ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DOTTORATO DI RICERCA IN

INGEGNERIA ELETTRONICA, TELECOMUNICAZIONI E TECNOLOGIE DELL'INFORMAZIONE

Ciclo 37

Settore Concorsuale: 09/F2 - TELECOMUNICAZIONI

Settore Scientifico Disciplinare: ING-INF/03 - TELECOMUNICAZIONI

OPEN RAN APPROACHES FOR 6G NON-TERRESTRIAL NETWORKS

Presentata da: Riccardo Campana

Coordinatore Dottorato

Davide Dardari

Supervisore

Alessandro Vanelli Coralli

Co-Supervisore

Carla Amatetti

Esame finale anno 2025

To my beloved family.

Contents

List of Figures

List of Tables

List of Acronyms

3GPP	Third Generation Partnership Project
A1AP	A1 Application Protocol
ADAM	Adaptive Moment Estimation
AFS	Adaptive Functional Split
AI	Artificial Intelligence
ANR	Automatic Neighbour Relationship
API	Application Programming Interface
ARIMA	Auto Regressive Integrated Moving Average
ARQ	Automatic Repeat Request
BCH	Broadcast Channel
BLER	Block Error Rate
BS	Base Station
CCC	Cell Configuration and Control
CDMA	Code Division Multiple Access
CL	Clutter Loss
CN	Core Network
CNF	Cloud-Native Network Function
CNN	Convolutional Neural Network
CoMP	Coordinated Multi-Point
CP	Control Plane
CPRI	Common Public Radio Interface
CSI	Channel State Information
CT	Core Terminals
CU	Centralized Unit
DL	Downlink
DL-SCH	Downlink Shared Channel
DQL	Deep Q-Learning
DRL	Deep Reinforcement Learning
DU	Distributed Unit
E2AP	E2 Application Protocol
E2S	E2 Service Model

eCPRI	enhanced Common Public Radio Interface
EI	Enrichment Information
eMBB	enhanced Mobile BroadBand
eMTC	enhanced Machine Type Communication
ESE	Elementary Signal Estimator
FCAPS	Fault, Configuration, Accounting, Performance, Security
FDD	Frequency Division Duplexing
FDMA	Frequency Division Multiple Access
FDS	Frequency Domain Spreading
FEC	Forward Error Correction
FFR	Full Frequency Reuse
FFT	Fast Fourier Transform
FH	Fronthaul
FML	Federated Meta-Learning
FoV	Field of View
FR1	Frequency Range 1
FR2	Frequency Range 2
FSL	Free Space Loss
GEO	Geostationary Orbit
GNN	Graph Neural Networks
GNSS	Global Navigation Satellite Systems
GOCA	Group Orthogonal Coded Access
GRU	Gated Recurrent Unit
GS	Ground Station
GSO	Geosynchronous Orbit
GW	Gateway
HAPS	High Altitude Platform Systems
HARQ	Hybrid Automatic Repeat reQuest
ICT	Information and Communication Technology
IETF	Internet Engineering Task Force
IMUX	input multiplexer
INLs	Inter-Node Links
IoT	Internet of Things
ISD	Inter-Site Distance
ISL	Inter Satellite Links
KPI	Key Performance Indicator
KPM	Key Performance Measurement
LBEF	Load Balancing Efficiency Factor
LBSH	Load Balancing Satellite Handover

LCRS	Low Code Rate Spreading
LDS	low-density spreading
LEO	Low Earth Orbit
LoS	Line of Sight
LS	Least Squares
LSTM	Long Short Term Memory
LTE	Long Term Evolution
MAC	Medium Access Control
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MDT	Minimisation of Drive Tests
MEO	Medium Earth Orbit
MIB	Master Information Block
MIMO	Multiple Input Multiple Output
ML	Machine Learning
MLOps	Machine Learning operations
MMSE	Minimum Mean Square Error
mMTC	massive Machine Type Communications
MnS	Management Services
ModCod	Modulation and Coding Scheme
MPA	Message Passing Algorithm
MSE	Mean Square Error
MUD	Multi-User Detection
MUSA	Multi-User Shared Access
NCMA	Non-Orthogonal Coded Multiple Access
Near-RT	Near Real time
NETCONF	Network configuration protocol
NFV	Network Function Virtualisation
NG	Next Generation
NGMA	Next Generation Multiple Access
NGSO	Non-Geosynchronous Orbit
NI	Network Interface
NIB	Network Information Based
NLOS	Non-Line-of-Sight
NN	Neural Network
NO	Network Operator
NOCA	Non-Orthogonal Coded Access
NOMA	Non Orthogonal Multiple Access
Non-RT	Non Real time
NR	New Radio
NTN	Non-Terrestrial Network

OFDMA	Orthogonal Frequency Division Multiple Access
OISL	Optical Inter Satellite Links
OMA	Orthogonal Multiple Access
OMUX	output multiplexer
O-RAN	Open Radio Access Network
PCI	Physical Cell ID
PD	Power domain
PDCP	Packet Data Convergence Protocol
PDMA	Pattern Division Multiple Access
PDU	Protocol Data Units
PHY	Physical Layer
PLFS	Physical Layer Frequency Signals
PLMN	Public Land Mobile Network
PLMNs	Public Land Mobile Networks
PNF	Physical Network Function
PRB	Physical Resource Block
PTP	Precision Time Protocol
QoS	Quality of Service
RAN	Radio Access Network
RAT	Radio Access Technology
RC	Ran Control
ReLU	Rectified Linear Unit
RF	Radio Frequency
RIC	RAN Intelligent Controller
RL	Reinforcement Learning
RLC	Radio Link Control
RMa	Rural Macro
RMSE	Root Mean Square Error
RRC	Radio Resource Control
RRM	Radio Resource Management
RRUR	Radio Resource Utilization Ratio
RSMA	Rate Splitting Multiple Access
RT	Real time
RTD	Round Trip Delay
RU	Radio Unit
SatCom	Satellite Communication
SCMA	Sparse Code Multiple Access
SCS	Sub-Carrier Spacing
SDAP	Service Data Application Layer
SDL	Shared Data Layer

SDMA	Space Division Multiple Access
SFCs	Service Function Chains
SI	Split Interface
SIB	System Information Block
SIC	Successive Interference Cancelation
SINR	Signal-to-Interference plus Noise Ratio
SLB	Swap-based Load Balancing
SLS	System-Level Simulation
SMO	Service Management and Orchestration
SNR	Signal-to-Noise Ratio
SoA	State-of-the-Art
SON	self-Organising Networks
SSB	Synchronisation Signal Block
SSP	Sub-Satellite Point
TDMA	Time Division Multiple Access
TN	Terrestrial Network
TNs	Terrestrial Networks
TSG	Technical Specification Groups
TTC	Tracking, Telemetry, and Command
UE	User Equipment
UMLB	Utility-based Mobility Load Balancing algorithm
UP	User Plane
uRLLC	ultra Reliable Low Latency Communication
VLEO	Very LEO
VNF	Virtual Network Function
vRAN	Virtualized-RAN
VSAT	Very Small Aperture Terminal
WI	Work Item

Abstract

The transition from the fifth to the sixth generation of mobile communication systems represents a significant leap in global connectivity, programmability, sensing, trustworthiness, and sustainability. At the heart of this transformation lies the native unification of Non-Terrestrial Network (NTN)s with Terrestrial Networks (TNs). In 6G NTN components are poised to deliver high throughput and low latency services, seamlessly integrating with terrestrial telecommunication networks. This integration will enhance system capacity, and offer flexibility and agility to address uncertainties and shifts in market demands, while also improving system availability. These advances aim to provide robust and scalable solutions that minimize both costs and energy usage. In addition, by providing reliable backhaul for next-generation networks, NTNs ensure seamless and ubiquitous service. This integration aligns with the European Union's sustainability goals, which emphasize reducing energy consumption and emissions while scaling connectivity to bridge the digital divide. The integration of TNs and NTNs into a unified 6G framework demands a paradigm shift in Satellite Communication (SatCom) system design. Traditional SatCom systems, built on static and closed architectures, cannot meet the dynamic and diverse demands of 6G. These systems lack the flexibility to adapt to real-time changes or support seamless interoperability with terrestrial counterparts. To overcome these limitations, future NTNs must adopt modular, virtualized, and intelligent designs that enable adaptability and efficient operation across varying environments. Advanced decision-making frameworks, powered by Artificial Intelligence (AI), can process telemetry data and dynamically optimize network performance. Open standards and shared design principles further enhance interoperability, fostering innovation and collaboration, while reducing reliance on proprietary solutions. Virtualized and software-driven technologies will decouple network functions from hardware, ensuring cost-effectiveness, energy efficiency, and the rapid deployment of services.

The Open Radio Access Network (O-RAN) framework plays a critical role in enabling this transformation. Introduces modularity, programmability, and interoperability into network design, bridging the gap between static legacy systems and the dynamic requirements of 6G. By disaggregating traditional base stations into components such as the central unit, the distributed unit and the radio unit, O-RAN supports flexible deployment and scaling. Intelligent control mechanisms, facilitated by RAN Intelligent Controller (RIC)s, analyze data and adjust network operations to optimize performance. Open interfaces ensure seamless communication

across heterogeneous network elements, while cloud-native virtualization enhances scalability and adaptability. Despite challenges such as adapting to satellite-specific constraints, O-RAN provides a solid foundation for integrating NTNs into the 6G ecosystem, ensuring sustainability, inclusion and innovation for future networks.

In this framework, the PhD research activities aimed to explore the applicability of the O-RAN paradigm and architecture to NTNs, evaluating its potential benefits in achieving flexible, interoperable, and scalable network solutions while addressing the unique challenges posed by the dynamic nature, resource constraints, and operational complexities of NTNs. Thus, the first part of this thesis is devoted to an introduction of the O-RAN standard and to an overview of the key aspects of SatCom system architectures, with a particular focus on NTN architectures as defined by the Third Generation Partnership Project (3GPP) standardization process. Then, the discussion addresses three fundamental aspects of the O-RAN integration in NTNs: i) the NTN architecture design and compatibility with O-RAN; ii) the Radio Access Network (RAN) virtualization and disaggregation implementation in NTNs; and iii) the exploitaion of near-real-time and real-time control loops to enhance NTN RAN performances.

In detail, the PhD research proposes three distinct architectural models to integrate O-RAN principles into NTNs addressed as: Full-gNB, Orbit-split, and Feeder-split architecture. These models address the unique challenges of satellite-based networks, including high mobility and diverse latency requirements.

The study follows with a thorough investigation of functional splits for RAN disaggregation, assessing their impact on NTN performance. Eight options, ranging from physical layer (PHY) splits to higher-layer protocol splits, are analyzed. The findings indicate that lower-layer splits, such as Options 7 and 6, pose significant challenges due to their high throughput demands and stringent latency requirements. In contrast, higher-layer splits offer greater flexibility in addressing NTN-specific impairments, although with reduced granularity in control. The challenges imposed by NTN-specific constraints on the split interface, such as interface capacity and latency, are examined, and adaptive solutions are proposed to mitigate these problems.

The research proceeds outlining the benefits of a virtualized 3GPP NTN system, emphasizing how software-driven RAN and Core Network (CN) solutions simplify upgrades to support new 3GPP features. Virtualization's flexibility allows operators to efficiently deploy and update networks as standards evolve, avoiding hardware-centric overhauls. The chapter details the NTN standardization path in 3GPP, focusing on architectures from Releases 17 and 18, ongoing efforts, and expected developments in Release 19. It also demonstrates how virtualization facilitates the adoption of new functionalities and analyzes the software updates required to integrate Release 18 NTN features into a Release 17 software-based gNB, User Equipment (UE), and CN.

Finally, the integration of non-real-time (non-RT), near-real-time (near-RT), and real-time (RT) control loops within NTN RIC is assessed. In the context of near-RT control loops, a novel AI-based framework for TN-NTN traffic offloading is developed, utilizing deep Q-learning to dynamically allocate resources between terrestrial and non-terrestrial layers. Numerical results demonstrate substantial improvements in system throughput with an increase of system spectral efficiency of 7% in the 70% of cases, achieving optimal load balancing under diverse network conditions. The research also exploits real time control loops to design a channel state information (CSI) prediction techniques to counteract the impact of channel aging caused by satellite mobility. A neural network-based CSI prediction model enhances decoding accuracy and improves the performance of Sparse Code Multiple Access (SCMA) in NTN environments. The numerical results demonstrate that the proposed solution can achieve high prediction accuracy compared with adopting the outdated CSI directly. Ultimately, the concept of control loop is enhanced with a proactive CNF orchestration framework for NTN environments, utilizing AI-driven time-series forecasting to predict and allocate critical resources efficiently. The approach minimizes delays and optimizes energy consumption, ensuring seamless scalability in dynamic NTN scenarios. Results demonstrate the superiority of multivariate Long Short Term Memory (LSTM) models, achieving a Mean Absolute Percentage Error (MAPE) of 0.28, significantly outperforming baseline methods like Auto Regressive Integrated Moving Average (ARIMA), and highlighting the benefits of incorporating multiple variables for improved prediction accuracy and resource management.

Chapter 1

Introduction

All sectors of industry and society already experience the benefits of 5G networks, and research and development interests are now directed towards enhancing 5G features through 5G-Advance (5G-A), [1]. The latter is expected to bring 5G to its full potential by solving its last points of attention and by providing its connectivity to all possible scenarios and devices. The enhancement of 5G features is paving the way for 6G, the next generation of mobile communications.

The evolution of 5G into beyond 5G (B5G) and 6G networks aims to address society's growing demand for continuous and ubiquitous connectivity services across all facets of life and industry [2]. This shift represents a significant leap beyond the goals of 5G, aspiring to provide much more than just fast mobile internet access. Indeed, 6G ambitions extend to redefining the interaction between the physical, human, and digital worlds through advanced technologies and innovative frameworks.

The vision for 6G encompasses several transformative objectives, [3]. It will support emerging paradigms such as digital twinning, immersive communication, cognition, and connected intelligence, enabling a seamless convergence of the physical, human, and digital realms. Central to its design is programmability, which will ensure the flexibility required to adapt to diverse and dynamic user needs. Moreover, 6G networks must deliver deterministic end-to-end services, guaranteeing predictable and reliable performance for critical applications. High-accuracy localization and high-resolution sensing services will be achieved through the integration of sensing and communication, while trustworthiness will underpin the infrastructure, making it a secure and dependable foundation for future societies.

Sustainability is another cornerstone of 6G. It aims to minimize the energy and resource footprint of mobile networks while promoting sustainability across other sectors of society and industry. Achieving global inclusivity is paramount; 6G must be scalable and affordable to ensure access for all, reducing resource waste and bridging the digital divide. Additionally, 6G will need to push beyond the performance benchmarks of 5G, significantly enhancing Key Performance Indicator (KPI)s to meet future demands.

1.1 Motivations

This section aims to address three key points that motivated the thesis: i) how to pursue a greener 6G with NTNs, ii) how SatComs need to be enhanced in future NTNs, and iii) why O-RAN is the choice.

1.1.1 NTN for a sustainable 6G Network

A critical pillar of the 6G framework is the integration of NTNs with TNs. NTNs excel in providing connectivity in areas that are challenging or costly to serve with terrestrial infrastructure, such as rural regions, aviation, and maritime environments. They also hold the potential to extend crucial KPIs, such as reducing communication latency over long distances. In the 6G ecosystem, NTNs will play an integral role, not only as part of the access network but also as a vital component of the backhaul for next-generation information and communication systems [4]. This capability makes NTNs indispensable for delivering anywhere, anytime service availability, continuity, and scalability over wide geographical areas.

Sustainable development remains a priority in Europe, with EU policies emphasizing the interconnection of economic, social, and environmental dimensions. Within the context of Information and Communication Technology (ICT), sustainability has two primary facets: its direct and enabling effects [3]. The direct effect involves minimizing the energy consumption, resource usage, and emissions associated with 6G, ensuring the technology aligns with broader sustainability goals. One of the greatest challenges for 6G will be mitigating the growth of energy consumption as traffic volumes rise, a task often referred to as flattening the energy curve.

The enabling effect of 6G highlights its transformative potential to drive innovation and efficiency across industries and communities by delivering unparalleled connectivity and intelligence. In this context, NTNs play a pivotal role. It is widely acknowledged that terrestrial networks alone cannot meet the stringent requirements for flexibility, scalability, reliability, and coverage necessary to enhance global energy efficiency. The integration of NTNs within the 6G architecture is therefore a key enabler of these objectives, advancing the twin goals of technological advancement and sustainable development.

1.1.2 SatCom Paradigm Change for TN-NTN Integration

The integration of terrestrial and non-terrestrial networks into a unified 6G architecture demands a transformative paradigm shift in SatCom system design and operation. Existing SatCom networks are predominantly built on closed, static architectures that tightly couple hardware and software. This rigidity results in inflexible systems that struggle to adapt to evolving technological demands and diverse application scenarios, presenting challenges such as limited scalability, constrained service differentiation, and difficulties in achieving seamless interoperability with terrestrial counterparts.

Unlike terrestrial networks, NTN systems introduce dynamic challenges inherent to their nature. Satellites operate in various orbits and interact with air-borne nodes, such as High Altitude Platform Systems (HAPS), creating a multi-layered architecture interconnected by Inter-Node Links (INLs). This third dimension of connectivity significantly increases the complexity of network management and interoperability, as it must address rapidly changing environmental conditions, mobility patterns, and diverse service demands over vast geographical areas. This complexity amplifies the need for a more adaptive and intelligent approach to managing and operating NTN systems within the 6G framework.

A dynamic and adaptable approach is critical to overcoming these challenges and enabling NTN systems to meet the requirements of next-generation networks. First, transitioning from monolithic to modular architectures will allow future NTNs to adopt flexible designs where network functions and components operate independently. This modularity facilitates scalability, upgradeability, and customized deployments that align with specific service needs and geographic constraints.

Second, the integration of advanced decision-making frameworks is essential for enabling NTNs to adapt dynamically to real-time conditions. By analyzing vast amounts of KPI data and proactively adjusting operations, these frameworks will ensure optimal performance in diverse and rapidly evolving environments. This intelligence is critical to addressing the inherent dynamism of NTNs while meeting the complex demands of a unified TN-NTN system.

Third, seamless interoperability between heterogeneous network elements will require a move toward shared, interoperable design principles. This approach involves adopting open and standardized communication frameworks, enabling NTNs to integrate efficiently with terrestrial networks. Enhanced interoperability will reduce dependency on proprietary solutions and foster a collaborative and innovative ecosystem, promoting broad technological advancements.

Lastly, the use of virtualized and software-driven technologies will decouple network functions from underlying hardware, introducing flexibility and efficiency into NTNs. This separation will enable dynamic resource allocation, adaptive traffic management, and the rapid deployment of new services, all while improving cost-effectiveness and energy efficiency. These attributes are essential to align NTNs with the broader sustainability objectives of 6G networks.

By addressing these interconnected aspects, NTNs can evolve into a flexible, interoperable, and resource-efficient cornerstone of the 6G ecosystem.

1.1.3 O-RAN

Building upon the transformative shifts required for SatCom and NTN integration, the O-RAN framework offers a practical solution to the challenges outlined in the previous sections. By addressing key requirements such as modularity, programmability, and interoperability, O-RAN provides a blueprint for enabling flexible, efficient, and intelligent systems that align with the ambitions of a unified TN-NTN 6G architecture.

- **Disaggregated Architecture**: At the heart of O-RAN is its disaggregated architecture, which splits traditional monolithic base stations into the Centralized Unit (CU), Distributed Unit (DU), and Radio Unit (RU). This modular approach supports independent scaling and deployment, allowing each component to address specific service and geographic requirements. For NTNs, such disaggregation is critical in accommodating the unique demands of multi-layered satellite networks and dynamic operating conditions.

- **Programmable and Intelligent Control**: O-RAN introduces intelligent control mechanisms through RICs, enabling both real-time and strategic network optimization. These controllers analyze KPI data, predict network behavior, and dynamically adjust operations to ensure efficient resource utilization and adaptability. This capability is particularly valuable for NTNs, where rapidly changing environmental and service conditions demand a high degree of operational flexibility and responsiveness.

- **Open Interfaces for Interoperability**: By standardizing communication interfaces such as E2, A1, and O1, O-RAN ensures seamless integration between heterogeneous network elements. This interoperability is fundamental for uniting terrestrial and non-terrestrial networks, enabling cohesive operation across diverse environments. Open interfaces reduce dependency on proprietary systems and encourage collaboration among vendors, fostering innovation across both TN and NTN ecosystems.

- **Cloud-Native Virtualization**: O-RAN's adoption of cloud-native principles brings virtualization into the forefront of network design, decoupling network functions from hardware. This flexibility allows dynamic resource allocation, efficient traffic management, and rapid deployment of new services. For NTNs, virtualization addresses critical needs for scalability, cost-effectiveness, and energy efficiency, enabling their alignment with the broader sustainability goals of 6G.

By incorporating these principles, O-RAN bridges the gap between the static architectures of existing SatCom systems and the flexible, interoperable infrastructure envisioned for next-generation NTNs. It provides a cohesive framework that supports the evolution of TN-NTN integration, ensuring that both terrestrial and non-terrestrial networks contribute effectively to the goals of a unified, sustainable 6G ecosystem. However, implementing the O-RAN framework in NTNs presents unique challenges, including adapting its principles to the dynamic nature of satellite systems and addressing the latency and resource constraints inherent to spaceborne networks. Overcoming these difficulties is essential to ensure the robust and efficient operation of O-RAN in NTN environments.

1.2 Thesis contributions and organization

The main objective of this thesis is to explore the applicability of the O-RAN paradigm and architecture to NTNs, evaluating its potential benefits in achieving flexible, interoperable, and scalable network solutions while addressing the unique challenges posed by the dynamic nature, resource constraints, and operational complexities of NTNs.

The research presented in this thesis is organized into three primary activities: i) the development of three O-RAN-compliant NTN architectures; ii) the evaluation of RAN virtualization and disaggregation within the proposed NTN architectures, focusing on the performance impact of selecting different functional split options and on the software implementation of the Release 18 NTN features; iii) the analysis of the impact of exploiting Real time (RT), Near Real time (Near-RT), and Non Real time (Non-RT) control loops for RAN optimization, achieved through the implementation and evaluation of Near-RT and RT Machine Learning (ML) optimization algorithms.

