering algorithms proved effective in managing dual-RAT resources, several extensions could further enhance the flexibility and applicability of the approach. One direction for future work is the integration of additional RATs, such as LTE or mmWave, into the same scheduling framework. This would enable the development of more complex and adaptive multi-RAT orchestration strategies. Furthermore, we aim to extend our policies utilizing also network slicing not solely within the WiFi DU, but also across the 5G DU component. In the current architecture, the 5G interface relies on a static fair-sharing model for distributing resources among UEs, which may limit adaptability under heterogeneous traffic demands. Introducing dynamic slicing mechanisms to both DUs would enable more granular and policy-driven resource allocation, enhancing the system's ability to support differentiated service requirements and improving overall isolation and efficiency in multi-tenant environments.

Bibliography

[1] A.K. Salkintzis. Interworking techniques and architectures for WLAN/3G integration toward 4G mobile data networks. *IEEE Wireless Communications*, 11(3):50–61, 2004.

[2] Intel and Qualcomm. New WI proposal: LTE-WLAN Radio Level Integration and Interworking Enhancements, 2015.

[3] Alcatel-Lucent. Revised WID: LTE-WLAN Radio Level Integration and Interworking Enhancements, 2015.

[4] 3GPP. 3GPP TR 23.501 v15.3.0 (2018-09), 3rd Generation Partnership Project; Technical Specification; 5G; System Architecture for the 5G System , 2018.

[5] M. Luglio, M. Quadrini, C. Roseti, D. Verde, and F. Zampognaro. Performance evaluation of untrusted non-3GPP Access to a 5G Core Network via satellite. In *2022 International Symposium on Networks, Computers and Communications (ISNCC)*, pages 1–6, 2022.

[6] Mario Teixeira Lemes, Cristiano Bonato Both, Antonio Carlos De Oliveira Júnior, and Kleber Vieira Cardoso. A Tutorial on Trusted and Untrusted non-3GPP Accesses in 5G Systems - First Steps Towards a Unified Communications Infrastructure. *CoRR*, abs/2109.08976, 2021.

[7] 3rd Generation Partnership Project (3GPP). System architecture for the 5G System (5GS); Stage 2. Technical Report TS 23.501, 3GPP, 2020. Version 16.4.0.

[8] Yoohwa Kang and Changki Kim. A Multi-Access Session Management for ATSSS Support in 5G Network. In *2019 25th Asia-Pacific Conference on Communications (APCC)*, pages 409–412, 2019.

[9] Chengdong Yang, Tianyu Wang, Zhe Xiao, and Shaowei Wang. A Suggest-And-Improve RAN Slicing Strategy for Mobile Networks. In *2022 14th International Conference on Wireless Communications and Signal Processing (WCSP)*, pages 748–752, 2022.

[10] Matteo Nerini and David Palma. 5g network slicing for wi-fi networks. In *2021 IFIP/IEEE International Symposium on Integrated Network Management (IM)*, pages 633–637, 2021.

[11] Yiannis Yiakoumis, Kok-Kiong Yap, Sachin Katti, Guru Parulkar, and Nick McKeown. Slicing Home Networks. In *Proceedings of the 2nd ACM SIGCOMM Workshop on Home Networks*, HomeNets '11, page 1–6, New York, NY, USA, 2011. Association for Computing Machinery.

[12] Matías Richart, Javier Baliosian, Joan Serrat, Juan-Luis Gorricho, and Ramón Agüero. Slicing With Guaranteed Quality of Service in WiFi Networks. *IEEE Transactions on Network and Service Management*, 17(3):1822–1837, 2020.

[13] Sven Zehl, Anatolij Zubow, and Adam Wolisz. Hotspot slicer: Slicing virtualized home Wi-Fi networks for air-time guarantee and traffic isolation. In *2017 IEEE 18th International Symposium on A World of Wireless, Mobile and Multimedia Networks (WoWMoM)*, pages 1–3, 2017.

[14] Aymen Zreikat. Performance evaluation of 5g/wifi-6 coexistence. *International Journal of Circuits, Systems and Signal Processing*, 14:903–913, 2020.

[15] Mohamed Khadmaoui-Bichouna, Jose M Alcaraz-Calero, and Qi Wang. Empirical evaluation of 5g and wi-fi mesh interworking for integrated access and backhaul networking paradigm. *Computer Communications*, 209:429–443, 2023.

[16] Ilias Chatzistefanidis, Nikos Makris, Virgilios Passas, and Thanasis Korakis. ML-based Traffic Steering for Heterogeneous Ultra-dense beyond-5G Networks. In *2023 IEEE Wireless Communications and Networking Conference (WCNC)*, pages 1–6. IEEE, 2023.

[17] 3GPP. 3GPP TR 38.801 v14.0 (2017-03), 3rd Generation Partnership Project; Technical Report; Technical Specifications for Cloud Radio Access Network, 2017.

[18] Nikos Makris, Christos Zarafetas, Pavlos Basaras, Thanasis Korakis, Navid Nikaein, and Leandros Tassiulas. Cloud-Based Convergence of Heterogeneous RANs in 5G Disaggregated Architectures. In *2018 IEEE International Conference on Communications (ICC)*, pages 1–6, 2018.

[19] Dimitris Kefalas, Nikos Makris, Thanasis Korakis, and Serge Fdida. Design, integration and slicing of non-3gpp dus in 5g networks: a hybrid 5g-du approach. In *IEEE INFOCOM 2024 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 1–6, 2024.

[20] Darijo Raca, Dylan Leahy, Cormac J. Sreenan, and Jason J. Quinlan. Beyond throughput, the next generation: a 5G dataset with channel and context metrics. New York, NY, USA, 2020. Association for Computing Machinery.