The work is organized as follows:

- Chapter 2 sets the overall framework of the thesis by first providing a comprehensive overview of the key aspects of SatCom system architectures, with a particular focus on NTN architectures as defined by the 3GPP standardization process and then by detailing the NTN-specific features introduced in 3GPP Releases 17 and 18, along with an account of ongoing work and anticipated developments for Release 19. For the sake of the analysis perfomed in the Thesis, this chapter also examines the primary impairments characteristic of NTN channels, offering insights into the challenges they pose.

- Chapter 3 provides an overview of the O-RAN framework focusing on the architectural concepts, the open interfaces, and the orchestration framework. Additionally, a literature review is presented, highlighting existing research on O-RAN-based non-terrestrial networks and identifying gaps.

- Chapter 4 explains how virtualizing 3GPP NTN systems enables rapid and flexible updates to RAN and CN software as new standards emerge. By separating network functions from hardware, operators can quickly adopt features introduced in 3GPP Releases 17 and 18 without major overhauls, and stay ahead of ongoing work for Release 19. It also describes how a virtualized architecture simplifies the transition to new functionality, outlining the software changes needed to integrate Release 18 NTN features into an existing Release 17 gNB solution.

- Chapter 5 investigates the principle of RAN disaggregation in the context of NTNs, focusing on its impact on system flexibility, scalability, and performance. It presents the proposed NTN system architectures, focusing on the integration of O-RAN architectures and interfaces. It provides a detailed analysis of functional split options, ranging from low-layer splits to high-layer splits, highlighting their trade-offs in terms of throughput, and latency. Additionally, it addresses the specific challenges posed by NTN environments, including interface capacity and latency constraints, and explores potential solutions. The chapter concludes with a forward-looking perspective on leveraging dynamic functional split optimization.
- Chapter 6 focuses on O-RAN-based RAN optimization to address the integration of NTNs into the 6G ecosystem, tackling challenges such as dynamic topologies, variable latencies, and resource constraints. Leveraging RICs and AI-driven optimization, it explores innovative solutions for load balancing, traffic steering, and resource optimization. Two key analyses are presented: first, a Deep Q-Learning-based framework for TN-NTN traffic offloading, optimizing system throughput by dynamically allocating resources between TN and NTN layers; second, a neural network-based model for channel prediction to address Channel State Information (CSI) aging caused by satellite mobility, enhancing decoding accuracy and improving the performance of Sparse Code Multiple Access (SCMA) in NTN environments. These approaches pave the way for intelligent and sustainable 6G network management. Ultimately, this chapter investigates the enhancement of control loop functionality through CNF orchestration, utilizing AI to facilitate proactive resource management and scalable, robust network operations designed for 6G.
- Finally, Chapter 7 concludes the thesis.

1.3 PhD Research Outcomes

This section describes the outcomes of the PhD research period in terms of paper publications and contribution to EU or ESA funded projects. The research activity on O-RAN-compliant NTN architectures has primarily resulted in the publication of the paper titled "O-RAN Based Non-Terrestrial Networks: Trends and Challenges", presented at the EuCNC Conference 2023, [5]. Subsequently, the work on NTN architectures was further developed and refined within the SNS-2022-STREAM-B-01-03 6G-NTN project, contributing to the design of its 3D multi-layered NTN architecture and Very Low Earth Orbit (VLEO) space segment, as detailed in deliverables 6G-NTN Deliverable 3.4 and D3.5, [6], [7]. Additionally, the analysis of O-RAN-based NTNs was leveraged within the HORIZON-JU-SNS-2022 5G-STARDUST project, contributing to deliverables on system requirements and specifications, [8], and AI-based radio resource management and onboard processing, [9].

The research activity focused on RAN disaggregation within the proposed NTN architectures culminated in the publication of the paper "RAN Functional Splits in NTN: Architectures and Challenges", [10], and supported the analysis of functional split options within the 6G-NTN project, as documented in deliverable D3.5, [7].

The research activity on O-RAN control loops for RAN optimization followed two distinct branches. The first branch contributed to the 6G-NTN deliverable 4.2, titled "Report on 6G-NTN Radio Controller", [11] with a focus on TN-NTN traffic offloading, and led to the publication of the paper "Emerging Advancements in 6G NTN Radio Access Technologies: An Overview", presented at the EuCNC/6G Summit 2024, [12]. The second branch concentrated on NTN channel prediction algorithms within the context of the ESA-funded EAGER project “tEchnologies And techniques for satcom beyond 5G nEtwoRks”. This work contributed to the success of the project and resulted in the publication of the paper "Cell-Free Multiple Input Multiple Output (MIMO) in 6G NTN with AI-predicted CSI", presented at SPAWC 2024, [13]. Furthermore, the ML-based algorithms developed during this activity were presented in the tutorial "AI in Non-Terrestrial Networks" at WiSEE 2023.

List of Publications

[P1] "Emerging Advancements in 6G NTN Radio Access Technologies: An Overview", H. Shahid, C. Amatetti, R. Campana, et al, *EuCNC/6G Summit 2024*

[P2] "Cell-Free MIMO in 6G NTN with AI-predicted CSI", B. De Filippo, R. Campana, A. Guidotti, C. Amatetti and A. Vanelli-Coralli, *SPAWC 2024*

[P3] "RAN Functional Splits in NTN: Architectures and Challenges", R. Campana, C. Amatetti, A. Vanelli-Coralli, *arXiv 2023*

[P4] "O-RAN Based Non-Terrestrial Networks: Trends and Challenges", R. Campana, C.Amatetti, A.Vanelli-Coralli, *EuCNC 2023*

[P5] "Towards the Future Generation of Railway Localization Exploiting RTK and GNSS", D. Mikhaylov, C.Amatetti, T.Polonelli, E. Masina, R.Campana, K. Berszin, C.Moatti, D.Amato, A. Vanelli-Coralli, M.Magno, L.Benini, *IEEE Transactions on Instrumentation & Measurement* 2023

[P6] "Neural Network based Non-Orthogonal Random Access for 6G NTN-IoT", C.Amatetti, R.Campana, A.Georganaki, A.Vanelli-Coralli, *IEEE GLOBECOM 2022*

List of Project Deliverables

SNS-2022-STREAM-B-01-03 **6G-NTN** “6G Non-Terrestrial Networks for the full integration of NTN component into 6G”:

[D3.3] "Report on software defined payload and its scalability", *2024*

[D3.4] "Report on VLEO space segment (1st version)", *2024*

[D3.5] "Report on 3D multi layered NTN architecture (2nd version)", *2024*

[D4.2] "Report on 6G-NTN radio controller (1st version)", *2024*

HORIZON-JU-SNS-2022 **5G-STARDUST** “Satellite and Terrestrial Access for Distributed, Ubiquitous, and Smart Telecommunications”

[D3.1] "System Requirements Analysis and Specifications", *2023*

[D4.4] "Preliminary report on AI-based radio resource management and onboard processing", *2024*

Tutorial

“AI in Non-Terrestrial Networks”, R.Campana, B. De Filippo, *WiSEE 2023*

Chapter 2

Non Terrestrial Networks

SatCom represent the result of collaborative advancements in communications and space technology research, aimed at extending capabilities and coverage while minimizing costs [14]. This chapter provides an overview of the key aspects of SatCom systems, discusses the typical impairments associated with satellite channels, reviews the NTN standardization process with a focus on Release 17, 18, and the anticipated developments in Release 19, and examines the State-of-the-Art (SoA) of O-RAN-based SatCom systems.

2.1 SatCom System Architecture

A SatCom system, shown in Fig. 2.1, is organized in the following way: i) a *space segment*, containing one or different active and spare satellites organized into a constellation; ii) a *control segment*, including all ground facilities for the control and monitoring of the satellites, known as Tracking, Telemetry, and Command (TTC) stations, and for the management of the traffic and the associated resources on board the satellite; and iii) a *ground segment*, including all the traffic Earth stations , [15].

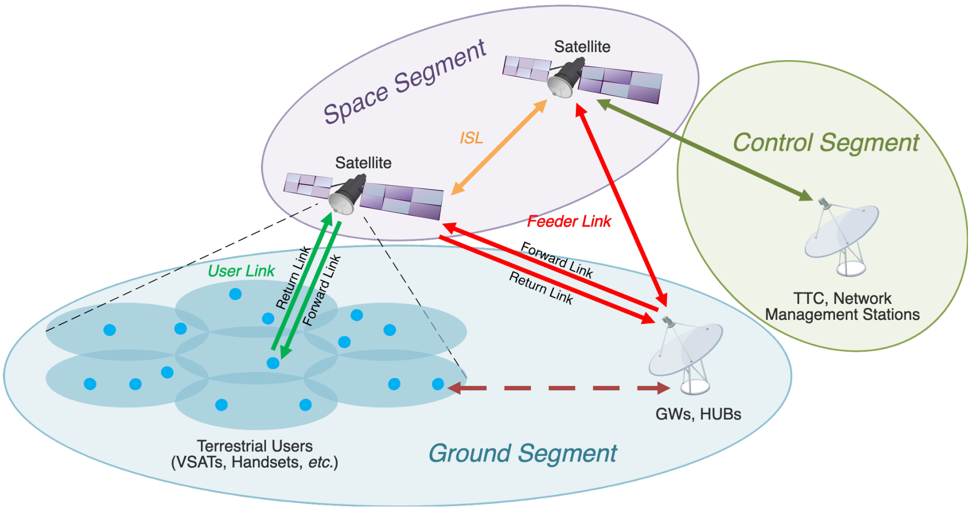


FIGURE 2.1: SatCom System high-level architecture.

Satellite networks are distinguished by their topology, the types of links they support, and the connectivity they offer among earth stations. In terms of configurations, a SatCom system could be organized: i) in a meshed network, where every node[1] is capable of communicating with every other node. A meshed satellite network is made up of a set of earth stations that are able to communicate with one another using satellite links consisting of radio-frequency carriers. ii) In a star network, where each node is able to communicate only with a single central node, named hub. iii) In a multi-star topology, various hubs (*i.e.*, central nodes) are identified. The other nodes can communicate only with those central nodes.

2.1.1 Orbits

In this Section the different types of satellite orbits and constellations are presented. 3GPP has focused on Geostationary Orbit (GEO) and LEO constellations for the deployment of NTN, [16].

2.1.1.1 Geostationary Orbit

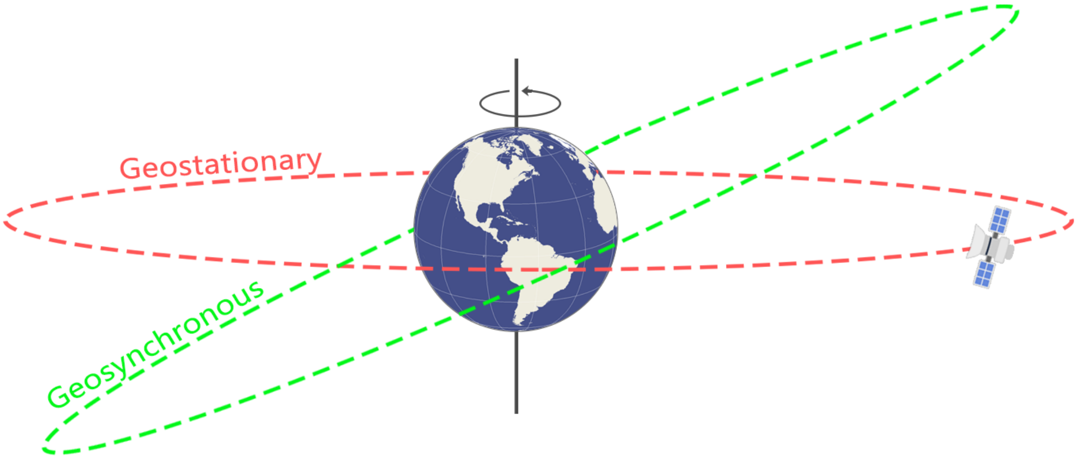


FIGURE 2.2: Geosynchronous and geostationary orbits.

A Geosynchronous Orbit (GSO), represented in Figure 2.2, is characterized by an orbital period of 24 hours, meaning that the satellite movement along its orbit is synchronous with the Earth's rotation. When a geosynchronous satellite is placed on the equatorial plane, at altitude of 35.787 Km, the orbit is said geostationary. In this case, an observer from earth sees the geostationary satellite in a fixed position in the sky. This equatorial geosynchronous orbit is very crowded due to its unique characteristics. Thus, in order to avoid interference between adjacent GEO satellites, each satellite is assigned a well delimited orbital slot, which corresponds to a distinct longitudinal position.

[1] A node is either a redistribution point or a communication endpoint. A satellite, as well as a Ground Station (GS) or a ground terminal, can act as network nodes.

2.1.1.2 Low Earth Orbit

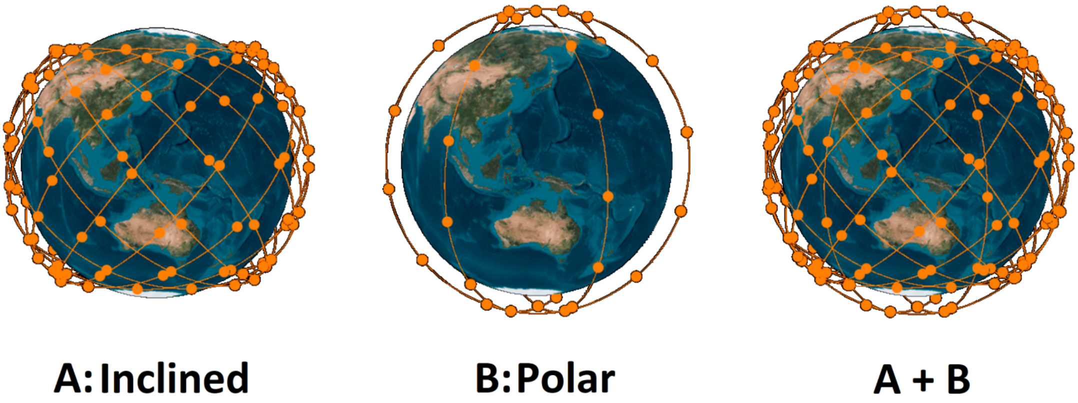
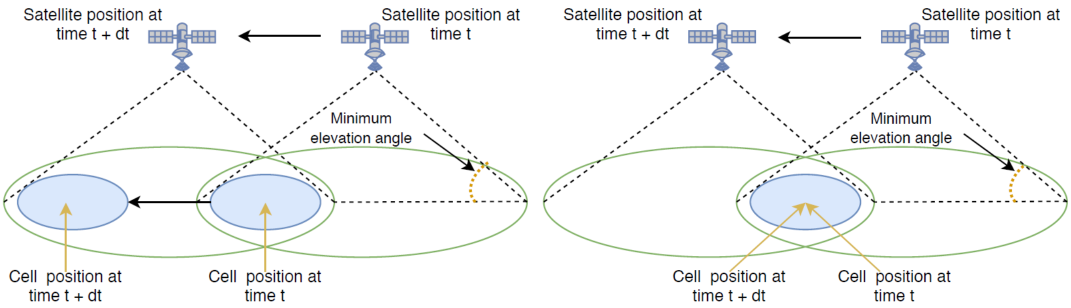


FIGURE 2.3: LEO at different inclinations.

LEO satellites altitudes range between 500 and 2000 km, involving orbital periods between 94 and 127 minutes and fall under the category of Non-Geosynchronous Orbit (NGSO) satellites. The orbit inclination is the angle of the orbit in relation to Earth's equator. A satellite that orbits directly above the equator has zero inclination. If a satellite orbits from the geographic north pole to the south pole, its inclination is 90 degrees. LEO satellites can be deployed at any inclination angle as seen in Figure 2.3. However, the deployment of a satellite at inclination value of either 28.5 degrees or 57 degrees can be easier, due to typical launch locations, [17].

LEO orbits are usually exploited for earth observation and communication satellites. For what concerns the satellite footprint, two types of deployment have been considered by 3GPP: Earth-moving beams and Earth-fixed beams. In the former, as depicted in Figure 2.4a, the satellite beams follow the satellite movement along the orbit, implying a coverage of a fixed point on Earth for just a few seconds. In the latter, as depicted in Figure 2.4b, the satellite realizes spotbeams that remain fixed on Earth[2].


Satellite position at time t + dt
Satellite position at time t
Minimum elevation angle
Cell position at time t + dt
Cell position at time t
Satellite position at time t + dt
Satellite position at time t
Minimum elevation angle
Cell position at time t + dt
Cell position at time t

(A) Earth-moving cells.

(B) Earth-fixed cells.

FIGURE 2.4: LEO satellite footprint typology, [18].

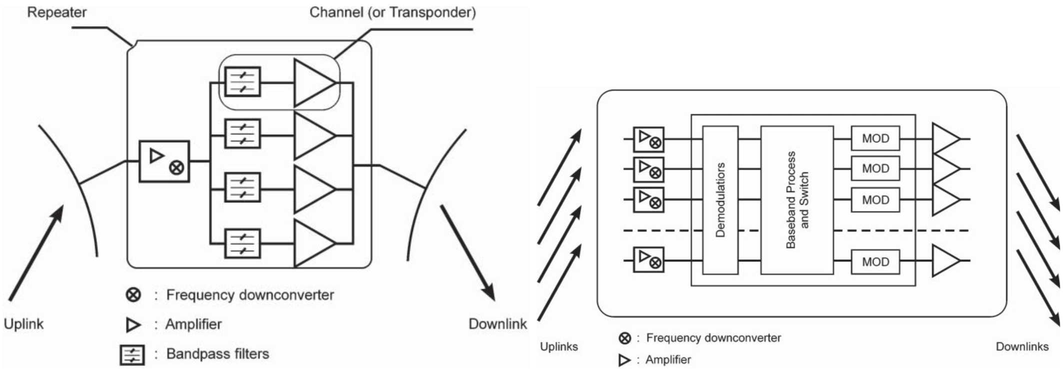
[2] The satellite can steer the spotbeams by tilting its antenna or by the means of attitude modifications.

2.1.1.3 GEO and LEO Scenarios Comparison

Making use of GEO it is possible to obtain a full earth coverage, polar region excluded, deploying only three satellites and gateways. By contrast, LEO presents a totally different scenario. In this case, in order to provide real-time global Earth coverage, the number of satellites and gateways required is much larger, and depends on the chosen orbit and the surface covered by the satellite antenna. In order to decrease the deployment and operational costs of the system one can reduce the number of LEO satellites selecting an inclined orbit. This ensures good coverage of the Earth after a given number of spacecraft revolutions; thus, it is a viable solution only when the message delivery latency is not an issue. Moreover, the number of terrestrial gateways can be reduced to one using a regenerative payload, hence deploying all eNB functionalities on the satellite.

LEO is characterized by high Doppler shift with compared to GEO. However, LEO constellations provide reduced round trip time compared to GEO. Moreover, LEO is characterized by a time variant geometry between the satellite and the UE that increases the probability of instantaneous Line of Sight (LoS) condition compared to GEO, allowing information delivery. On the other hand, LEO can cause temporary link unavailability causing an increase in the data delivery time.

2.1.2 Space Segment



(A) Transparent

(B) Regenerative

FIGURE 2.5: Types of payload [19].

As mentioned above, in the space segment one or several active and spare satellites orbit the Earth. Each of them is composed of the *payload* and the *platform*. The former is equipped with receiving and transmitting antennas and all the electronic equipment supporting the signal transmissions. Fig. 2.5 shows the two types of payload: *transparent* and *regenerative*.

A transparent payload, shown in Fig. 2.5a, (also known as bent pipe) only acts as a mirror between the elements of the ground segment, since it only amplifies the carrier power and down converts the central frequency. The power gain is of the order of 100-130 dB to increase the power level of the received carrier from a few tens of picowatts to the power level of the carrier fed to the transmitting antenna. To improve the separation between the receiving and the transmitting signals, frequency conversion is necessary. Moreover, since the payload is power limited, the overall bandwidth in the satellite is split into different sub-bands, and the carriers in each sub-band are amplified by a dedicated power amplifier. The amplifying chain associated with each sub-band is called a satellite channel, or *transponder* [19]. The sub-bands are obtained by means of a series of filters called the input multiplexer (IMUX). The amplified carriers in the output are then recombined in the output multiplexer (OMUX). The transparent payload, shown in Fig. 2.5a, could be carried either on i) a single-beam satellite, where each antenna generates one beam only; or ii) in a multi-beam satellite, where multiple-beam antennas generate multiple beams.

Referring to Fig. 2.5b, a multiple-beam regenerative payload is depicted. Differently from the bent-type payload, the uplink carriers are demodulated and re-routed to the output. Having the baseband signals allows *on-board processing* and routing of information from upbeam to downbeam through *on-board switching at baseband*. The frequency conversion is performed by modulating carriers, generated on board, at downlink frequency, which are then amplified and delivered to the destination in downlink.

Each beam identifies a beam coverage area, also known as *footprint*, on the Earth's surface. In the case of multi-beam, the aggregate beam coverage areas define the coverage area. If a satellite has several multiple-beam antennas, its coverage defines the satellite access area, which is directly related to the Field of View (FoV) of the satellite. Moreover, there are two types of beams: "Earth-fixed" and "Earth-moving". "Earth-fixed" beams, also named steerable beams, are used to cover stationary areas on the surface of the Earth that have fixed boundaries. These beams can be directed to different locations on the Earth. In contrast, "Earth-moving beams" are used in systems where the beam generated by the payload moves constantly as the spacecraft moves along its orbit.

In terms of coverage, it is possible to distinguish between *instantaneous* and *long-term* system coverage (shown in Fig. 2.6). The first one includes the aggregated coverage areas at a given time of the individual satellites participating in the constellation. The latter defines the area on the Earth scanned over time by the satellites in the constellation. When considering real-time services, the instantaneous system coverage should at any time have a footprint covering a geographical region where at least an element belonging to the GS is present. For store-and-forward services, this condition is not necessary, however, long-term coverage of the service zone is necessary. In order to provide continuous global coverage, a large number of satellites is required for LEO and VLEO satellite constellations. With GEO systems, three satellites are enough to achieve global coverage (except for the polar areas).

The platform consists of all the subsystems that allow the payload to operate, such as: i) the fuel system which is responsible for making the satellite run for

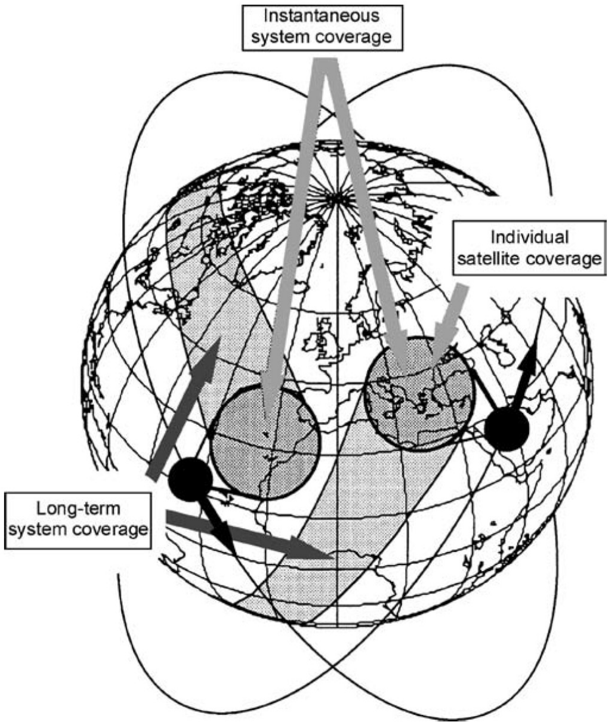

Instantaneous system coverage
Individual satellite coverage
Long-term system coverage

FIGURE 2.6: Types of coverage [19].

years; ii) the solar panels that provide the required energy for the satellite's operation; iii) the satellite and telemetry control system used to transmit commands to the satellite as well as to send the status of the onboard systems to the ground stations.

2.1.2.1 Non-Terrestrial Nodes

The next generation NTN topology will incorporate two types of non-terrestrial nodes: deterministic nodes, with a stable and predictable orbit (GSO and NGSO), and flexible nodes such as HAPs or specialized heavy drones. These flexible nodes may appear at different times and locations to boost coverage or network capacity but are deployed only as needed and are not intended for constant global presence.

Satellites positioned at varying orbits are referred to as deterministic nodes. The envisioned 3D 6G-NTN network comprises multiple layers, specifically:

- An upper GSO layer consists of minimum three satellites in geostationary orbits (GEO). Three such satellites can almost cover the entire globe, except for polar areas where they are not visible (i.e., near or below the horizon). The actual coverage depends on the minimum elevation, the lowest angle at which the satellite is visible above a user's local horizon on Earth, as illustrated in Figure 2.7. Inclined GSO orbits are omitted as they eliminate the benefit of GEO: no need for tracking antennas—while maintaining tight link budgets and high delays. The GSO is mainly anticipated to complement NGSO, concentrating on:

  - Broadcast and multicast operations aimed at fixed ground stations located, for instance, at the edge of coverage, though this is not the 6G-NTN project's primary goal.
  - Broadband access offering lower data rates and higher delay compared to NGSO, thus suited as backup or complementary capacity for hotspots (with dual steer/connectivity between GSO and NGSO links considered).
  - Offloading non-delay-sensitive traffic from the NGSO network facilitated by inter-satellite links connecting NGSO and GSO layers.
  - Delivering essential control and management functionalities to the NGSO fleet when feeder links or the ground segment are unavailable, enabling resilient and autonomous network operation (albeit with reduced capabilities) in the event of major disruptions to ground infrastructure.
  - Ensure resilience and link recovery in instances, such as failure, of lower constellations.

- A foundational layer comprised of NGSO satellites, where LEO includes Earth-centered circular orbits with altitudes up to 2,000 km, allowing them to orbit the Earth more rapidly than the planet's rotation. NGSO satellites are primarily intended to provide broadband access to handheld devices and VSAT-like UEs.

- MEO satellites, which usually orbit at approximately 10,000 km altitude, have been excluded to reduce the variety of architectural options for analysis. A MEO layer might be considered as a substitute for the GSO layer.

Flexible nodes essentially consist of HAPs or specialized heavy drones that can be temporarily utilized to enhance capacity in particular regions. Notable instances include disaster-stricken areas lacking terrestrial infrastructure or regions anticipating a brief surge in capacity demand, such as major concerts or sporting events, whether located in urban centers or remote areas. It is important to note that these nodes are not intended to form a permanent network; instead, they are expected be strategically deployed as required.

2.1.3 Ground Segment

Earth stations are categorized into the following classes: i) *user stations*, *e.g.*, handsets, mobile stations, and Very Small Aperture Terminal (VSAT), which allow the user to directly access the space segment; ii) *interface stations*, also named Gateways (GW), which link the space segment to a terrestrial network; and iii) *service stations*, like hub or feeder stations, which are in charge of collecting or distributing information from and to user stations through the space segment.

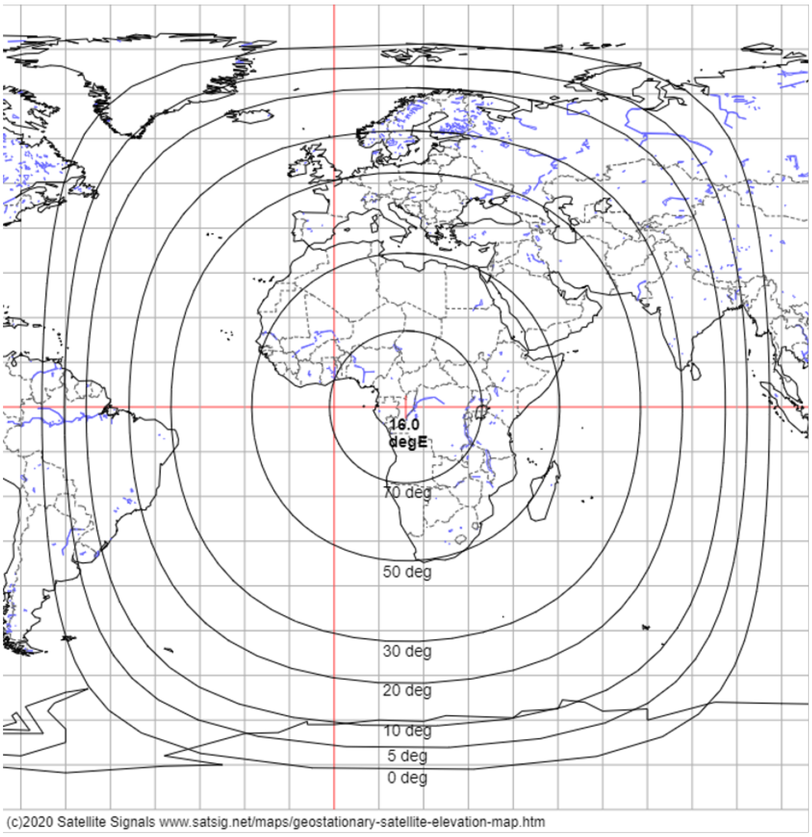

16.0
degE
70 deg
50 deg
30 deg
20 deg
10 deg
5 deg
0 deg
(c)2020 Satellite Signals www.satsig.net/maps/geostationary-satellite-elevation-map.htm

FIGURE 2.7: Exemplary coverage of a GEO satellite for different minimum elevation angles, [7].

The ground segment includes all the earth stations connected to the end user's terminal by a terrestrial network or, in the case of small station VSAT, directly connected to the end user's terminal. According to the volume of traffic on the satellite link and the type of traffic, Stations have different sizes.

2.1.4 Communication Links

The Non-terrestrial nodes communicate by means of the following links:

- **Feeder Links (FLs)** serve to connect deterministic or flexible nodes with a Ground Station (GS) or Gateway (GW) on Earth. GSs, equipped with large antennas, face fewer link budget constraints than UEs, resulting in FLs typically achieving a high availability of around 99.5% due to advanced fading countermeasures like power control, Adaptive Coding and Modulation (ACM), predictive handovers, etc. However, data rates can fluctuate during deep fading events, such as rain. FLs can function as both DL – Space to Earth, and UL – Earth to Space.

- **Inter-Node Link (INL)** connect non-terrestrial nodes. When both nodes involved are satellites, this link is called Inter-Satellite Link (ISL). If the link employs optical communication technologies, it is known as an Optical Inter-Satellite Link (OSIL); otherwise, it assumes the use of conventional RF technologies.

- **Service Link (SL)** connect deterministic or flexible nodes with a ground-based UE or one mounted on a drone, plane, or HAP. Like FLs, SLs can be both DL – Space to Earth and UL – Earth to Space.

2.1.4.1 Inter-Node Links

Five distinct types of INLs can potentially be considered, [7], namely:

1. Establishing links between HAPs and LEO satellites are considered to utilize optical technology. HAPs are designed as independent, adaptable network nodes not always within view of a ground station, so they must connect to the LEO constellation.

2. HAPs are considered to connect to GEOs using optical technology. Given the vast distance, their feasibility and value with respect to the data rate warrant further evaluation.

3. Connections between LEOs and GEOs are considered to occur in the Ka-band. Determining the benefit of all LEO satellites including ISL technology for GEO connections would requires additional analysis. Optical technology may replace Radio Frequency (RF) solutions if needed.

4. LEO satellite interconnections are considered to be established via optical technology.

5. GEO satellite links are considered to also use optical technology. Due to the vast distance (approximately 90,000 km with three evenly spaced GEOs), assessing their feasibility and significance concerning data rate would require further investigation.

To summarize: LEO-GEO in Ka-band, LEO-LEO using optical technology, and HAP-LEO with optical technology remain as baselines. Further analysis is needed for both GEO-GEO and HAP-GEO with optical technology.

2.1.4.2 Feeder links

While Q/V-band is generally preferred due to its available bandwidth and minimal spectral congestion, [7], Ka-band is still widely used and E band is considered especially for future use. Currently, these links are predominantly employed for GEO missions, where their directional beams simplify interference management. HAPs are considered to have feeder links only when a ground station is within sight. As a result, HAPs are expectet to only connect opportunistically to ground stations and will not have dedicated ones. Instead, they may use any LEO ground station within visibility, or alternatively, route traffic through the LEO network.

TABLE 2.1: NTN reference scenarios per system type [22]

System	Transparent	Regenerative
GEO	A	B
LEO steerable beams	C1	D1
LEO fixed beams	C2	D2

2.2 3GPP NTN system architectures for 5G

When considering the NTN integration into 5G systems, the RAN needs to communicate with the CN through the space segment, which includes the space-borne, *i.e.*, satellite-based communication platforms, and the air-borne, *i.e.*, HAPS. Thus, this heterogeneous system (*i.e.*, New Radio (NR)-NTN) envisages the mapping of these two entities to offer services, [15].

The 5G RAN, *i.e.*, the Next Generation Radio Access Network (NG-RAN), is described by its own interfaces and architectures, which are reported in TR 38.801 [20]. In detail, the 5G NR conceives the functional split of the gNodeB (gNB) in CU and DU. The goal of the functional split is to improve scalability, adaptability to various use cases and vertical services, and performance in network management. It is also the baseline approach for Network Function Virtualization (NFV) and Software Defined Networking (SDN). As stated in TS 38.401: i) a gNB can be split into a CU and one or more DU; ii) a DU communicates with only one CU; iii) the DU and CU are connected through the F1 air interface, which is a logical air interface. It is implemented by means of any existing standard, as long as specific signaling operations are ensured [21]. In this case, the CU hosts the Radio Resource Control (RRC) and Packet Data Convergence Protocol (PDCP) layers and the DU hosts the Radio Link Control (RLC), Medium Access Control (MAC), Physical Layer (PHY), and RF layers.

In 5G-NTN, the 3GPP identified six different macro scenarios (shown in Table 2.1) for the NTN providing access to the UE based on the type of orbit, payload, and its capabilities. In all of them, the link between the UE and the satellite (*i.e.*, the user service link) works either in S-band (*e.g.*, 2 GHz) or in Ka-band (*e.g.*, 20 GHz in the downlink and 30 GHz in the uplink). Moreover, LEO satellites can be distinguished based on whether they provide fixed or moving beams for coverage on the ground. It is worth emphasizing that only scenarios D1 and D2 include ISL, as shown in Table 2.1.

Therefore, based on i) the location of the gNB, thus, based on the type of payload, either transparent or regenerative, and ii) the type of user access link, either direct or relay-based, it is possible to distinguish several architectural options, described in the following sub-sections.

2.2.1 Direct access

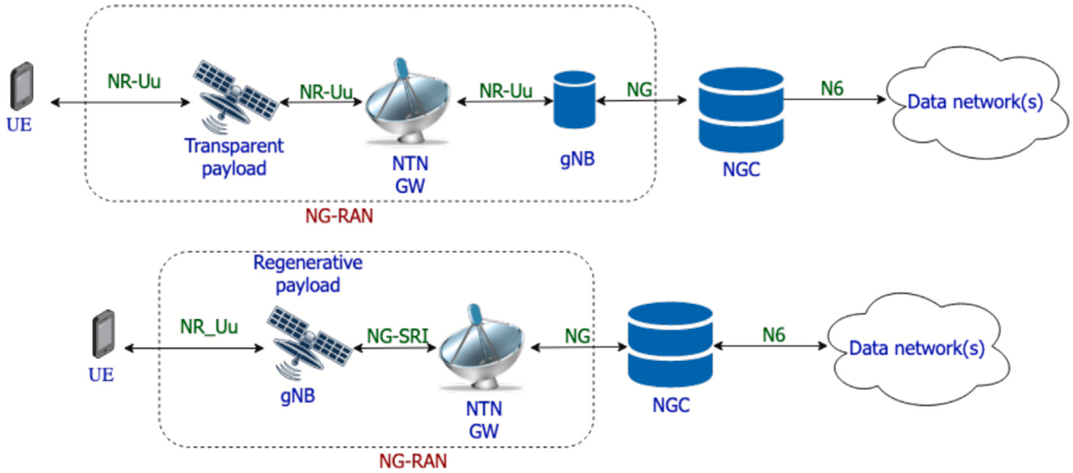


FIGURE 2.8: Architectural options with direct user access with transparent (top) and regenerative (bottom) payload [23].

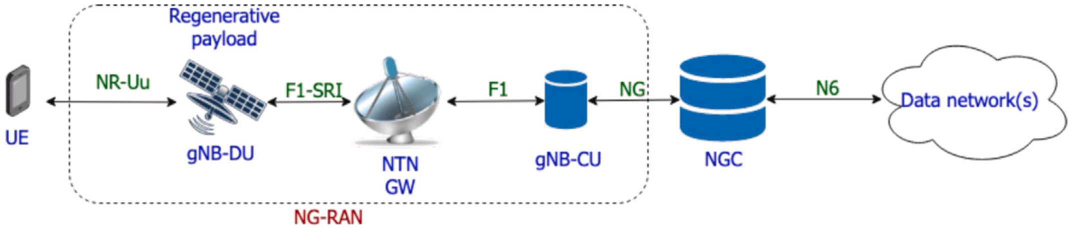


FIGURE 2.9: Architectural option with direct user access, regenerative payload, and functional split. Applicable to B, D1, and D2. [23].

In Fig. 2.8 the architectures with direct access are shown. In particular, the top diagram depicts a transparent payload and the bottom one a regenerative payload. Clearly, in this architecture, the users are directly connected to the satellite, [24].
In the transparent architecture, the platform only relays the NR signal between the gateway and the user terminal. Here, the gNB is conceptually located at the gateway, which means that all NR procedures are completed on the ground. The NR-Uu air interface is implemented on both links. Since it is a terrestrial interface, the typical impairments of the satellite channel, such as the large Round Trip Delay and the Doppler shift, must be taken into account on both the feeder and user links. In terms of the number of connections, each gNB is capable of managing a few tens of beams. Since in multi-beam NTN systems, each satellite might be able of generating even hundreds of beams, depending on the mission requirements, multiple gNBs might be needed in order to manage the NTN node. In case the satellites generate a reduced number of beams, then a single gNB can serve multiple payloads.
The regenerative architecture foresees the gNB or part of it on the flying platform. Even if it is a more expensive and complex solution, it allows for greater flexibility in adapting to the NTN channel.

When the full gNB is implemented on board, the NR-Uu protocols are entirely terminated on-board, *i.e.*, this Air Interface is only present on the user service link. The GW basically acts as a Transport Network layer node, terminating all transport protocols and connecting to the 5GC and the onboard gNB via the NG interface. This Air Interface is logical, *i.e.*, it can be implemented by means of any Satellite Radio Interface (SRI), as, for instance, the DVB-S2, DVB-S2X, or DVB-RCS2, [25].
As previously mentioned, the regenerative payload also allows for the possibility of implementing a functional split between the DU on the satellite and the CU on the ground (shown in Fig. 2.9). Currently, the adopted solution for the split is option 7.2, in which the user access link uses the Uu air interface and the DU and CU are connected by the F1 air interface on the feeder link. Usually, one CU is responsible for managing multiple DUs. However, it is worth noting that a single satellite can generate hundreds of beams, so when multiple satellites in a swarm serve the same area, multiple CUs are necessary to manage the swarm This architecture poses a challenge related to the F1 interface (*i.e.*, the interface between the CU and the DU) since it requires a persistent connection between the gNB-DU and the gNB-CU and it cannot be closed and re-activated on-demand; as such, with moving satellites as in a non-stationary scenario, all of the connections towards the served UEs would be dropped once the satellite is outside of the visibility of the current gNB-CU. Thus, smart implementations of the F1 interface and/or the functional split in NTN shall thus be designed.

2.2.2 Integrated Access and Backhaul

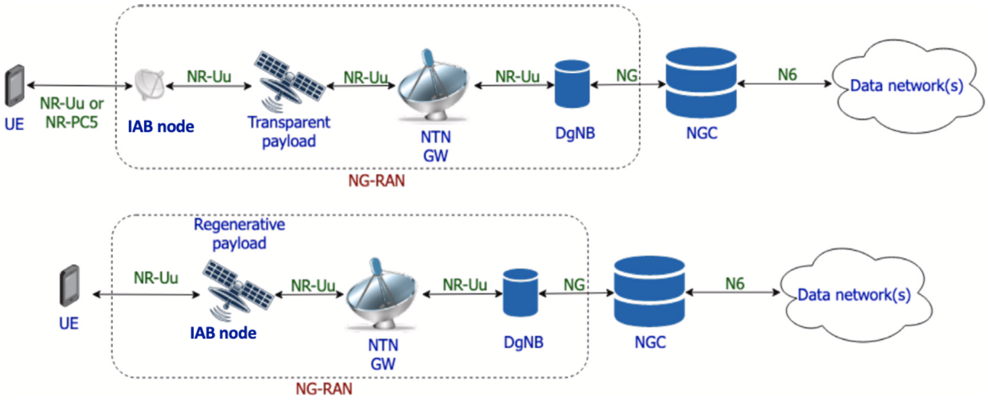


FIGURE 2.10: Architectural options with IAB with transparent (top) and regenerative (bottom) payload. [23].

In a relay-based access solution, the UEs do not directly connect to the gNB, but to an Integrated Access and Backhaul (IAB) node. The IAB allows NR access technologies to be used not only on the link between the gNB and terminals but also on wireless backhaul links. The architecture, in Fig. 2.10, includes one IAB donor node and multiple IAB nodes. The donor node provides connectivity to the CN, to which it is connected by means of conventional methods, such as fiber or microwaves, and serves the IAB nodes and the terminal directly linked to it. Each IAB node consists of a Mobile Termination (MT) part and a DU part. The former connects the IAB node to a parent DU, which could be the IAB donor or the DU part of another node, while the DU part communicates with UEs or the MT part of child IAB nodes. It is meaningfulness to highlight that: i) the MT part acts as a UE when communicating with the parent DU. ii) From the UE perspective, the DU part works as a normal gNB. In fact, it is not able to differentiate between normal gNBs and IAB nodes, because the IAB node's architecture is transparent to the user.

When considering an NTN element, the IAB can be implemented using either a transparent or regenerative payload. The architecture with relay nodes and transparent payloads (top diagram in Fig. 2.10) is more complex compared to direct access scenarios due to the introduction of a potentially large number of on-ground IAB nodes acting as gNBs. Because the IAB node can terminate protocols up to layer 3, no modification is foreseen on the user service link. In this case, the impact of typical satellite channel impairments only needs to be assessed on the forward link. When an onboard IAB node is implemented, *i.e.*, the regenerative solution is used, and the payload cost increases but there is the advantage of terminating protocols up to layer 3 on the satellite, as in the regenerative direct access architecture. For the feeder link, the connection between the IAB node and the donor gNB must be implemented using an NR-Uu air interface [26]. However, these architectures are currently under investigation for NTNs [22].

2.3 NTN impairments

This section reports some considerations about the delay, Doppler, and path losses elaborated in "NTN support for mMTC: Architectural and channel model considerations" [27].

Notably, compared to a terrestrial channel, satellite links pose challenging issues related to large delays, Doppler shifts, and path losses.

In the following, the system shown in Fig. 2.11 is considered as the baseline, where the following assumptions hold: i) a beam center located on the uv-plane at (u_{BC}, v_{BC}); ii) a beam radius $r_{uv} = \sin \vartheta_{3dB}$, where ϑ_{3dB} is the 3 dB half beam width, defined based on the NTN configuration sets in TR 38.821 [22]; and iii) a satellite orbiting the spherical Earth at an altitude h_{sat}. The conversion between (u, v) coordinates and satellite look angles is given by:

$$\begin{cases} u = \sin \vartheta \cos \varphi \\ v = \sin \vartheta \sin \varphi \end{cases} \tag{2.1}$$

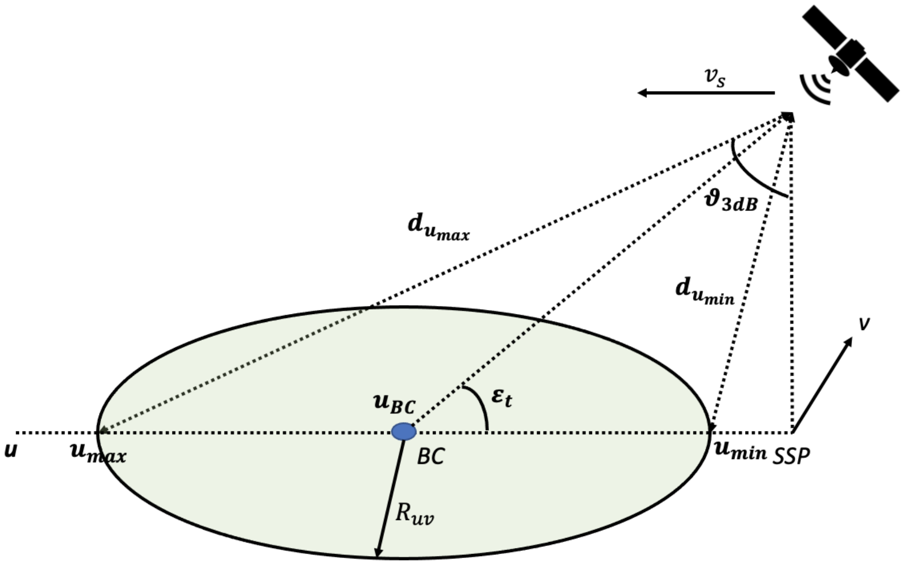


FIGURE 2.11: Reference system for delay and Doppler in *u*-*v* coordinate system.

$$\begin{cases} \vartheta = \arcsin\left(\sqrt{u^2+v^2}\right) \\ \varphi = \arctan\left(\frac{v}{u}\right) \end{cases} \tag{2.2}$$

For the sake of clarity, only a single beam is considered, but this assumption does not impact the generality of the proposed mathematical framework. In fact, when assessing the impact of the channel impairments on the NB-Internet of Things (IoT) procedure, it shall be noticed that such procedures involve users from the same beam.

2.3.1 Delay

To characterize the satellite channel in terms of latency, it shall be noticed that: i) compared to terrestrial communications, the propagation delay is predominant; and ii) the feasibility of most of the NB-IoT procedures is impacted by the Round Trip Delay (RTD) and the differential delays (the maximum, in particular) of users belonging to the same beam. In the following, the processing delay at the network elements is considered negligible and the propagation delay on the user link is considered the only contribution to the overall latency. Thus, the goal of this paragraph is to compute the maximum differential delay, *i.e.*, the maximum differential slant range, between two generic users. By definition, the differential delay is the difference in delay between a generic user and the user experiencing the minimum delay in the same beam.

Indeed, the one-way propagation (T_{ow}) delay of two or more terminals in the same beam can be split into two parts:

$$T_{ow} = \Delta\tau + T_{ul} \tag{2.3}$$

where T_{ul} describes the delay component common to all the UEs in the same beam (please, note that the feeder link and the Inter Satellite Links (ISL) are not considered, otherwise they belong to the common delay in addition to the delay on the user link) and $\Delta\tau$ defines the differential component of the delay.
It is worth emphasizing that all sources of the common delay can be pre-compensated, [23]. By means of orbital and geometrical considerations, the maximum difference in slant range between any two beam users is obtained when: i) the beam center is located on the satellite ground track, *i.e.*, $v_{BC} = 0$; and ii) the two users are located at the intersections between the ground track and the beam edge, *i.e.*, $v_{min} = v_{max} = 0$, where min and max define the users located closest and farthest from the satellite, respectively. In a given time instant along its orbit, the satellite is seen at an elevation angle ε_t from the beam center; thus, from Eq.(2.1), it is possible to write:

$$u_{BC} = \frac{R_E}{R_E + h_{sat}} \cos\varepsilon_t \tag{2.4}$$

where R_E is the Earth's radius. Since the beam radius in uv coordinates is known, it is possible to obtain the coordinates of the two users as:

$$\begin{cases} u_{min} = u_{BC} - r_{uv} = \frac{R_E}{R_E+h_{sat}} \cos\varepsilon_t - r_{uv} \\ u_{max} = u_{BC} + r_{uv} = \frac{R_E}{R_E+h_{sat}} \cos\varepsilon_t + r_{uv} \end{cases} \tag{2.5}$$

Moreover, it is worth highlighting that: i) when the Sub-Satellite Point (SSP) is in the beam, the minimum slant range location is given by the SSP itself; and ii) when the maximum slant range point is over the satellite's FoV, the maximum slant range location is on the FoV. From these observations, it is possible to write:

$$u_{min} = \begin{cases} \frac{R_E}{R_E+h_{sat}} \cos\varepsilon_t - r_{uv}, & u_{BC} > r_{uv} \\ 0, & u_{BC} \leq r_{uv} \end{cases} \tag{2.6}$$

and

$$u_{max} = \begin{cases} \frac{R_E}{R_E+h_{sat}} \cos\varepsilon_t + r_{uv}, & u_{BC} + r_{uv} \leq \frac{R_E}{R_E+h_{sat}} \\ \frac{R_E}{R_E+h_{sat}}, & u_{BC} + r_{uv} > \frac{R_E}{R_E+h_{sat}} \end{cases} \tag{2.7}$$

From the u-axis coordinates, the corresponding elevation angles are:

$$\varepsilon_i = \arccos\left(\frac{R_E + h_{sat}}{R_E} u_i\right) \tag{2.8}$$

with $i = min, max$. Knowing the elevation angle ε_i and the nadir angle $\vartheta_i = \arcsin u_i$, the Earth central angle can be computed as $\lambda_i = \pi/2 - \varepsilon_i - \vartheta_i$. This allows obtaining the slant range as:

$$d_i = R_E \frac{\sin \lambda_i}{\sin \vartheta_i} \tag{2.9}$$

Thus, the maximum differential slant range, Δd_{max}, is given by:

$$\Delta d_{max} = R_E \left(\frac{\sin \lambda_{max}}{\sin \vartheta_{max}} - \frac{\sin \lambda_{min}}{\sin \vartheta_{min}} \right) \tag{2.10}$$

From the above equation, it is straightforward to obtain the maximum differential delay as follows:

$$\Delta \tau_{max} \left(\varepsilon_t, h_{sat}, r_{uv} \right) = \frac{R_E}{c} \left(\frac{\sin \lambda_{max}}{\sin \vartheta_{max}} - \frac{\sin \lambda_{min}}{\sin \vartheta_{min}} \right) \tag{2.11}$$

with c being the speed of light. Eq. 2.11 clearly highlights that the maximum differential delay is a function of the elevation angle at the beam center, the beam radius, and the satellite altitude since both the nadir and Earth central angles depend on these parameters from Eq.(2.6), (2.7), and (2.8).

2.3.2 Doppler shift

The Doppler shift consists of the change in the carrier frequency due to the relative motion between the satellite and the user terminal. Similarly to the above analysis for the propagation delay, only the mathematical framework of the maximum differential Doppler shift is discussed since all terms introducing a common shift can be pre-compensated assuming, as per 3GPP current analyses, that: i) the terminals are equipped with Global Navigation Satellite Systems (GNSS) capabilities, and ii) the satellite ephemeris are known.

By means of orbital and geometric considerations, the worst-case scenario (in terms of maximum differential shift between any two users) arises when the beam's major semi-axis lies on the satellite ground track, [28]. For a terminal not located at the beam center, but on the beam major semi-axis, the same Doppler curve applies with a horizontal shift given by the time instant at which that UE will see the satellite at $\varepsilon = \pi/2$. The differential Doppler between any two users is obtained by evaluating the Doppler shift at the corresponding elevation angles and computing the difference. Therefore, the maximum variability is obtained when the UEs are at the two beam edges on the beam major semi-axis. In [29], the authors proposed a simplified formula for the Doppler shift experienced at locations on the satellite orbit projection on the ground as a function of the elevation angle:

$$f_d \left(\varepsilon_i \right) = f_c \frac{\omega_s R_E \cos \varepsilon_i \left(\varepsilon_t \right)}{c} \tag{2.12}$$

where ω_s is the satellite's angular speed and f_c the carrier frequency. It is worth emphasizing that in the above formulation, the elevation angle at the user location ε_i is a function of the elevation angle at beam center ε_t, as a consequence of the dependency of the u-axis coordinate from the beam center elevation angle, discussed above. To provide the maximum differential Doppler shift, the elevation angles at the minimum and maximum Doppler shift locations, *i.e.*, Eq. (2.6) and (2.7), must be computed. From these, it is possible to write:

$$\begin{aligned}\Delta f_{d,max} &= f_d\left(\varepsilon_{max}\right) - f_d\left(\varepsilon_{min}\right)\\ &= 2f_c\frac{R_E + h_{sat}}{c}\omega_s r_{uv}\end{aligned} \tag{2.13}$$

From Eq. 2.13, it is clear that the maximum differential Doppler depends on the satellite altitude and the beam radius only, and not on the elevation angle at the beam center.

2.3.3 Path Loss

For a satellite communication link between a UE on ground and the satellite, the overall losses, L, can be computed as:

$$L = PL + L_E = \underbrace{L_B + L_A + L_{POL}}_{PL} + \underbrace{L_F + L_D}_{L_E} \tag{2.14}$$

where PL represents losses due to the channel impairments while L_E are the losses related to the equipment configuration. In details: *i*) L_B is the basic path loss, which combines free space, clutter, and shadowing losses; *ii*) L_A represents the losses due to atmosphere; *iii*) L_{POL} is the polarisation mismatch loss; *iv*) L_F represents the losses in the equipment; and *v*) L_D represents the depointing losses.

The *basic path loss* is the combination of Free Space Loss (FSL), L_{fs}, clutter loss, L_{cl}, and log-normal shadowing, L_σ:

$$L_B = L_{cl} + L_{fs} + L_\sigma \tag{2.15}$$

The clutter loss models the attenuation of signal power caused by surrounding buildings and objects on the ground. It depends on the elevation angle, ε, computed in Eq. 2.8, the system operating frequency, f_c, and the environment. Typical values for this parameter can be found in [30] for different scenarios and it can be always assumed null in Line of Sight (LoS) conditions. For a generic user located at slant range d from the satellite, the FSL is given by:

$$L_{fs} = 20\log_{10}\left(\frac{4\pi d f_c}{c}\right) \tag{2.16}$$

where c is the speed of light. As for the shadowing loss, L_σ is modeled as a log-normal random variable with zero mean and variance related to the harshness of the shadowing environment, *i.e.*, $L_\sigma \sim (0, \sigma_s^2)$, and the values of, σ_s^2 are provided by 3GPP for dense urban, urban, and rural scenarios as a function of the elevation angle in [30].

Atmospheric losses take into account the atmospheric gases absorption, L_{gas}, the rain/snow fall and cloud attenuation, L_{rain}, and the scintillation losses, L_s:

$$L_A = L_{gas} + L_{rain} + L_s \tag{2.17}$$

Atmospheric gas absorption depends mainly on frequency, elevation angle, altitude above sea level, and water vapor density (absolute humidity). In particular, in order to be compliant with 3GPP standardization, losses are computed as provided in Annex 2 of ITU-R P.676 for slant paths. The atmosphere is modelled with temperature 288.15 K, pressure 1013.25 hPa, and water vapour density 7.5 g/m^3 [31]. Rain and cloud attenuations are dependent on the geographical location of the ground terminal. Section 2.2 of ITU-R P.618-13 describes a method to estimate the long-term statistics of attenuation due to rain [32]. For 3GPP System-Level Simulation (SLS), the baseline is to consider clear sky conditions only and in any case, rain attenuation is considered negligible for frequencies below 6 GHz [30]. Scintillation is a variation of the amplitude of received carriers caused by variations in the refractive index of the troposphere and the ionosphere. The tropospheric scintillations, impacting signals in Ka-band, are modeled as a fixed term depending on the user elevation angle [30]. These values are obtained by means of the procedure described in ITU-R P.618 [32]. The ionospheric scintillations, impacting signals in S-band, are modeled as a fixed term of 2.2 dB as from [30], and described in ITU-R P.531-13 [33].

It is also necessary to consider the *polarisation mismatch loss* observed when the receiving antenna is not oriented with the polarisation of the received wave because propagation through the atmosphere can also affect the polarization. In fact, the ionosphere introduces a rotation of the plane of polarization of an angle, $\Delta\psi$ which is inversely proportional to the square of the frequency. This rotation is particularly dangerous for linear polarization. Furthermore, with linear polarisation, the receiving antenna may not have its plane of polarisation aligned with that of the incident wave. In general, the polarization mismatch loss can be defined as [19]:

$$L_{POL} = -20 \log_{10}(\cos \Delta\psi) \tag{2.18}$$

The *equipment losses* represents the losses in the transmitting and receiving equipment respectively, in the feeder between the power amplifier and the antenna:

$$L_F = L_{FTX} + L_{FRX} \tag{2.19}$$

In particular, L_{FTX} is the feeder loss between the transmitter and the antenna, while L_{FRX} is the feeder loss between the antenna and the receiver.

Finally, the *depointing losses* are functions of the transmission and reception angles misalignment, θ_T and θ_R respectively, with respect to the antenna boresight. The result is a fallout of antenna gain with respect to the maximum gain on transmission and on reception, which can be formulated as a function of the θ_{3dB}, later explained in this chapter:

$$L_D = L_T + L_R = 12\left(\frac{\theta_T}{\theta_{3dB}}\right) + 12\left(\frac{\theta_R}{\theta_{3dB}}\right) \tag{2.20}$$

Chapter 3

O-RAN Specification

This chapter provides an overview of the O-RAN specifications, focusing on its core architectural principles: disaggregation, virtualization, open interfaces, and data-driven control. It examines the functional split of base stations into CU, DU, and RU, and the virtualization of network components within the O-Cloud. Key O-RAN interfaces—E2, A1, O1, and Fronthaul—are discussed, highlighting their role in enabling interoperability and modularity. The chapter also explores the RAN Intelligent Controllers (non-RT RIC and near-RT RIC) and their applications (rApps and xApps), which utilize AI/Machine Learning (ML) to enable closed-loop control and optimize network performance.

3.1 Architectural concepts

The O-RAN architecture is based on the principles of: i) disaggregation; ii) virtualization; iii) data-driven RAN control; and iv) open interfaces. These are presented below.

3.1.1 Disaggregation

The disaggregation principle extends the functional disaggregation paradigm proposed by 3GPP for the NR gNB, [34], effectively splitting base stations into different functional units. Referring to Figure 3.1, the gNB is split into a Central Unit (O-CU), a DU, and a RU, with the CU divided also in the CP and UP. Thanks to this logical split, different functionalities can be deployed at different network elements and on specialized or general-purpose hardware platforms. The functional-split option is defined by how the protocol-stack functions are allocated between the CU and DU within the gNB. In O-RAN the selected split is 7.2x. In this case, the RU takes care of the Fast Fourier Transform (FFT) and of the cyclic prefix addition/removal operations. The remaining functions of the physical link are then implemented in the DU, together with the MAC and RLC layers. Finally, all the remaining functions of the 3GPP protocol stack are centralized in the CU, i.e., the RRC, Service Data Application Layer (SDAP), and the PDCP.

3.1.2 Virtualization

As defined in [35], all the O-RAN architecture components shown in Figure 3.1 can be deployed as virtualized components on a hybrid cloud computing platform called O-Cloud. This platform specialises the virtualization paradigm of O-RAN combining physical nodes, software components, and management and orchestration functionalities. The virtualization concept enables: i) the decoupling between software and hardware components; ii) the definition of standardized hardware capabilities to use in the O-RAN infrastructure; iii) sharing of the computational capabilities of the hardware; and iv) the automated deployment of the RAN functions on the hardware platforms.

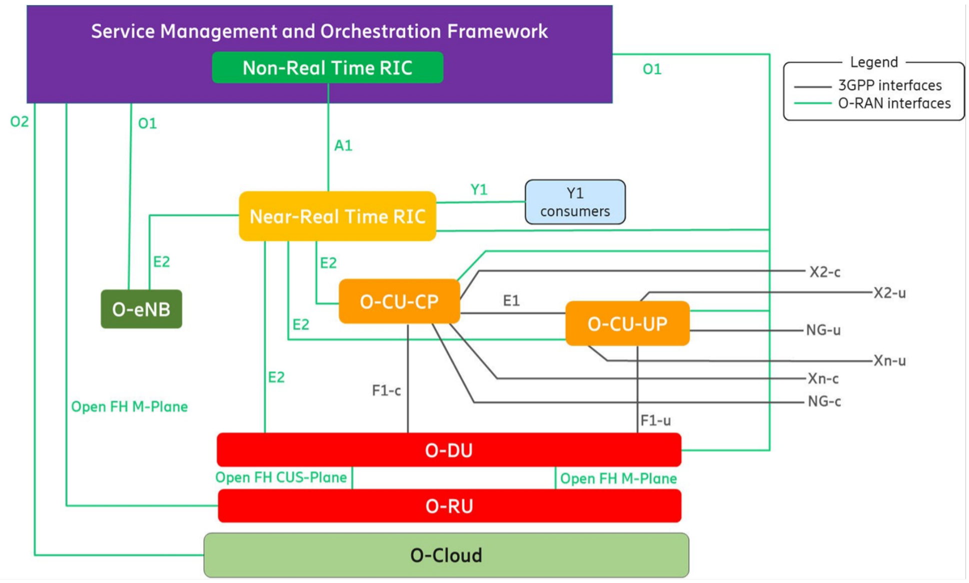


FIGURE 3.1: O-RAN architecture (components and interfaces).

3.1.3 Open Interfaces

O-RAN introduces technical specifications describing open interfaces connecting different components of the architecture. Figure 3.1 shows the O-RAN open interfaces interconnecting the architecture components along with the 3GPP defined interfaces. Indeed, O-RAN includes and leverages 3GPP defined interfaces to additionally foster the disaggregation of the RAN. Leveraging open interfaces is possible to deploy the O-RAN architecture described in Figure 3.1 selecting different network locations for the virtualized components, with multiple possible configurations. An additional benefit of open interfaces is to break the vendor lock-in inside the RAN, fostering market competitiveness, faster components upgrade, and facilitating the introduction of additional virtualized components in the RAN. We will provide a description of each interface in section 3.2.

3.1.4 Closed-Loop Control

In order to orchestrate the RAN, O-RAN introduces the RICs. These, thanks to data pipelines that stream the KPI of system nodes, have an abstract and centralized point of view on the network. By processing this data and exploiting AI and ML algorithms, the RICs can optimize and apply the control policies of the RAN in a closed loop. With reference to Figure 3.2, O-RAN foresees the non Real Time (non-RT) RIC and the near Real Time (near-RT) RIC, differentiated on the role and on the timescale of intervention. The former operates on a time scale longer than 1 s and provides guidance, enrichment information, and management of ML models for the near-RT RIC. The latter consists of multiple applications supporting custom ML models deployed at the edge and operating on a time scale between 10 ms and 1 s. In Figure 3.2 are highlighted the closed-loop controls on the disaggregated O-RAN infrastructure enabled by the different RICs. Control loops that operate in the real-time domain (below 10 ms) are also considered, even if they are not yet included in the current O-RAN architecture and are mentioned for further study. These could be used for Radio Resource Management, beam management, or detection of physical layer parameters.

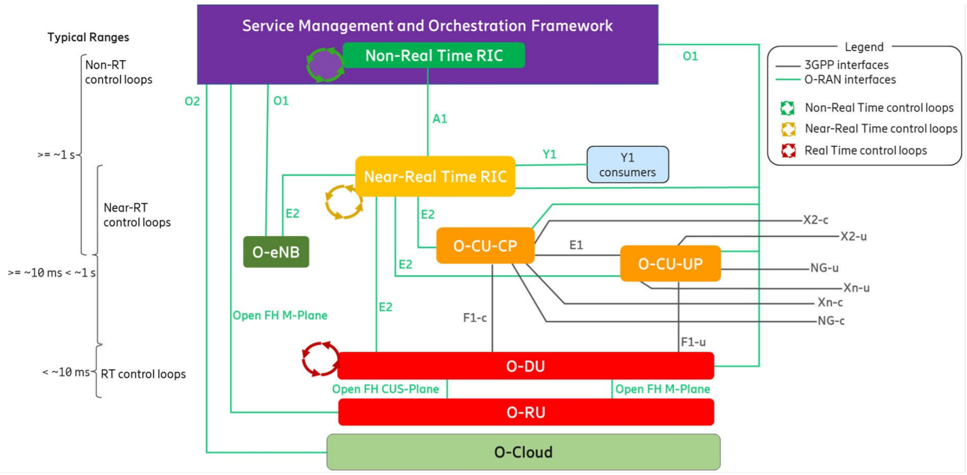


FIGURE 3.2: O-RAN Control loops.

3.2 O-RAN Open Interfaces

In this section we review the interfaces standardized by the O-RAN Alliance so far, detailing their logical abstractions and procedures. More specifically, the E2, O1, A1, and Fronthaul interfaces are detailed below, together with other O-RAN and 3GPP defined interfaces.

3.2.1 E2 Interface

The E2 interface is an open interface interconnecting the near-RT RIC with the E2 nodes, i.e., CUs, DUs, and Long Term Evolution (LTE) eNBs designed to be O-RAN compliant. This interface enables the RIC to collect metrics from the RAN components periodically or after trigger events. The metrics can then be used by the RIC to control functionalities and procedures of the E2 nodes. The data collection and control services can have different granularities, i.e., specific UEs, one cell, multiple cells, or Quality of Service (QoS) classes. The single UE and groups of UEs are identified by the O-RAN Alliance using a variety of unique identifiers. To identify the gNBs, QoS classes, and slices, OO-RAN relies on identifiers based on 3GPP specifications, while specific UEs are identified relying on O-RAN introduced user identifier UE-ID. Each E2 node exposes different capabilities and the services it supports, i.e., different DUs can expose different parameters to be tuned, along with their capability collect specific network metrics. Additionally, the capabilities of each node are clearly separate, and the RIC-RAN interaction is clearly defined relying on a publish-subscribe mechanics. The E2 Application Protocol (E2AP), [36], coordinates the communication between the near-RT RIC and the E2 nodes, and provides a set of services:

- E2 Setup: is used to establish the E2 interface between the Near-RT RIC and an E2 Node.
- E2 Reset: is used by either the E2 Node or Near-RT RIC to reset the E2 interface.
- Near-RT RIC Service Update: is used by the E2 Node to inform the Near-RT RIC of any change to the list of supported RIC services and mapping of services to functions within the E2 Node.
- E2 Node configuration update: is used by the E2 Node to inform the Near-RT RIC of any change to the list of supported RIC services and mapping of services to functions within the E2 Node.
- E2 Removal: is used by either the E2 Node or Near-RT RIC to release the E2 signalling connection.

After the connection is established, the following E2AP services can be exploited to implement the E2 service models.

- Report: Near-RT RIC uses a RIC Subscription and/or RIC Subscription Modification procedures to request that E2 Node sends a report message to Near-RT RIC and the associated procedure continues in the E2 Node after each occurrence of a defined RIC Subscription procedure Event Trigger.
- Insert: Near-RT RIC uses a RIC Subscription and/or RIC Subscription Modification procedures to request that E2 Node sends an insert message to Near-RT RIC and suspends the associated procedure in the E2 Node after each occurrence of a defined RIC Subscription procedure Event Trigger.

- Control: Near-RT RIC sends a control message to E2 Node to initiate a new associated procedure or resume a previously suspended associated procedure in the E2 Node.

- Policy: Near-RT RIC uses a RIC Subscription and/or RIC Subscription Modification procedures to request that E2 Node executes a specific policy during functioning of the E2 Node after each occurrence of a defined RIC Subscription procedure Event Trigger.

- Query: Near-RT RIC sends a query message to the E2 node to retrieve RAN-related and/or UE-related information from the E2 Node.

- E2 Service Model (E2S), [37]: The E2S is inserted as payload in one of the E2AP messages. The O-RAN Alliance has standardized three service models:

- E2S Key Performance Measurement (KPM), [37], reports the RAN performance metrics, exploiting E2 report messages. Precisely, the procedure is as follows: i) in the E2 setup procedures, the metrics exposed by the E2 nodes are advertised; ii) an xApp in the RIC sends a subscription message indicating which KPMs are of interest; iii) the E2 node streams the selected KPMs trough Indication messages of type Report.

- E2S Network Interface (NI), [37], takes the messages received by the E2 node on the network interfaces and delivers them to the near-RT RIC exploiting the E2 report messages. The E2 node advertises which interfaces it supports during the subscription procedure, and they include X2 (which connects LTE eNBs), Xn (which connects different NR gNBs), and F1, which connects DUs and CUs).

- E2S Cell Configuration and Control (CCC), [38], controls and re-configure the E2 nodes at a cell or node level, e.g., for the bandwidth part configuration. It relies primarily on E2 report and control messages. The technical specification for CCC, in its current version, specifies the control of the selection of X2 and Xn neighbours, RAN slicing, and parameters related to the bandwidth part and the synchronization signals of a cell.

- E2S Ran Control (RC) implements control functionalities through E2 control services. Compared to E2S CCC, it focuses on more granular control (up to the UE or bearer level). It also provides capabilities for UE identification and UE information reporting.

3.2.2 O1 Interface

O1, [39], is an open interface that standardizes operations and maintenance practices. It interconnects the O-RAN managed elements (including the near-RT RIC, RAN nodes) to the Service Management and Orchestration (SMO) and the non-RT

RIC. O1 interface uses a combination of REST/HTTPS Application Programming Interface (API)s and Network configuration protocol (NETCONF), [40], which is a protocol standardized by theInternet Engineering Task Force (IETF) for the lifecycle management of networked functions. The O1 interface enablesManagement Services (MnS) including: i) O-RAN components' lifecycle management; ii) trace collection trough KPI reports and performance assurance; and iii) software and file management. To this aim, this interface connects a MnS provider, i.e., the node managed by the SMO to one MnS consumer, i.e., the SMO.

The defined management services are described below:

- Provisioning Management services: they allow a Provisioning MnS Consumer to configure attributes of managed objects on the Provisioning MnS Provider that modify the Provisioning MnS Provider's capabilities in its role in end-to-end network services and allows a Provisioning MnS Provider to report configuration changes to the Provisioning MnS Consumer. NETCONF is used for the Provisioning Management Services to Create Managed Object Instance, Delete Managed Object Instance, Modify Managed Object Instance Attributes and Read Managed Object Instance Attributes. A REST/HTTPS event is used to notify the Provisioning MnS subscribed Consumers when a configuration change occurs.

- Fault Supervision Management Services: they allow a Fault Supervision MnS Provider to report errors and events to a Fault Supervision MnS Consumer and allows a Fault Supervision MnS Consumer to perform fault supervision operations on the Fault Supervision MnS Provider, such as get alarm list.

- Performance Assurance Management Services: they allow a Performance Assurance MnS Provider to report file-based (bulk) and/or streaming (real time) performance data to a Performance Assurance MnS Consumer and allows a Performance Assurance MnS Consumer to perform performance assurance operations on the Performance Assurance MnS Provider, such as selecting the measurements to be reported and setting the frequency of reporting.

- Trace Management Services: they allow a Trace MnS Provider to report file-based or streaming trace records to the Trace MnS Consumer. Trace Control provides the ability for the Trace Consumer to start a trace session by configuring a Trace Job via the Trace Control IOC or by establishing a trace session that will propagate trace parameters to other trace management providers via signalling. There are multiple levels of trace that can be supported on the provider. The Trace Provider may be configured to support filebased trace reporting or streaming trace reporting.

- File Management Services: they allow a File Management MnS Consumer to request the transfer of files between the File Management MnS Provider and the File Management MnS Consumer.

- Heartbeat Management Services: they allow a Heartbeat MnS Provider to send heartbeats to the Heartbeat MnS Consumer and allow the Heartbeat MnS Consumer to configure the heartbeat services on the Heartbeat MnS Provider.
- Physical Network Function (PNF) Startup and Registration Management Services: they allow a physical PNF Startup and Registration MnS Provider to acquire its network layer parameters either via static procedures (pre-configured in the element) or via dynamic procedures (Plug-n-Play) during startup. During this process, the PNF Startup and Registration MnS Provider also acquires the IP address of the PNF Startup and Registration MnS Consumer for PNF Startup and Registration MnS Provider registration. Once the PNF Startup and Registration MnS Provider registers, the PNF Startup and Registration MnS Consumer can then bring the PNF Startup and Registration MnS Provider to an operational state.
- PNF Software Management Services: they allow a PNF Software MnS Consumer to request a physical PNF Software MnS Provider to download, install, validate and activate a new software package and allow a physical PNF Software MnS Provider to report its software versions. O-RAN will utilize the liaison to 3GPP to initiate enhancements to the 3GPP specifications for PNF Software Management. Until those enhancements are put in place, O-RAN PNF Software Management will be described in this specification. Software management described in this document is modelled on the O-RAN Fronthaul Management Plane Specification.

3.2.3 A1 Interface

The A1 is an open interface interconnecting the two O-RAN specific components (the non-RT RIC and the near-RT RIC), [40]. Exploiting the A1 interface, the non-RT RIC is able to:

- deploy policy-based guidance to the near-RT RIC, e.g., to set optimization goals;
- manage ML models used in xApps;
- orchestrate and negotiate the collection of enrichment information from the network to the near-RT RIC.

The network policies, ML models, and enrichment information can be applied at different levels in order to refer a group of UEs or to a single UE. A1 interface relies on the A1 Application Protocol (A1AP), designed for policy deployment and network functions combining REST APIs over HTTP for the transfer of JSON objects. As the A1-based ML model management is still considered for further studies, [41], the A1 interface functions are:

- **Policy Management**: the purpose of A1 policies is to enable the Non-RT RIC function in the SMO to guide the Near-RT RIC function, and hence the RAN, towards a better fulfilment of the RAN intent. By utilising the observability over O1, and the A1 policy feedback, the Non-RT RIC can conclude that the RAN intent is not achieved. The NonRT RIC can then decide to use A1 policies that enable the Near-RT RIC to, e.g., optimize Radio Resource Management (RRM) for a single UEs or for group of UEs. There are different types of A1 policies referred to as A1 policy types. A Non-RT RIC need not use all A1 policy types, and a specific function in the Near-RT RIC may only support one specific A1 policy type. Non-RT RIC can discover available A1 policy types over the A1 interface. An A1 policy type is identified by the policy type identifier (PolicyTypeId). Different policy types have different PolicyTypeIds. Based on the PolicyTypeId, schemas are identified and used for creation, validation, and formulation, and for query of the status, of A1 policies of that type. An A1 policy is identified by a policy identifier (PolicyId) that shall be assigned by Non-RT RIC. The PolicyId shall be locally unique within Non-RT RIC and sent in the policy request operations that carry representations of A1 policies. A1 policies consist of a scope identifier and one or more policy statements. The scope identifier represents what the policy statements are to be applied on (e.g., UEs, QoS flows, or cells). The policy statements represent the goals to the Near-RT RIC and covers policy objectives and policy resources.

- **Enrichment Information (EI)**: the purpose of A1 enrichment information is to enable the Near-RT RIC to improve its RAN optimization performance by utilising information that is not available within the RAN. The information sources can be O-RAN internal and O-RAN external, and the derived A1 enrichment information can be provided by the Non-RT RIC over the A1 interface. There are different types of A1 enrichment information referred to as EI types. A Near-RT RIC may not need to use all EI types, and a specific function in the Near-RT RIC may only need one specific type of A1 enrichment information. An EI type is identified by the EI type identifier (EiTypeId). Based on the EiTypeId, information can be provided about the A1 enrichment information properties and how to request delivery of the A1 enrichment information. The Near-RT RIC can discover available EiTypeIds over A1 and request delivery of A1 enrichment information related to an available EiTypeId. The Non-RT RIC controls the access to A1 enrichment information and how the connection for delivery of the enrichment information can be made. The enrichment information function is used by the Non-RT RIC to produce and make A1 enrichment information available to the Near-RT RIC. The Non-RT RIC is responsible for exposure and secure delivery of A1 enrichment information.

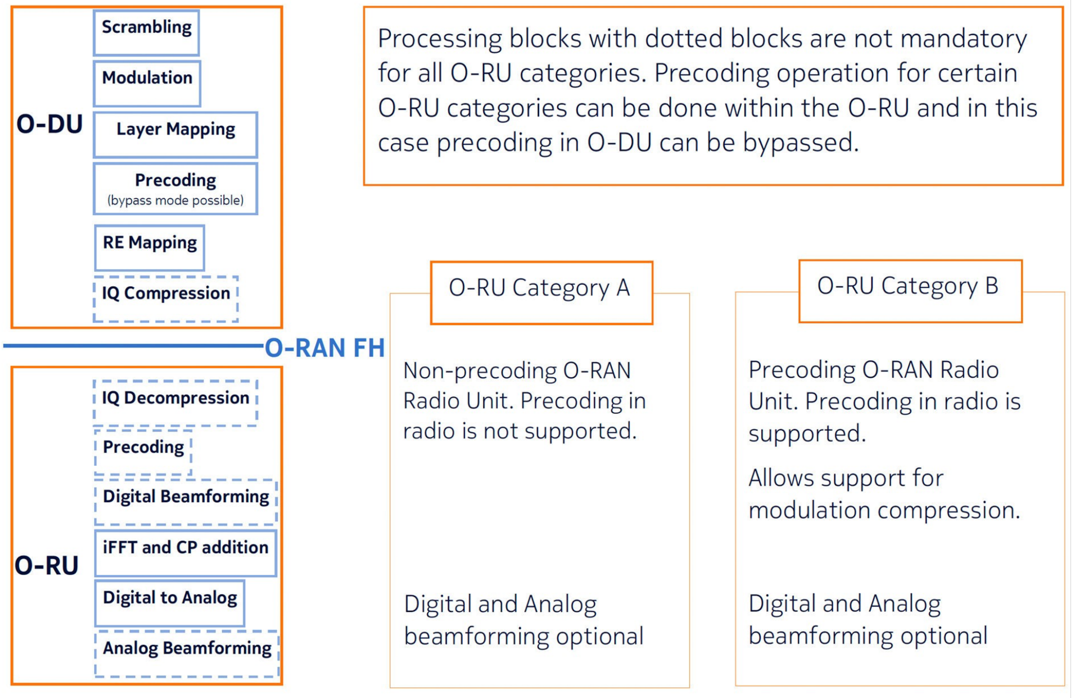

O-DU
Scrambling
Modulation
Layer Mapping
Precoding
(bypass mode possible)
RE Mapping
IQ Compression
O-RAN FH
O-RU
IQ Decompression
Precoding
Digital Beamforming
iFFT and CP addition
Digital to Analog
Analog Beamforming
Processing blocks with dotted blocks are not mandatory for all O-RU categories. Precoding operation for certain O-RU categories can be done within the O-RU and in this case precoding in O-DU can be bypassed.
O-RU Category A
Non-precoding O-RAN Radio Unit. Precoding in radio is not supported.
Digital and Analog beamforming optional
O-RU Category B
Precoding O-RAN Radio Unit. Precoding in radio is supported.
Allows support for modulation compression.
Digital and Analog beamforming optional

FIGURE 3.3: Split Point and Category A and Category B O-RAN Radio Units.

3.2.4 O-RAN Fronthaul

The O-RAN Fronthaul (FH) interface connects a DU to one or multiple RUs inside the same gNB, [42], and enables the distribution of the physical layer functionalities between the RU and the DU, and to control RU operations from the DU. When considering the functional split defining a fronthaul interface there are two competing interests: the simplicity of the interface and of the RU design, and the data rate required for fronthaul transport. To resolve this, O-RAN has selected a single split point, known as "7-2x" but allows a variation, with the precoding function to be located either "above" the interface in the O-DU or "below" the interface in the O-RU. For the most part the interface is not affected by this decision, but there are some impacts namely to provide the necessary information to the O-RU to execute the precoding operation. O-RUs within which the precoding is not done (therefore of lower complexity) are called "Category A" O-RUs while O-RUs within which the precoding is done are called "Category B" O-RUs. See Figure 3.3 for a depiction of this dual O-RU concept. The FH interface can be based on Ethernet or UPD/IP encapsulation, carrying either an enhanced Common Public Radio Interface (eCPRI), [43], or an IEEE 1914.3, [44], payload. Since one DU can support multiple Rus, the FH specification introduces and additional component with the aim to multiples the fronthaul stream to multiple RUs. As an alternative, the RUs can be connected in a chain. The O-RAN FH specification foresees four different communication planes detailed below.

C-plane: this plane transfers commands from the DU and the RU, i.e., from the high-PHY to the low-PHY, including:

- scheduling and beamforming configurations;
- management of different NR numerologies in different subframes;
- downlink precoding configuration;
- spectrum sharing control.

The C-plane messages are encapsulated in eCPRI or IEEE 1914.3 protocols, with specific fields and commands for different control procedures.

U-plane: This communication plane is mainly used to transfer I/Q samples in the frequency domain between the DU and the RU. Typically, the U-plane messages follow a C-plane one that specifies scheduling and beamforming configurations, so that the I/Q samples can be transmitted in the corresponding transmission opportunities. An additional functionality of the U-plane is to take care of the transmission timing of the packets so that the RU has enough time for processing before transmission. Additionally, the U-plane specifies the digital gain of the samples and, for more efficient transfer, it can compress them.

S-plane: this communication plane takes care of synchronization between the DU and RU in terms of time, frequency, and Phase between the clocks. Having a shared clock reference enable the DU and RU to properly align time and frequency resources for the transmission of data and control channels. This is of paramount importance in a slotted distributed system. The topologies foreseen by the O-RAN specification differ on the type of interconnection between the DU and RU that can be either a direct link or an indirect link trough a fabric of ethernet switches. Additionally, the synchronization can be based in different protocol with different precisions: i) Physical Layer Frequency Signals (PLFS) or Precision Time Protocol (PTP), with sub-microsecond accuracy.

M-plane: This protocol enables the initialization of the DU and RU, and the management of their interconnection. It is also in charge of the configuration of the RU, [45]. Specifically, it takes care of:

- managing the RU start-up, during which the RU establishes the management with the DU and/or the SMO;
- enabling software updates, configuration management, performance and fault monitoring, and file management for bulk transfer of data;
- managing the registration of the RU as PNF, the parameters of the RU-to-DU link, and the update of beamforming vectors.

M-plane relies on a IPv4 or IPv6 tunnel with dedicated endpoints in the DU and RU, and runs in parallel to the C-, U-, and S-planes. The specification foresees two architectural options for the M-plane implementation:

- Hierarchical: the SMO manages the DU and the DU manages the RU;
- Hybrid: the SMO manages the DU and can also interact directly with the RU.

Contrary to the C-/U-/S-planes, the M-plane is end-to-end encrypted through SSH and/or TLS.

3.3 Orchestration Framework and Non-RT RIC

Non-Real Time RAN Intelligent Controller (Non-RT RIC) is the intelligent RAN operation and optimization functionality internal to the SMO framework in O-RAN overall architecture. Indeed, it represents a subset of functionality of the SMO framework. Non-RT RIC logically terminates the A1 interface, and it provides policybased guidance, enrichment information, and AI/ML model management to the Near-RT RICs. It can also access other SMO framework functionalities, for example influencing what is carried across the O1 and O2 interface. Non-RT RIC is comprised of:

- Non-RT RIC Framework – Functionality internal to the SMO Framework that logically terminates the A1 interface to 1 the Near-RT RIC and exposes set of R1 services to NonRT RIC Applications (rApps);
- rApps: Applications that leverage the functionalities available in the Non-RT RIC Framework / SMO Framework to provide value added services related to RAN operation and optimization. The scope of rApps includes, but is not limited to, radio resource management, data analytics, and providing enrichment information.

In Figure 3.4 is shown the high-level architecture of the SMO. The O-RAN specification do not set a strict boundary between the SMO and non-RT RIC functionalities but group them into three distinct sets:

- functions anchored inside the non-RT RIC (in blue);
- functions anchored outside the non-RT RIC (in orange);
- functions not anchored to any SMO component or spanning multiple components (in white).

In this section we describe the SMO framework functionalities and interfaces.

3.3.1 Non-RT RIC

The non-RT RIC enables closed-loop control of the RAN (with time scales larger then 1 s) supporting the execution of third-party applications, the rApps. These are used to enable value added services to support RAN optimization and RAN operations, including: i) policy guidance; ii) enrichment information; iii) configuration management; and iv) data analytics. Inside the non-RT RIC is defined the R1 termination,

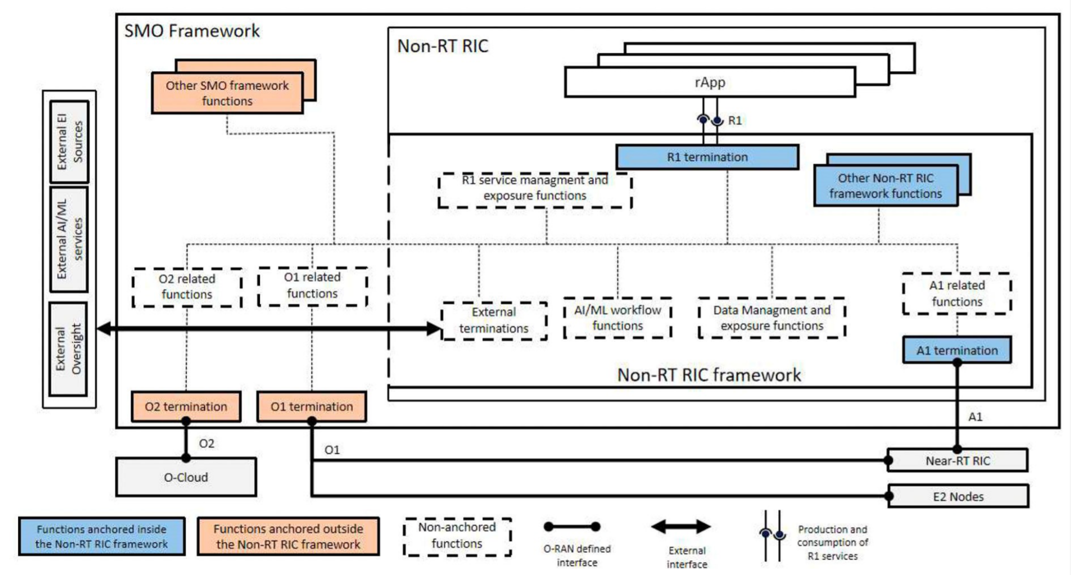


FIGURE 3.4: SMO high-level architecture.

shown in Figure 3.4, which interfaces rApps with the non-RT RIC and allows them to obtain access to: i) data management and exposure services; ii) AI/ML functionalities; and iii) A1, O1 and O2 interfaces through the internal messaging infrastructure. Although rApps are very similar to xApps in the control functionalities they can provide, they have been designed to provide control policies that operate at a larger scale, such as RAN sharing, performance diagnostics, frequency and interference management, and network slicing.

3.3.2 Near-RT RIC

The Near-RT RIC is deployed at the edge of the network to operate control-loops over the CUs and DUs in the RAN, as well as over O-RAN compliant eNBs. Usually, the near-RT RIC controls multiple RAN nodes, so its closed-loop control function is associated with the UEs of several cells. The control functionality of the near-RT RIC is delegated to the xApps, multiple application deployed inside the RIC that support custom logic. The xApps receive KPIs data from the RAN at all different layers, i.e., user, cell, or slice, and computes and applies control policies. NearRT RIC shall consist of multiple xApps and a set of platform functions that are commonly used to support the specific functions hosted by xApps.

An overview of the architecture of the O-RAN standardized near-RT RIC is provided in Figure 3.5. The architecture includes:

- **Conflict mitigation**: in the context of Near-RT RIC, Conflict Mitigation is about addressing conflicting interactions between different xApps. An application will typically change one or more parameters with the objective of optimizing a specific metric. Conflict Mitigation is necessary because xApps objectives may be chosen/configured such that they result in conflicting actions. The

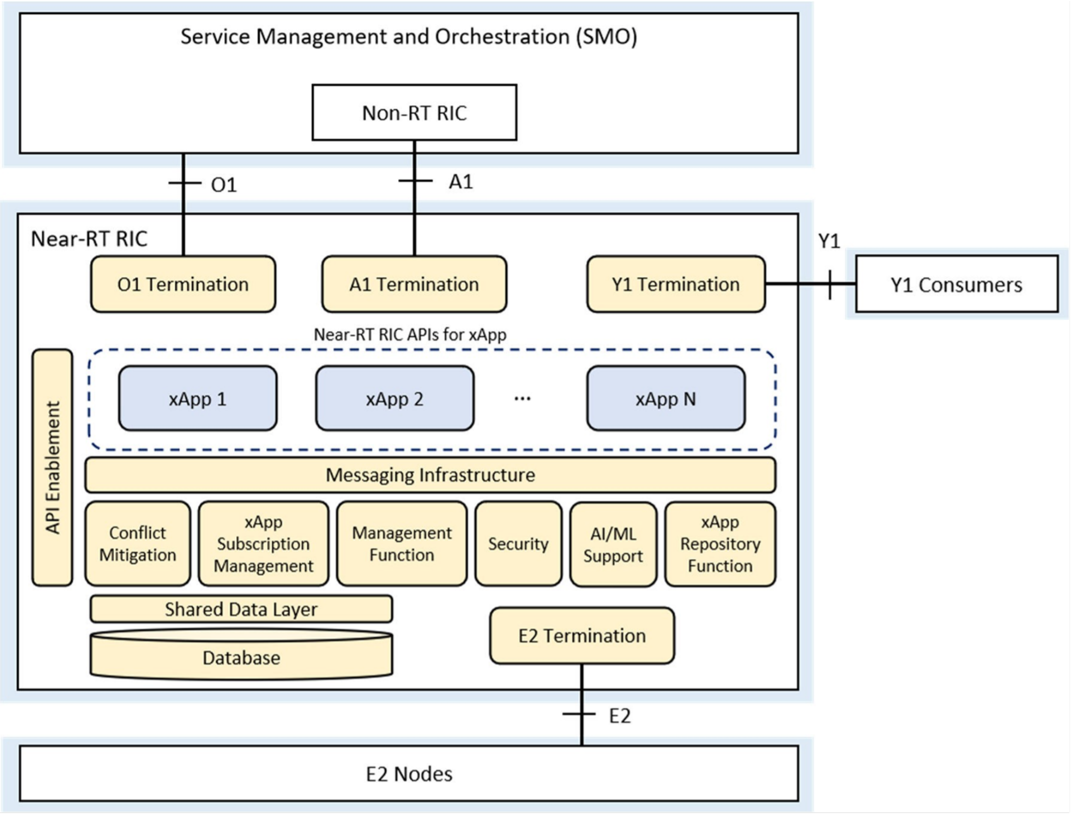


FIGURE 3.5: Near-RT RIC Internal Architecture.

control target of the radio resource management can be a cell, a UE or a bearer, etc. The control contents of the radio resource management can cover access control, bearer control, handover control, QoS control, resource assignment and so on. The control time span indicates the valid control duration which is expected by the control request. The conflicts of control can be illustrated as: i) Direct Conflicts, that can be observed directly by Conflict Mitigation; and ii) Indirect Conflicts, that cannot be observed directly, nevertheless, some dependence among the parameters and resources that the xApps target can be observed. Conflict Mitigation may anticipate the possible conflicts and take actions to mitigate them.

- **Internal messaging infrastructure**: it provides low-latency message delivery service between Near-RT RIC internal endpoints. It needs to support: i) registration message used from endpoints to register themselves to the messaging infrastructure; ii) discovery message used to discover endpoints by the messaging infrastructure initially and registered to the messaging infrastructure; and iii) deletion message to delete endpoints one they are not used anymore. It also provides APIs to send and receive messages from the xApps. This APIs can rely on point-to-point communications or publish/subscribe mechanisms.

- **Subscription manager**: the subscription management functionality manages subscriptions from xApps to E2 Nodes and enforces authorization of policies controlling xApp access to messages. It also enables merging of identical subscriptions from different xApps into a single subscription toward an E2 Node.

- **Security**: to prevent malicious xApps from leaking sensitive RAN data or from affecting the RAN performance. The details of this component are still left for further studies;
- **Network Information Based (NIB) Database and Shared Data Layer API**: the RAN NIB contains information about the E2 nodes and the UE-NIB contains the identification of the UEs and their entries. The Shared Data Layer (SDL) is used by xApps to subscribe to database notification services and to read, write and modify information stored on the database. UE-NIB, R-NIB and other use case specific information may be exposed using the SDL services.
- **xApp management**: this service features automated life-cycle management of the xApps. It accounts for onboarding, deployment, termination, and tracing and logging of Fault, Configuration, Accounting, Performance, Security (FCAPS).
- **AI/ML support**: the AI/ML data pipeline in Near-RT RIC offers data ingestion and preparation for applications (xApps). The input to the AI/ML data pipeline may include E2 node data collected over E2 interface, enrichment information over A1 interface, information from applications, and data retrieved from the Near-RT RIC database through the messaging infrastructure. The output of the AI/ML data pipeline may be provided to the AI/ML training capability in Near-RT RIC. The AI/ML training in Near-RT RIC offers training of applications (xApps) within Near-RT RIC, [35]. The AI/ML training provides generic and use case-independent capabilities to AI/ML-based applications that may be useful to multiple use cases.

3.3.2.1 xApps

The Near-RT xApp platform allows the roll-out of smart apps for management and optimisation of the RAN. These applications can have access to the RAN data as never before. They can use this to feed near-real time control functions, leveraging the benefits of AI and Big Data. This open platform allows third-party apps to complement the RAN vendors portfolio. It can implement functions such as the following:

- **Intelligent and autonomous Neighbour Manager**: Manage frequency planning or force handover to another cell when two UEs are close to the cell boundary and close to each other to avoid interference in MU-MIMO network.
- **Radio resource management**: xApps can be used to optimize the use of radio resources in the O-RAN architecture, such as by allocating frequency bands or scheduling transmissions.

- **Mobility management**: xApps can be used to manage the movement of devices within the O-RAN architecture, such as by providing handover support or tracking device location.
- **QoS**: xApps can be used to implement QoS control functions, such as traffic management or congestion control, to optimize the performance of the O-RAN architecture.
- **Near zero-touch network provisioning**: Adding new sites, CU, DU or RU will automatically provision the network to integrate it and optimise their use.
- **Smarter Handover manager**: Manage handover not only on radio quality but from past handovers using machine learning algorithms, using knowledge of success of failure of past handovers
- **Automated Physical Cell ID (PCI) Allocation**: Avoid and learning from past PCI collisions by a smart allocation and re-allocation
- **Security**: xApps can be used to implement security functions, such as authentication and encryption

3.4 O-RAN in NTN: State of the Art Analysis

The idea of having an open RAN has gained a lot of interest in recent years in the TN community, which has performed a vast analysis. Its open interfaces and closed-loop control enable proactive management elements that take care of networking tasks. The research on these applications has focused on three main categories, based on their expected latency, as specified in [46]: *i*) non-real-time (non-RT) applications, with a latency larger than 1 second; *ii*) near-real-time (near-RT) applications, with latency between 10ms and 1s, and *iii*) real-time applications.

In the non-RT field, the interest has been mainly in network orchestration. The authors in [47] present a novel orchestration framework that builds upon the Open RAN paradigm. It evaluates the optimal set of data-driven algorithms and the best execution location and functional split in an automatic way, to meet the needs of the Network Operator (NO). In [48] the authors propose a reinforcement learning dynamic functional split to choose the optimal splitting point, while in [49] the same problem is solved with a heuristic algorithm.

For what concerns the near-RT applications, the research has focused on different aspects of network edge management. Handover management is tackled in [50], the authors propose a new approach to Automatic Neighbour Relationship (ANR) optimization exploiting O-RAN architecture and develop a ML based optimization technique to improve gNB handovers. A similar approach has been followed in [51], the authors optimize the handover performances through an intelligent access control scheme with Deep Reinforcement Learning (DRL). Resource allocation optimization in O-RAN is tackled in [52] where network slicing is used to study the service-aware baseband resource allocation and Virtual Network Function (VNF) activation.

Real-time applications are still not fully supported by the O-RAN standard but are gaining interest to implement specific AI algorithms in the far network edges. An example is the industrial project [53], where a cognitive MAC layer to predict user Equipment's mobility was built.

The O-RAN functional split architecture brings flexible and scalable network deployments, and it relies on network interfaces with limited throughput and latency performances, as underlined in [54]. Therefore, the Authors in [54] provide a survey of the 3GPP and O-RAN fronthaul compression techniques. [55] addresses the channel information aging of MIMO base stations due to the fronthaul interface latency. The channel information being present at one side of the split takes time to reach the other side to implement uplink beamforming, causing air-interface performance degradation. Additionally, the work in [56] focuses on open interfaces security, since they may be exposed to a plethora of threats, emphasizing missing authentication and authorization vulnerabilities. In [57], MACsec, a standard security protocol that operates in the data-link layer bringing high datarate performance, is proposed as a potential protection solution for the O-RAN Fronthaul interface.

The exploitation of O-RAN in NTN has been analyzed but requires more research effort to reach maturity. Precisely, the O-RAN application to UAVs applications has gained more appeal compared to its exploitation in SatCom. In [58], to minimize the inter-cell interference generated by a video streaming UAV, the authors propose a closed-loop control system based on O-RAN that optimizes UAV's location and its transmission direction. The authors in [59] take into consideration a flying base station on a UAV system and propose a method to jointly optimize the UAVs trajectory and the offloading tasks based on O-RAN-enabled AI. In [60], the opportunities of exploiting O-RAN latency consciousness of the network segments to enable UAV manual real-time control and autonomous drive are analyzed.Surprisingly, the O-RAN exploitation in the SatCom field has been addressed only in a single work, [61]. Here the authors exploit closed-loop feedback and open interface standards to control the interference between congested terrestrial and non-terrestrial systems. The work presents a spectrum-sharing architecture between terrestrial 5G and LEO military satellite systems based on a spectrum sensor. It is important to underline that this inter-system overall interference management would not be possible without the exploitation of the O-RAN architecture. Indeed, while O-RAN increases the system complexity, it equally increases the system design degrees of freedom and enables otherwise unreachable optimizations. This understanding, along with the recent history of O-RAN success in TNs illustrated in this paragraph, has motivated us to start this novel research branch about O-RAN exploitation in NTNs. Indeed, in our opinion, the O-RAN architecture can enhance the NTN systems with a plethora of unexplored applications. Among all use cases, we underline: *i*) the full exploitation of AI models must rely on data-collection pipelines and centralized intelligence provided by open RAN approaches; *ii*) the optimal allocation of RAN functions to the different network nodes requires the O-RAN enabled disaggregation and virtualization of the RAN, along with central network-status knowledge, and *iii*) the leveraging of the O-RAN standardized and open network interfaces to pursue NTN systems interoperability.

Chapter 4

Virtualized 3GPP NTN System Implementation

The following chapter presents the benefits of implementing a virtualized 3GPP NTN system, highlighting how software-driven solutions for both RAN and CN can be more easily upgraded to accommodate new 3GPP features. Due to the rapid pace of standardization initiatives, the intrinsic adaptability of virtualization proves to be exceptionally valuable. By decoupling network functions from fixed hardware platforms, operators can deploy and update networks more quickly as standards evolve, avoiding the extensive overhauls typically associated with hardware-based systems. This the standardization path of NTN in 3GPP, with a particular focus on NTN architectures as defined by the 3GPP standardization process. It details the NTN-specific features introduced in 3GPP Releases 17 and 18, along with an account of ongoing work and anticipated developments for Release 19. Then it demonstrates how a virtualized architecture eases the transition to new functionality, presents an analysis of the software adaptations required to incorporate all Release 18 NTN features into a Release 17 software-based gNB implementation.

4.1 Network Virtualization Trend

The rapid increase in data traffic is basically changing the nature of things for mobile operators, especially with the shift towards 5G and beyond. This increase in traffic comes with some crucial challenges that operators must handle in order to maintain and enhance network performance. Of these, one is efficient resource utilization. With the demand for data dramatically increasing, operators must ensure that network resources are allocated and managed in a manner that allows them to cater to the enhanced load at maximum efficiency.

Another constraint that operators are going to deal with is reliable service. As the amount of data increases and networks become more complex, stronger mechanisms are needed to ensure continuity and reliability of transmission when peak loads are experienced or when there are disruptions to networks. Another critical challenge is latency reduction. Given the ever-increasing demand for real-time applications, it becomes very important to minimize the delays.

The reduction in power consumption driven by the push towards greener communications is also another critical aspect. Operators are under increasing pressure to support the global drive toward energy-efficient technologies and practices that lower the environmental impact of their networks. Besides these operational and technical challenges, cellular operators are grimly taking a look into the issue of Operational Expenditure and Capital Expenditure reduction. As the complexity of networks increases, so do the costs associated with site acquisition, network densification, and equipment installation and maintenance.

A very promising technique to deal with such challenges is represented by Network Function Virtualization. This is a technology, brought to evidence by the recent formation of the VNF Task Group by the European Telecommunications Standards Institute, which aims to virtualize cellular network functions, operations, coordination, and management. Network Function Virtualisation (NFV) manages this by decoupling network functions from physical hardware. Instead, VNF are deployed as software instances running on virtual machines. These VMs are then run on high-capacity, NFV-enabled hardware servers, switches, and routers, and are distributed over different geographical locations.

In addition to optimizing resource utilization, NFV significantly streamlines the upgradability of virtualized functions, allowing operators to swiftly integrate new features or adopt the latest 3GPP releases without the prolonged downtime typically seen in hardware-centric solutions.

Compared with the traditional hardware installations, network virtualization accelerates service deployment, hence gives operators more flexibility, scalability, and capacity [62]. Furthermore, NFV provides slicing of networks, which means that on top of one physical network, there could exist multiple virtual networks manageable independently. This feature helps an operator effectively utilize the available resources to provide services in order to meet special needs and achieve optimal network performance.

4.1.1 Softwarized NTN RAN and CN

Traditionally, RANs were built using proprietary hardware with tightly integrated functions, which served earlier generations of mobile networks well. However, as data demands have grown and applications have diversified, the limitations of traditional RAN architectures have become apparent. These challenges are even more pronounced in NTNs, where the dynamic nature of satellite and aerial platforms demands a more flexible and adaptive approach to RAN deployment, [63] [64].

With the advent of 5G, these pressures have driven changes in RAN design, leading to the emergence of Virtualized-RAN (vRAN). The fundamental concept of vRAN lies in the decoupling of network functions from the physical hardware on which they traditionally operated, enabling their deployment on general-purpose computing platforms, often cloud-based, [64]. This separation introduces significant advantages over the rigid hardware-specific architectures of earlier RAN systems

[65]. For NTNss, this decoupling is particularly advantageous, allowing network functions to be deployed closer to satellite gateways or edge computing nodes, enhancing performance and reducing latency.

The primary benefits of vRAN are scalability and manageability. Traditional RANs required significant investment in new hardware to scale for increased traffic or higher data rates, making expansion both costly and inflexible. In contrast, vRAN allows operators to scale efficiently by reallocating or adding virtual resources in the cloud, limiting the need for physical upgrades, [66] . This approach simplifies the deployment of new features and optimization algorithms, ensuring that the network remains agile and responsive to the evolving demands of modern applications. For NTNss, the ability to dynamically scale and optimize RAN functions based on satellite coverage areas or orbital conditions is crucial to maintaining seamless connectivity and high-quality service delivery.

The exponential increase in mobile network traffic, fueled by data-intensive services such as IoT and video streaming, presents a significant challenge for communication service providers, [67]. Effective traffic management now requires careful consideration of factors like demand, bandwidth, and quality of service. vRAN enables the flexibility and scalability to address these diverse needs. In the context of NTNss, vRAN can dynamically adapt to varying user densities across terrestrial and non-terrestrial coverage areas, ensuring efficient resource utilization and service continuity.

In addition, the transition from traditional RAN to vRAN marks a fundamental evolution in wireless network architecture. Beyond providing flexibility and scalability, vRAN significantly reduces costs, ensuring networks are equipped to meet the dynamic requirements of modern applications. Moreover, vRAN paves the way for constructing future networks capable of supporting the broad array of services and applications envisioned in the 5G ecosystem. For NTNs, vRAN plays a pivotal role in enabling the integration of satellite and terrestrial networks, offering a unified architecture that meets the stringent requirements of low latency, high throughput, and diverse application support critical for 5G and beyond.

With respect to the core network, it is undergoing a significant transformation in the 5G era, driven largely by the adoption of VNFs, [68]. Previously, core networks were implemented as dedicated hardware-based elements. However, with 5G, these functions are now software-driven and operate on virtual machines. This transition leverages the principles of software-defined networking and network function virtualization to create a more adaptable, scalable, and efficient architecture. Network virtualization enables VNFs to be deployed on standard servers, which are often integrated into Fog or Cloud Computing environments [69]. This softwarization of the core network allows dynamic management, allocation, and scaling of network functions to meet diverse service requirements, [68]. For instance, certain 5G services may demand low latency and high reliability, while others may prioritize storage capacity. Virtualized core networks adapt to these requirements by deploying

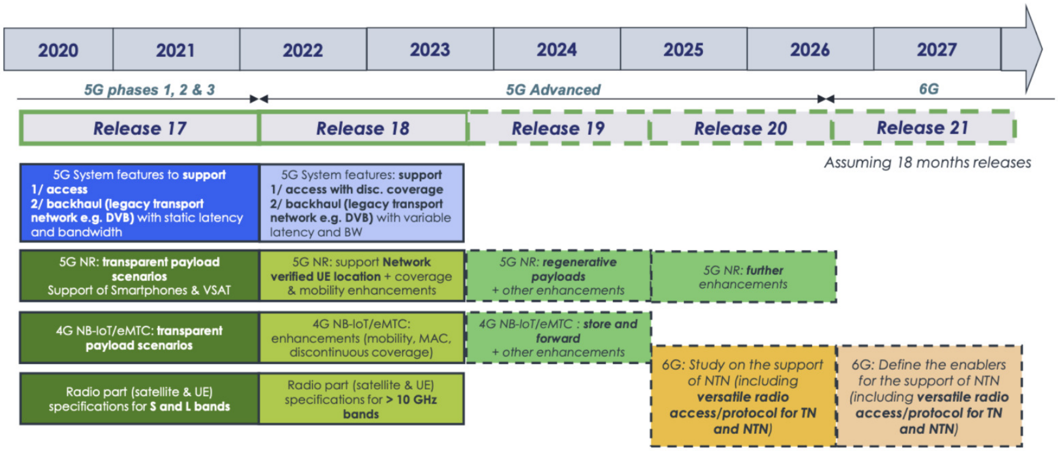

2020
2021
2022
2023
2024
2025
2026
2027
5G phases 1, 2 & 3
5G Advanced
6G
Release 17
Release 18
Release 19
Release 20
Release 21
Assuming 18 months releases
5G System features to support 1/ access 2/ backhaul (legacy transport network e.g. DVB) with static latency and bandwidth
5G System features: support 1/ access with disc. coverage 2/ backhaul (legacy transport network e.g. DVB) with variable latency and BW
5G NR: transparent payload scenarios Support of Smartphones & VSAT
5G NR: support Network verified UE location + coverage & mobility enhancements
5G NR: regenerative payloads + other enhancements
5G NR: further enhancements
4G NB-IoT/eMTC: transparent payload scenarios
4G NB-IoT/eMTC: enhancements (mobility, MAC, discontinuous coverage)
4G NB-IoT/eMTC : store and forward + other enhancements
Radio part (satellite & UE) specifications for S and L bands
Radio part (satellite & UE) specifications for > 10 GHz bands
6G: Study on the support of NTN (including versatile radio access/protocol for TN and NTN)
6G: Define the enablers for the support of NTN (including versatile radio access/protocol for TN and NTN)

FIGURE 4.1: NTN 3GPP roadmap [70].

VNFs in optimal locations—closer to the edge for low-latency services or in centralized cloud environments for resource-intensive applications. This virtualization approach not only enhances flexibility but also facilitates the creation of network slices: virtualized, isolated segments of the network tailored for specific applications. By customizing resources to meet the performance needs of various services, the core network supports a diverse range of use cases, fulfilling one of 5G's key promises.

4.2 The NTN path in 3GPP standardization

The 3GPP recognized the importance of NTN and included it in its Rel.17 to explore how the inclusion of NTN component enables planned 5G services and to push toward 5G-Advanced that will lead to Sixth-generation (6G) systems. Fig. 4.1 shows the roadmap of NTN standardisation, proposed in [70], covering up to Rel. 21.
The 3GPP started the work on NR over NTN in 2017 with two study items entitled i) "Study on NR support for Non-Terrestrial Networks", and ii) "Study on using the satellite access in 5G". The former, led by the RAN group with the support of RAN1 (*i.e.*, related to Layer 1) focused on the deployment scenario and adaptation of the 3GPP channel models for NTN, identifying the potential key impact areas on the NR and proposing solutions for the identified impacts on the RAN protocols and architectures. The outcomes of this study were documented in the Technical Report (TR) 38.811 [30]. Following this initial study, the 3GPP focused on the definition of the use cases for Satellite-based NR as part of the study item "Study on NR support for Non-Terrestrial Networks" under the supervision of SA working group with the support of the SA1 (Service). This study item, started in 2017 but finalized into Release 16 and associated with the Work Item (WI) on "Integration of Satellite Access in 5G", led to the definition of three main categories of use cases for satellite-based NTN:

- Service Continuity: the terrestrial network alone cannot provide 5G services, thus the 5G system shall support service continuity between 5G terrestrial access network and 5G satellite access networks. The NTN use cases mapped to the Terrestrial Network (TN)- NTN service continuity include stationary UE (enhanced Mobile BroadBand (eMBB)), Pedestrian UE (eMBB), Machine terminals (massive Machine Type Communications (mMTC)), stationary/ vehicular relay UE (eMBB), relay UEs on vehicles, ships, or high-speed train;

- Service Ubiquity: the aim of this use case is to extend the terrestrial network coverage into un-served or under-served geographical areas. Typical Examples of ubiquity use cases are enhanced Machine Type Communication (eMTC) (*e.g.*, agriculture, asset tracking, metering), public safety (*i.e.*, emergency networks), and home access;
- Service Scalability: this use case foresees the multicast or the broadcast of content to a large area by leveraging the large coverage area of satellites. An example of this use case is the distribution of rich TV content (*i.e.*, Ultra High-Definition TV).

After concluding Rel.15 on the scenario and channel models for NR to support NTN, 3GPP focused on both the system and architecture aspects and access technologies with a follow-up Rel. 16. With respect to the former, three activities were started within Rel. 16: i) a study item on "Study on architecture aspects for using satellite access in 5G" within the Working Group 2 (Architecture); ii) a WI on "Integration of Satellite Access in 5G", under the supervision of Working Group 1; and iii) a study item on "Study on management and orchestration aspects with integrated satellite components in a 5G network", within SA5 (Management). The SA1 identified the critical areas related to the integration of NTN in NR and provided solutions and requirements when considering the use cases described in TR 22.822 [71]:

- Roaming between terrestrial and satellite networks;
- 5G fixed backhauling between satellite-enabled NR-RAN and the 5G Core.

These requirements were added directly to the existing SA1 5G specification, TS 22.261 "Service Requirements for the 5G System" [72]. In addition, the SA5 addressed aspects related to management and orchestration for NTN, whose results are reported TR 28.808 [73].

The studies on the access technologies, guided by the RAN3 group (Interface), are captured in the study item "Study on solutions for NR to support NTN" completed at the end of 2019. These activities defined a baseline for NR functionalities aimed at supporting LEO and GEO satellites. Indeed, within this Release, a set of required adaptations enabling NR technologies and operations in the NTN context were identified, covering several issues in RAN1 (Physical layer), RAN2 (Layers 2 and 3), and RAN3. In particular, the performance assessment of NR over GEO and LEO satellites was provided at both system and link levels, together with a preliminary set of potential solutions for NR adaptations at Layers 2 and 3. Moreover, some architecture aspects were modified with respect to TR 38.811 [30], which is superseded from this point of view by TR 38.821 [22].

Based on the outcome of the Release 16 study, 3GPP initiated the following activities within Release 17: i) a WI named "Solutions for NR to support non-terrestrial networks" within the RAN2; ii) a study item named "Architecture Aspects for Using Satellite Access in 5G" under the supervision of SA2; iii) study item "Study on Public Land Mobile Network (PLMN) selection for satellite access" led by the Core Terminals (CT) working groups WG1, WG3, and WG4; and iv) study item on "Narrow-Band Internet of Things (NB-IoT) / eMTC support for NTN" under RAN1. The activities within the WI are devoted to the normative work and aim at specifying the necessary enhancements for LEO and GEO while also targeting support for HAPS and air-to-ground networks. This involves the physical layer aspects, protocols, and architecture as well as the radio resource management, Radio Frequency requirements, and frequency bands to be used. The WI is based on the following principles:

- Transparent payload architecture;
- Earth fixed tracking areas with Earth fixed and moving cells;
- Frequency Division Duplexing (FDD);
- The type of terminals supported are handheld devices in Frequency Range 1 (FR1), *i.e.*, below 7.125 GHz, and VSAT with external antenna operating in Frequencies band above 7.125 GHz (Frequency Range 2 (FR2)). The terminals are assumed to have GNSS capabilities.

The WI specified feature enhancements in RAN1, RAN2, RAN3, and RAN4. In particular, in the first two Working Groups, the objective is to address issues related to long propagation delays, large Doppler Shift effects, and moving cells in NTN. This implies enhancements on the PHY and MAC layers aspects, both on the User Plane (UP) and Control Plane (CP). The last two Working Groups focus on architectural enhancements, such as feeder link switchover and cell-related aspects, as well as aspects related to the UE RRM and RF requirements.

Building on these foundational studies, 3GPP has progressively refined and expanded the support for NTN across subsequent releases, aiming to address the challenges and opportunities presented by the integration of satellite and airborne systems into the 5G ecosystem. Each release introduced specific enhancements targeting key aspects of NTN deployment, performance, and capabilities.

In Release 17, the focus was on establishing the baseline framework for NTN integration. This involved defining the primary use cases, addressing unique challenges such as long propagation delays and Doppler shifts, and specifying adaptations for the physical and protocol layers. The release also explored frequency allocations and initial architectural solutions for both direct satellite access and backhaul scenarios.

Release 18 marked the transition to 5G-Advanced, where further enhancements were introduced to improve system robustness and versatility. These included support for a wider range of frequency bands, uplink performance optimizations, and advanced mobility management techniques to ensure service continuity between terrestrial and satellite components. Additionally, Release 18 addressed the integration of satellite backhaul and edge computing capabilities, enabling more dynamic and efficient network operations.

Looking ahead, Release 19 builds on these foundations to explore new frontiers for NTN. This includes expanding frequency allocations, supporting innovative use cases such as store-and-forward communications and direct UE-to-satellite-to-UE interactions, and introducing enhancements for regenerative payloads and positioning. These efforts aim to further align NTN capabilities with the evolving demands of next-generation networks.

The subsequent sections provide a detailed overview of the NTN features introduced in Releases 17, 18, and 19, highlighting the specific enhancements and their implications for 5G and beyond.

4.2.1 Release 17

Rel. 17 provides the first normative definition of the new or enhanced features of the 5G System (5GS) to support a NTN component, [14]. The targeted use cases include coverage extension,IoT via satellite, disaster communications, global roaming, and broadcasting, for which the target NTN KPIs have been introduced in TS 22.261,[74]. Severalstudy item and WIs carried out in the RAN, SA, and CT Technical Specification Groups (TSG) to study and select the solutions capable to cope with satellite-specific key issues. In this Section, we review those defined in Rel. 17, which constitute the baseline for the NTN integration in 3GPP-based systems. The motivation for this description is that, in order to properly introduce and understand the enhancements of Rel. 18, the baseline NTN system in Rel. 17 shall be detailed. Please note that below we report only the enhancements or new features specifically introduced for NTN; references to the actual procedures are provided within the text and will be detailed during the Study. In this context, it shall be mentioned that the descriptions and definitions of the Layer 2 measurements to be performed by the network or the User Equipment, and transferred over the standardised NR interfaces, to support NR radio link operations, RRM, network Operation and Maintenance (O&M), Minimisation of Drive Tests (MDT), and self-Organising Networks (SON) are provided in TS 38.314, [75]; in addition, the detailed MAC layer procedures, channels, formats, and parameters are detailed in TS 38.321, [76].

4.2.1.1 Payload, User Equipment, and Frequency Allocations

In Rel. 17, only FR1 systems are defined for NTN. As such, only handheld terminals operating in L or S band can be supported transmitting at most 23 dBm (with

TABLE 4.1: Frequency allocations and duplexing in Rel. 17 NTN.

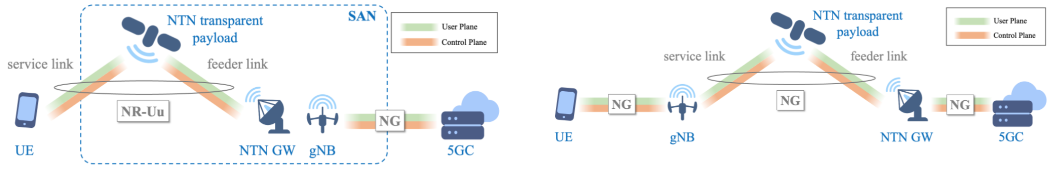
Band	UL (UE-to-SAN)	DL (SAN-to-UE)	Duplexing
n.256	1980-2010 MHz	2170-2200 MHz	FDD
n.255	1626.5-1660.5 MHz	1525-1559 MHz	FDD
n.1	1920-1980 MHz	2110-2170 MHz	FDD HAPS

±2 dB tolerance), in the frequency allocations indicated in Table 4.1 with the terminal characteristics defined in TR 38.821 Clause 6.1.1, [77], and TS 38.101-5, [78]. The maximum transmission bandwidth configurations per UE for each UE channel bandwidth and Sub-Carrier Spacing (SCS) are provided in TS 38.101-5, [78]: i) the allowed channels bandwidth configurations are 5 MHz, 10 MHz, 15 MHz, and 20 MHz; ii) the allowed SCSs are 15 kHz (but not for 5 MHz channels), 30 kHz, and 60 kHz.

With respect to the RF performance and RRM requirements, the following specifications apply: UE in TS 38.101-5, [78], and TS 38.108, [79], Satellite Access Node (SAN) in TS 38.108, High Altitude Platform Station (HAPS) SAN in TS 38.104, [80]. In addition, TS 38.133, [81], provides specific requirements for radio resource management in NTN.

4.2.1.2 SA and CT

Within the FS_5GSAT_ARCH study item, SA2 defined the architectures to provide satellite access in 5G, as well as the impact that such integration might bring to the 5GS and the required solutions, as reported in TR 23.737, [82]. Based on the outcomes of this study item, the 5GSAT_ARCH WI updated the architecture specifications in TS 23.501, [83], TS 23.502, [84], and TS 23.503, [85]. As for the CT activities, CT1 studied the impact of the selection procedure for Public Land Mobile Networks (PLMNs) in 5GSAT_ARCH_CT, leading to modifications reported in TS 23.122, [86], and TS 24.501, [87]. A satellite NG-RAN can be shared among multiple 5G PLMN Core Networks (5GCs) for two access types: direct access (i.e., the satellite NG-RAN is a new 3GPP access node) and backhaul connectivity of Terrestrial Network (TN)s to the 5GC, as represented in Figure 4.2. Transparent payloads are assumed for: i) spaceborne platforms, i.e., GSO orNGSO atLEO (300-1500 km), Medium Earth Orbit (MEO) (7000-25000 km), or GEO (35786 km); and ii) airborne platforms, i.e., HAPS, which include Lighter Than Air (LTA) and Heavier Than Air (HTA) Unmanned Aerial Systems (UAS) operating in quasi-stationary mode between 8 and 50 km. The NTN payload transparently forwards the radio protocol from/to the UE (via the service link) to/from the NTN Gateway (via the feeder link) based on the NR-Uu Air Interface. It shall be noticed that a single gNB may serve multiple NTN payloads and an NTN payload may be served by multiple gNBs. In this framework, it shall be noticed that Multi-Connectivity (MC) and Ultra Reliable and Low Latency



(a) NTN component for direct access

(a) NTN component for backhaul connectivity

FIGURE 4.2: NTN access solutions as per 3GPP Rel. 17

Communications (URLLC) are not considered in Rel. 17. For the UE, it is mandatory to be equipped with GNSS capabilities, motivated by the enhancements introduced at RAN level discussed in Section 4.2.1.3.

In direct access solutions, the coverage can be provided by means of:

- Earth-fixed beams: the beams continuously cover the same geographical area all the time (e.g., GSO).
- Quasi-Earth-fixed beams: the beams cover a geographical area for a limited period and a different area in the next limited period (e.g., NGSO satellites generating steerable beams).
- Earth-moving beams: the beams cover a fixed area with respect to the satellite, i.e., they are moving on the surface of the Earth along the satellite's movement on its orbit (e.g., NGSO with non-steerable beams).

These architectures produce significant impacts on several procedures and operations of the 5GS, such as mobility management, synchronisation, signalling, and O&M, due to the large size of the beam footprints, the potentially moving beams on the surface of the Earth, the increased over-the-air latency, and the Doppler shift. In this context, it is worthwhile highlighting that, for NTN operation, the Tracking Areas (TAs) and the Cell Identities (Cell IDs) refer to specific geographical areas (i.e., they are Earth-stationary) and, thus, the 5GS can exploit this information to represent the UE location. For both direct access and backhaul connectivity, new 5G Quality of Service Identifiers (5QIs) have been introduced in TR 23.501, as the QoS is impacted by the larger delays over NTN; an example if 5QI n. 10, with 1100 ms of packet delay budget, for Video (buffered streaming), TPC-based, and any other service that can be used over satellite with these characteristics, [83].

The satellite backhaul only assumes constant delays, which means that the Satellite Network Operator (SNO) shall mask any variations happening on the feeder/service links by exploiting the satellite ephemeris to pre-compute the variable delay to be tackled. In addition, the satellite backhaul category (referring to GEO, MEO, LEO, or others) is reported to the 5GC; in particular, in case the Access and Mobility Function (AMF) is aware that a satellite backhaul with large delays is being used, this information will be reported to the Session Management Function (SMF) within the PDU Session establishment procedure, detailed in TS 23.502, [84]. In case the backhaul category changes, the AMF will report the change to the SMF. For the UE, it shall be mentioned that the impacts at Non-Access Stratum (NAS) layer have been specified in CT1, including, e.g., a new NG-RAN satellite RAT type in the USIM, the extension of NAS supervision timers over satellite access (only for GEO and MEO, LEOs can exploit the legacy ones), new triggers for the PLMN selections upon the transition in or out of international areas, the support for multiple TACs for the same PLMN broadcast in the radio cell, etc., [86]-[87].

4.2.1.3 RAN

Within RAN, several NTN-specific aspects have been considered to identify the required enhancements and extensively discussed in TR 38.811, [30], and TR 38.821, [77]:

- Delay variation and Doppler variation, as well as the impact of potentially moving radio cells on the surface of the Earth, due to the space-/air-borne platform motion.
- Large over-the-air latencies, due to the altitude of the space-/air-borne platforms.
- Differential delays and possible multi-country cell coverage, due to large radio cell footprints on-ground.
- Satellite channel model characteristics, detailed in [30].
- Different radio unit performance due to specific space-/air-borne payload performance

Among the new features introduced for the support of NTN, a new System Information Block (SIB), SIB19, has been introduced. Its location is reported in the SIB1, the transmission of which is mandatory together with the Master Information Block (MIB) for all 5G systems. The MIB is transmitted on the Broadcast Channel (BCH) with a periodicity of 80 ms (or larger, if the Synchronisation Signal Block, Synchronisation Signal Block (SSB), is scheduled with larger periodicities) and it contains information to acquire the SIB1. SIB1 and SIB19 are transmitted on the Downlink Shared Channel (DL-SCH) and SIB1 provides the information to acquire other SIBs, including SIB19. The actions to be performed for the acquisition of the System Information, and the subsequent steps based on the specific SIB, are detailed in TS 38.331, [88], providing clear guidelines depending on the current UE status (RRC_IDLE, RRC_CONNECTED, RRC_INACTIVE). It shall be mentioned that, during the study phase in Rel. 15 and 16, it was established that no modifications are needed to the SSB for NTN support. The content of SIB19 is provided in TR 38.331, [88], and most of its parameters will be introduced in the next sections when discussing the enhancements introduced for NTN support. It is worth mentioning that a UE willing to exploit NTN connectivity shall always maintain a valid and updated copy of the

SIB19 information. In case the timer indicating its validity (T430 in TR 38.331) expires, the upper layers shall be informed that the UE lost synchronisation, and it shall start a new acquisition procedure.

TIMING, SYNCHRONISATION, AND HARQ ENHANCEMENTS

In Rel. 17, only UEs equipped with GNSS capabilities can operate. More specifically, the field uplinkPreCompensation-r17 indicates if the UE supports the uplink time and frequency pre-compensation and timing relationships enhancements, which TS 38.306 indicates to be mandatory. With respect to frequency synchronisation, the UE shall autonomously pre-compensate the Doppler shift. In fact, during the study phase, it was established that the downlink frequency pre-compensation on the service link is not performed at the gNB and it was left for implementation at the UE; in addition, any frequency compensation on the feeder link or management of transponder frequency errors is also operated at the gateway and left for system implementation. The first procedure to be implemented at the UE is the uplink synchronisation, in which the TA shall be autonomously pre-compensated by exploiting the common TA provided by the gNB, the satellite ephemeris, and the UE location information. This computation needs some details. The TA parameter is computed as:

$$T_{TA} = T_C \left(N_{TA} + N_{TA,offset} + N_{TA,adj}^{common} + N_{TA,adj}^{UE} \right)$$

where the term
$(N_{TA} + N_{TA,offset})$ comes from legacy NR and the term $\left(N_{TA,adj}^{common} + N_{TA,adj}^{UE} \right)$ has been introduced for NTN. $N_{TA,adj}^{common}$ includes any common delay computed at the network side (e.g., two-way feeder link) and $N_{TA,adj}^{UE}$ is the UE adjustment based on the ephemeris and GNSS information (two-way service link). The common TA corresponds to the Round Trip Time (RTT) between the NTN payload and the uplink timing synchronisation Reference Point (UTSRP)[1]. When the UTSRP is not at the NTN payload, the UE needs to compensate the feeder link RTT and any timing offset considered necessary by the network. Hence, the timing drift on the feeder link shall be compensated by the UE with sufficient accuracy. $N_{TA,adj}^{common}$ is derived from the higher-layer parameters ta-Common (TA_{Common}), ta-CommonDrift ($TA_{CommonDrift}$), and ta-CommonDriftVariation ($TA_{CommonDriftVariation}$), [89], if configured, as follows:

$$\Delta T_{common}(t) = \frac{TA_{Common}}{2} + \frac{TA_{CommonDrift}}{2}(t - t_{epoch}) + \frac{TA_{CommonDriftVariation}}{2}(t - t_{epoch})^2$$

[1] The UTSRP is defined as the point at which DL and UL are frame-aligned with an offset given by $N_{TA,offset}$. It can be located at the gNB, on-board the NTN platform, or at the gateway. If it is on the satellite, the parameters discussed above do not need to be provided, as the common delay is null.

where t_{epoch} is the epoch time of the parameters. If the required parameters are not available, $N_{TA,adj}^{common}$ it is set to 0. The TA applied by the UE should compensate the delay in the UL signal path from the UE to the UTSRP at the time of transmission of a given UL slot and the delay in the DL signal path from the UTSRP to the UE of the corresponding DL slot. Thus, for a given uplink slot n, the common TA is given by, [90]:

$$TA_{Common} = \Delta T_{common}(t_{UL,n}) + \Delta T_{common}(t_{DL,n})$$

where $t_{UL,n}$ and $t_{DL,n}$ denote the time of transmission of slot n on the uplink and the corresponding downlink timing, respectively. In case the UE has not a valid GNSS position and/or a valid SIB19 (for the satellite ephemeris), it is not allowed to communicate with the network until both are re-acquired; the validity is configured with a single value for both, and it can range between 5 seconds and 900 seconds (e.g., for GEO access). The UE can be configured to report the TA during the initial access procedure or in RRC_CONNECTED state; in the latter case, a combination of both open-loop (i.e., UE autonomous TA estimation and common TA estimation) and closed-loop (i.e., received TA commands in RAR message or MAC CE) is supported.

To accommodate the larger propagation delays over NTN, many timing relationships (e.g., timers extension in MAC/RLC/PDCP layers and RACH adaptation) have been extended by the common TA and two scheduling offsets: K_{offset}, a scheduling offset that shall be larger than or equal to the sum of the service link RTT and the common TA, and k_{mac}, an offset approximately corresponding to the RTT between the UTSRP and the gNB, [91]. K_{offset} supports the extension of the following timing relationships

- The transmission timing of RAR grant or fallbackRAR grant scheduled PUSCH.
- The transmission timing of PDCCH ordered PRACH.
- The transmission timing of DCI scheduled PUSCH, including CSI transmission on PUSCH.
- The timing of the first PUSCH transmission opportunity in type-2 configured grant.
- The transmission timing of HARQ-ACK on PUCCH, including HARQ-ACK on PUCCH to message B (MsgB) in two-step random access.
- The timing of the adjustment of uplink transmission timing upon reception of a corresponding TA command.
- The transmission timing of aperiodic SRS.
- The CSI reference resource timing.

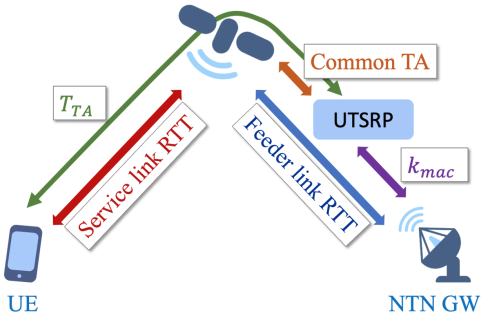


FIGURE 4.3: Timing relationships in Rel. 17

For this parameter, 3GPP also defined cell-specific and UE-specific values, as detailed in TS 38.321, [76]. As for k_{mac}, it has been introduced to extend the Random Access Response (RAR) window and the msg3 Response Window in the Random Access (RA) procedure. This parameter is provided by the SIB19 and it allows to estimate the UE-gNB RTT as $T_{TA} + k_{mac}$ (it equals the offset of the gNB's downlink and uplink frame timing), [76]-[88]. The timing relationships described above are shown in Figure 4.3.

Finally, to limit the impact of Hybrid Automatic Repeat request (HARQ) signalling, in NTN it is possible to either disable the HARQ feedback per HARQ process in the presence of ARQ transmissions at the RLC layer (e.g., in GEO systems) and/or to increase up to 32 the number of HARQ processes for retransmissions in the MAC layer, [88]. Due to the large delays in NTN, the number of HARQ processes might even expire before the first feedback is received; thus, to avoid HARQ stalling and degrade the peak throughput, their number can be increased. For instance, assuming a LEO satellite at 600 km (maximum RTT up to 25.77 ms), 16 HARQ processes would lead to a peak throughput reduction of approximately 43%, [92].

MOBILITY MANAGEMENT

Mobility within the NTN component or to/from terrestrial networks follows the same procedures and principles of legacy 5G networks, with some adjustments. In general, during the NTN-TN mobility, a UE is not required to be connected to both networks at the same time; in addition, a UE might be able to support mobility between satellite RATs on different orbits. The features introduced for NTN mobility are discussed below depending on the UE state: RRC_IDLE/RRC_INACTIVE and RRC_CONNECTED.

RRC_IDLE/RRC_ INACTIVE: in both states, the UE-controlled mobility is based on the network configuration and it is performed through cell selection and cell reselection procedures. This applies all of the service link types (Earth-fixed, Quasi-Earth fixed, and Earth-moving). The cell selection is used to identify the most appropriate cell for the UE to camp on and it is applicable: i) after the terminal has been switched on; ii) after the UE leaves the RRC_CONNECTED state; and iii) after the

UE is back into a coverage area. Cell selection is based on the following principles detailed in TS 38.304, [93]:

- The UE NAS layer identifies a selected PLMN(s) and it searches the supported frequency bands for each carrier frequency.
- The strongest cell is identified, and the UE reads the cell broadcast information to identify the PLMN(s) and other relevant parameters, such as cell restrictions.
- The UE searches a suitable cell and, if none can be identified, it looks for an acceptable cell.
  - suitable if: i) the measured cell attributes meet the cell selection criteria, based on the downlink radio signal strength/quality; ii) the cell belongs to the selected PLMN(s); and iii) the cell is not restricted, i.e., it is not barred or reserved or part of forbidden roaming areas;
  - acceptable: the measured cell attributes satisfy the cell selection criteria and the cell is not barred.
- Among the identifies suitable/acceptable cells, the UE selects the strongest one and it camps on that.
  - Please note that, if signalled/configured by the radio network, some frequencies might be prioritised for the UE camping.

The cell selection criteria (also known as "S" criteria, based on the parameters Srxlev and Squal) are detailed in TS 38.304, [93]. The cell re-selection procedure is based on the following principles:

- The procedure is always based on Cell Defining SSBs (CD-SSBs) located on the synchronization raster.
- The UE measures the attributes of the serving and neighbour cells to enable the re-selection process. For the search and measurement of inter-frequency neighbouring cells, only the carrier frequencies shall be indicated.
- The cell re-selection identifies the cell on which the UE should camp. Such identification is based on cell re-selection criteria, involving measurements of the serving and neighbouring cells.
  - intra-frequency re-selection is based on the ranking of cells;
  - inter-frequency re-selection is based on the absolute priorities where a UE tries to camp on the highest priority frequency available.
- It shall be noticed that:

  - a neighbour cell list can be provided by the serving cell to handle specific cases for intra-/inter-frequency neighbouring cells;
  - blacklists can be provided to prevent the UE from reselecting to specific intra-/inter-frequency neighbouring cells;
  - the cell re-selection can be speed dependent;
  - service-specific prioritization is implemented. The rules for NR inter-frequency and inter-RAT frequencies re-selection are provided in the system information and detailed in TS 38.133, [81].

In the RRC_IDLE state, the UE can autonomously perform the cell re-selection procedure without informing the network, as long as it remains within the registered Tracking Area (which corresponds to a fixed geographical area and the mapping is broadcasted by the network, as discussed above). The UE acquires the SIB1 after each re-selection to determine whether this is the case or not. In case the TA is changed, and it is outside the UE registration area, a mobility registration update procedure shall be performed. It shall be noticed that relying on Srxlev or Squal thresholds to trigger the UE measurements in NTN systems might not be sufficient. This is motivated by the fact that, unlike terrestrial systems, there is a small difference in the received signal strength between two beams in an overlap region. Therefore, new time-based and location-based rules have been introduced to limit the needed cell re-selection measurements, i.e., to let the UE decide when to perform the measurements:

- Location-based measurement: when Srxlev and Squal are above the thresholds, in case the distance between the UE and the serving cell reference location is below a threshold (distanceThresh[2]), the UE may not perform the measures for cell re-selection. The distance threshold and the reference location are broadcasted in SIB19 and are defined in TS 38.304, [93]; moreover, the UE shall support location-based measurement initiation, as described in TS 38.306, [94]. The guarantee that a valid location information is obtained by the UE is up to its implementation.

- Time-based measurement: this option is a new rule introduced for cell re-selection within an NTN cell with Quasi-Earth fixed beams, i.e., with beams covering a geographical area for a limited time. It is based on the expected time in which the cell will remain active as indicated in SIB19 (t-Service[3]), which indicates the time information related to when the cell is going to stop serving the current area. As per TS 38.304, the UE shall perform intra-frequency, inter-frequency, or inter-RAT measurements before t-Service, independently from the distance to the serving point or the values of Srxlev and Squal. These measurements can be prioritised as detailed in TS 38.133, [81]. This procedure is optional for the UE, as specified in TS 30.306, [94].

[2] Up to 65525 multiples of 50 meters as indicated in the SIB19, [88].

[3] Its value can be between 0 and 549755813887 multiples of 10 ms, [88].

TABLE 4.2: Measurement events in RRC_CONNECTED

Meas. event code	Description
A1	The serving cell measurement becomes better than a given threshold
A2	The serving cell measurement becomes worse than a given threshold
A3	The neighbouring cell measurement becomes <offset>better than the serving cell
A4	The neighbouring cell measurement becomes better than given threshold
A5	The serving cell measurement becomes worse than a given threshold and the neighbouring cell becomes better than another threshold
A6	The neighbouring cell measurement becomes <offset>better than a secondary cell
B1	The inter-RAT neighbouring cell becomes better than a given threshold
B2	The primary cell becomes worse than a given threshold and the inter-RAT neighbour becomes better than another threshold
D1	The distance between the UE and a reference location (referenceLocation1) becomes larger than a threshold (distanceThreshFromReference1) and the distance between the UE and a second reference location (referenceLocation2) becomes smaller than another configured threshold (distanceThreshFromReference2)

RRC_CONNECTED: the handover procedure refers to the transfer of an active UE connection from one radio channel to another. Network-controlled handover is implemented in 5G NR for users in RRC_CONNECTED state, and it can be categorised as follows: i) cell-level, in which explicit RRC layer signalling is triggered (inter-cell handover, handover request, handover acknowledgment, handover command, and handover complete procedure are supported between the source and the target cells, with the target gNB releasing the source gNB resources); and ii) beam-level, which does not require RRC signalling and is managed at lower layers (the RRC is not required to know which beam is being used at a given time). The network can configure the UE to perform measurements based on several events (see Table 4.2) and report them based on the measurement configuration.

It shall be mentioned that:

- The measurement reports include the measurement identify of the associated configuration that triggered the reporting.
- Cell and beam level measurements to be included in the reports are configured by the network.
- The number of non-serving cells to be measured and reported can be limited by the network.
- Beam measurements are configured by the network: beam identifier only, beam identifier and measurement result, or no beam reporting.

It is worthwhile mentioning that the RRM event D1 is a new measurement triggering events based on location. It is introduced to trigger location reporting based on the UE location in the NTN system. For this measurement event, the UE is configured with two reference locations and two threshold distances. In Rel. 15, an SSB-based Measurement Timing Configuration (SMTC) window was introduced. The SMTC configuration provides the UE with the measurement periodicity and timings of the SSBs that can be used for measurements during its window. As such, the UE is not allowed to search for the SSB outside this window, which has a periodicity in the same range of the SSB: 5, 10, 20, 40, 80, or 160 ms, with a duration of 1, 2, 3, 4, or 5 ms depending on the number of SSBs. The larger propagation delays in NTN might impact RRM measurements: if the SMTC measurement configuration does not consider this, the UE might miss the measurement window and it will be unable to perform the measurements on the configured reference signals. To deal with this, the SMTC was enhanced in Rel. 17 by allowing the network to configure multiple SMTCs in parallel per carrier and for a given set of cells depending on the UE capabilities, [77]. Finally, an enhancement was also introduced for Conditional Handover (CHO), i.e., a handover to be performed by the UE when one or more conditions (set by the serving gNB) are met. In terrestrial networks, up to two conditions can exist for CHO: A3 and A5 measurement-based events (see Table 4.2). In NTN, exploiting only the signal strength might not be sufficient, in particular with large NTN cells. Moreover, a very large number of UEs might need to perform the handover at a given time, leading to an impossible signalling overhead and service continuity challenges. Thus, in NTN, new time-based and location-based triggering conditions have been introduced, as well as event A4. The gNB can still configure up to two trigger conditions per target cell and one of these events shall be based on the signal strength/quality, [95]. Finally, it shall be mentioned that, upon network request and after the Access Stratum (AS) security is established for RRC_CONNECTED UEs, the terminals shall report their coarse location (most significant bits, guaranteeing an accuracy in the order of 2 km) to the NG-RAN, if available.

FEEDER LINK SWITCH-OVER

The change of the feeder link from a source NTN gateway to a target NTN gateway is a Transport Network Layer procedure, and it can be either soft or hard. In soft switch-over (Figure 4.4), the NTN payload temporarily connects to more than one gateway (i.e., during the feeder link transition), while in the hard procedure (Figure 4.5) the NTN payload can only connect to one gateway at a time and radio link interruptions might occur.

The moment in which a feeder link switch-over shall be performed is determined by the NTN Control function[4]. Depending on the specific gNBs' implementation and their configuration defined by the NTN Control function, the transfer of the

[4] The NTN Control function controls the space-/air-borne platforms and the radio resources of the NTN infrastructure; moreover, it provides control data (e.g., ephemeris) to the non-NTN infrastructure gNB functions, as described in TS 38.300 [91].

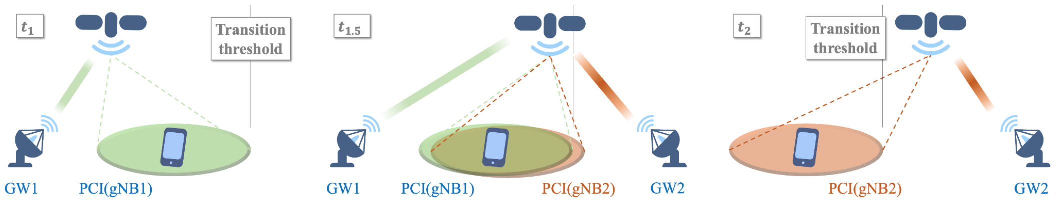


FIGURE 4.4: Soft feeder link switch-over in NTN

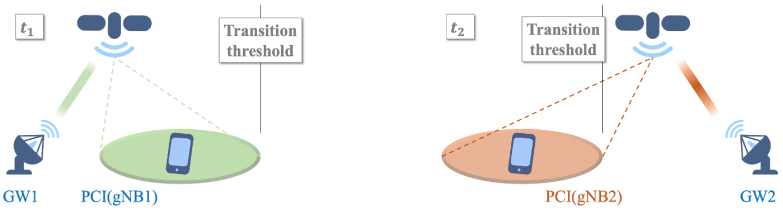


FIGURE 4.5: Hard feeder link switch-over in NTN

UEs' context between the two gNBs during the feeder link switch-over can be either NG-based or Xn-based, [91]; it shall be noticed that these handovers exploit the Cell ID (for which the enhancements are discussed below). In Quasi-Earth fixed scenarios, the NTN Control function shall provide the time window of successive switch-overs (both feeder and service), and the identifier and the time window of all serving platforms and NTN gateways; in Earth-moving scenarios, it shall provide the schedule of successive serving NTN gateways/gNBs and that of successive feeder/service link switch-overs.

NG-RAN SIGNALLING

The following assumptions and enhancements hold for NTN with respect to signalling in the NG-RAN. The gNB reports a Cell-ID to the 5GC in the UE Location Information (ULI). This Cell-ID is always the Mapped-Cell-ID, regardless of the NTN RAT orbit or the service-link types in use. This identifier is exploited for paging optimisation, area of interest, and public warning services, [91]. The mapping between the Mapped Cell ID and the geographical areas is configured in both the RAN and 5GC: the gNB is in charge of constructing the Mapped ID based on the ULI, if available. Please note that the mapping can be pre-configured (based on the operator's policy) or left up to implementation. The gNB reports the broadcasted TACs of the PLMN(s) to the AMF as part of the ULI. The Cell ID is defined in TS 38.413, [95], and TS 38.424, [96].

AMF (RE-)SELECTION

The selection of the appropriate NAS node is implemented at the gNB as specified in TS 38.410, [25]. When the gNB is configured to guarantee that a UE in RRC_CONNECTED state is connected to an AMF serving the country in which it is located, it might initiate a NG handover to change the serving AMF or initiate a UE context release towards the serving AMF, if the UE moves to a different country to that of the serving AMF.

OPERATION AND MAINTENANCE (O&M) REQUIREMENTS

All NTN related parameters provided in TS 38.300 clause 16.14.7 (ephemeris and their epoch, location of the NTN gateways, and any additional parameter enabling the gNB operation for service or feeder link switch-over) shall be provided by the O&M to the gNB providing NTN access. In addition, the NTN-related parameters in Annex B4 of TS 38.300 can be provided by the O&M to the serving gNB for its operation.

4.2.2 Release 18

Rel. 18 contains the first set of specifications dedicated to 5G-Advanced, [97]. In addition, in this framework, 3GPP also submitted the NTN Radio Interface Technology (RIT) to ITU-R for its inclusion in IMT-2020. As reported in TR 37.911, , two submissions were completed:

- A Set of Radio Interface Technologies (SRIT) with 2 components: NR NTN (for NR satellite access) and IoT NTN for (NB-IoT/eMTC satellite access).
- A RIT on NR NTN (for NR satellite access).

Below, we discuss the enhancements and/or new features introduced for NTN systems.

4.2.2.1 Payload, User Equipment, and Frequency Allocations

Rel. 18 introduced FR2 (WI on “NR NTN enhancements” (NR_NTN_enh)) and extended FR1 (WI on "Introduction of the satellite L-/S-band for NR" (NR_NTN_LSband)) for both GSO and NGSO systems, with the allocations reported in Table 3. With respect to the UEs, the following enhancements were introduced:

- The configuration of handheld terminals as per Rel. 17 still hold. In addition: i) an increased uplink performance is allowed through commercial smartphones with 5.5 dBi antenna gain and 3 dB polarisation loss per antenna port; ii) the maximum transmission bandwidth configuration can be as large as 30 MHz (for alignment with ITU-R evaluations).
- Very Small Aperture Terminal (VSAT) UEs have been introduced, as reported in Table 4 (please note that all classes assume that the UE can generate one beam pointing at one satellite in a given time instant). For these: i) the allowed SCSs are 60 kHz and 120 kHz; ii) the maximum transmission bandwidth configuration for each UE channel bandwidth and SCS are 50 MHz, 100 MHz, 200

TABLE 4.3: Frequency allocations and duplexing introduced in Rel. 18 NTN, as per TS 38.101-5, [3]

Band	UL (UE-to-SAN)	DL (SAN-to-UE)	Duplexing	Applicability
n.512	27.5-30.0 GHz	17.3-20.2 GHz	FDD	
n.511	28.35-30.00 GHz	17.3-20.2 GHz	FDD	CEPT ECC Decision(05)01 and ECC Decision (13)01
n.510	27.50-28.35 GHz	17.3-20.2 GHz	FDD	US subject to FCC 47 CFR part 25
n.254	1610-1626.5 MHz	2483.5-2500 MHz	FDD	Earth Station operations in the US subject to FCC 47 CFR part 25. No ESIM currently.

TABLE 4.4: New UE classes for NTN in Rel. 18

UE class	UE type	Description
Fixed VSAT	1	GSO and LEO with mechanical steering antenna
	2	GSO and LEO with electronical steering antenna
	3	LEO with electronical steering antenna
Mobile VSAT	4	GSO with mechanical steering antenna
	5	GSO with electronical steering antenna

MHz, and 400 MHz (not available for 60 kHz SCS); iii) for fixed VSATs, the maximum transmission power is 35 dBm (type 1) and 43 dBm (types 2 and 3); iv) for mobile VSATs the maximum transmission power is 35 dBm (type 4) and 43 dBm (type 5).

With respect to the UE types, it shall be noticed that, in Rel. 18, mobile VSAT terminals are only allowed for GSO systems, while fixed VSAT receivers can connect to both GSO and LEO satellites.

4.2.2.2 SA and CT

Within SA and CT, two important WIs were finalised in Rel. 18: i) "5G system with satellite backhaul" (5GSATB), focusing on the support of dynamic backhaul and User Plane Function (UPF) on-board GEO platforms; and ii) “(Stage 2 of 5GSAT_Ph2) 5GC/EPC enhancement for satellite access Phase 2” (5GSAT_Ph2), focusing on the support for discontinuous coverage. TS 22.261 reports the requirements for satellite access. With respect to the scenarios defined in Rel. 17, it is worthwhile mentioning that Rel. 18 added the video surveillance use case; this scenario refers to video surveillance data transmitted in uplink from an on-ground UE using satellite access, as well as the video surveillance-related configuration or control data that might be required. In addition, it shall be mentioned that TS 22.261 Clause 6.7.2 states that the case in which the backhaul connection presents dynamic changes of latency and/or bandwidth shall be considered for the 5GS, in particular to support proper mechanisms to determine the suitable QoS parameters for the traffic. This is in line with the WI on satellite backhaul, discussed below.

5G SYSTEM WITH SATELLITE BACKHAUL

A study item (FS_5GSATB) was conducted and the outcomes reported in TR 23.700-27, [98], identifying three key issues: “PCC/QoS control enhancement considering dynamic satellite backhaul” (related to the support of dynamic backhaul, e.g., issues related to multi-hop ISLs, variable delivery latency, or packets out-of-sequence), and “Support of Satellite Edge Computing via UPF on board” (to reduce latency and minimise the resource request on the backhaul) and “Support of Local Data Switching via UPF on-board” (to reduce the end-to-end delay for the communicating UEs). Based on the study, the WI 5GSATB was completed, in which SA1 added requirements for QoS control and charging when using satellites as transport/backhaul in 5G system to TS 22.261 and SA2 implemented modifications to TS 23.501 (architecture), [83], TS 23.502 (procedures), [84], and TS 23.503 (policy), [85]. The stage 3 aspects were then addressed within the CT3/CT4 WI “CTx aspects of 5GSATB (Satellite Backhauling in 5GS),” which modified TS 29.502, [99], TS 29.508, [100], TS 29.512, [101], TS 29.514, [102], TS 29.518, [103], TS 29.523, [104], TS 29.525, [105], and TS 29.571, [106], accordingly. The solutions for the three key issues are discussed below. Please note that the corresponding enhancements in terms of management are reported in TS 28.541, [107], and TS 28.552, [108].

Policy and Charging Control (PCC) and QoS control enhancement considering dynamic satellite backhaul: Notably, satellite backhauling might lead to variable delays and/or bandwidth when multiple ISLs or different platforms are involved. A change in the satellite backhaul has an impact on the PCC and QoS control, for which two enhancements have been introduced in Rel. 18

- The satellite backhaul categories introduced in Rel. 17 (see Section 1.1.2) have been extended with DYNAMIC_GEO, DYNAMIC_MEO, DYNAMIC_LEO, and DYNAMIC_OTHERSAT; these indicate to the involved network elements that the backhaul capabilities might vary over time. In case the AMF becomes aware that the category changes (e.g., due to handover), it reports the current category and indicates the change to the SMF. As per TS 23.501, the AMF is assumed to be capable of determining the backhaul category based on local configuration, e.g., based on Global RAN Node IDs associated with satellite backhaul.
- If a dynamic satellite backhaul category is indicated, the SMF or Policy Control (PCF) can decide to enforce the QoS monitoring to measure the packet delay between the UE and PDU Session Anchor (PSA) UPF (which is the UPF terminating the N6 interface of a PDU session within the 5GC). Such monitoring allows to measure the packet delay between the UE and the PSA UPF as a combination of the RAN part of the UL/DL packet delay (see TS 38.314 Clause 4.2.1.2, [75]) and the UL/DL packet delay between NG-RAN and PSA UPF. The PCF can calculate the Packet Delay Variation and the round-trip packet

delay when UL and DL are on different QoS flows, as specified in TS 23.501. The PCF handling of the backhaul category indication and the possible QoS monitoring are reported in TS 23.503.

Support of Local Data Switching via UPF on-board: to avoid or reduce the use of satellite backhauls, which might introduce long packet delivery latencies or limited bandwidth, a UPF on-board might be exploited. However, NGSO platforms would call for the management of UPF mobility, which might significantly impact the current 5GC procedures and architectures. As such, Rel. 18 focused on the enhancements to support the deployment of UPFs on-board GEO satellites to locally route the UE-to-UE traffic without going down to an on-ground gateway; moreover, it is worthwhile highlighting that the enhancements introduced in Rel. 18 assume Data Network Names (DNNs) and slices for 5G Virtual Networks (VNs). Two local switching types have been introduced, for which the details can be found in TS 23.501, [83]: PSA UPF on-board and UL CL/BP with local PSA UPF on-board.

Support of Satellite Edge Computing via UPF on board: for similar reasons as those who led to local data switching, edge computing solutions can be introduced. Also in this case, GEO satellites are assumed, and the UE can establish a PDU session with a PDU UPF on-board or with an UL CL/BP and local PSA UPF on-board (i.e., the two local switching types introduced in TS 23.501). The SMF selects the option based on the GEO satellite ID provided by the AMF, which is informed of the satellite category. This selection is performed during the PDU Session Establishment procedure or the PDU Session Modification procedure as per TS 23.502, [84].

SUPPORT FOR DISCONTINUOUS COVERAGE

In discontinuous coverage, the UEs: i) might be covered only at specific times/places; ii) the UEs' location may not be timely known at the network side for efficient paging; iii) may not always need to be awake, to reduce power consumption. The normative work for this scenario was performed within "Stage 2 of 5GSAT_Ph2: 5GC/EPC enhancement for satellite access Phase 2" (5GSAT_Ph2) in stage 2 and "CTx aspects of 5GSAT_Ph2" (5GSAT_Ph2) for stage 3, based on the outcomes of the study phase reported in TR 23.700-28, [109]. It shall be mentioned that the identified solutions shall be applicable for both Evolved Packet System (EPS) and 5GS, with the baseline being the discontinuous coverage solutions for the Evolved Packet Core network (EPC); the case for 5GS/EPS inter-working is not considered. In this context, two key issues were identified and addressed for discontinuous coverage scenarios:

- Mobility management enhancements, aimed at identifying the gaps in the Rel. 17 solution for EPS with respect to: i) study how the UE determines that it will remain with no service or that it shall register to available different RATs/PLMNs to receive service during the discontinuity phases of the current RAT; and ii) study how to reduce the impact to the target RAT or system due to the potentially large number of users requesting access to receive the normal service.

- Power saving enhancements, to address the issues that might still arise despite Rel. 17 allows the UE in EPSs to deactivate the Access Stratum (AS) without network coverage, e.g., whether de-registration occurs due to any inconsistency of CM states between the UE and the core network and, the usage of eDRX in CM-IDLE state, etc. Based on the coverage information of the UE, the study focused on understanding how to enhance the power saving mechanisms to make sure that the UE does not attempt PLMN access when out of coverage, while attempting it when in coverage as needed (e.g., transfer data, receive paging, etc.). In this context, TR 23.700-28 clearly states that network coverage can be provided by any RAT supported by the UE.

The solutions identified for the above aspects are provided in TS 23.501 for: i) mobility management and power saving optimisation (clause 5.4.13.1); ii) coverage availability information provisioning to the UE (clause 5.4.13.2) and to the AMF (clause 5.4.13.3); iii) paging (clause 5.4.13.4); and iv) overload control (clause 5.4.13.5).

4.2.2.3 RAN

The WI on "NR NTN enhancements" (NR/_NTN/_enh) introduced several enhancements: i) uplink coverage enhancements; ii) network verified UE location; iii) NTN-TN and NTN-NTN mobility and service continuity; and iv) NR-NTN deployment above 10 GHz bands (already described in Section 4.2.2.1).

UPLINK COVERAGE ENHANCEMENTS

To improve the uplink coverage, two enhancements have been introduced in Rel. 18, [91]:

- Physical Uplink Control Channel (PUCCH) repetition for Msg4 HARQ-ACK configured in the study item or dynamically in the Downlink Control Information (DCI) for Msg4 when multiple repetition factors are configured in the study item. The following operations and assumptions hold:
  - The supported number of transmissions are 1, 2, 4, and 8. In case a single value between 2 and 8 is configured in the SIB, then the repetition factor is applied. In case multiple values are selected between 1 and 8 in the SIB, one of them is indicated in Downlink Assignment Index (DAI) field of DCI format 1_0 with Cyclic Redundancy Check (CRC) scrambled by the Temporary Cell Radio Network Temporary Identifier (TC-RNTI). It shall be noticed that the repetition factor applied to Msg4 HARQ-ACK is applied to any PUCCH transmission before the dedicated PUCCH resource is approved.

  - The repetition slot counting is defined in TS 38.213, [89].
  - Frequency hop for PUCCH retransmission for Msg4 HARQ-ACK as per Rel. 15/16/17 is applied in every slot.
  - A Reference Signal Received Power (RSRP) threshold can be configured in the SIB when a certain number of repetitions is defined. In this case, the UE that can apply PUCCH repetitions for Msg4 HARQ-ACK reports this capability via a Msg3 PUSCH only when the RSRP is below the threshold. If the threshold is not configured, the repetition capability is reported without any condition.

- Improved channel estimation via NTN-specific PUSCH DMRS bundling enhancement, which enables DMRS bundling in the presence of a time drift with the UE maintaining phase continuity by considering the effects of the transmission delay variation between it and the UTSRP.

NETWORK VERIFIED UE LOCATION

The 5GC can request a network verification procedure for a UE in RRC_CONNECTED to confirm it is consistent with the network-based assessed location; it shall be noticed that the management of UEs not supporting the location verification is left to network implementation (the procedure is in fact optional, and it is up to the core network to decide when to initiate it). The verification of the UE's location is based on multi-RTT with a single satellite; at least the following measurements are reported, as per TS 38.215, [110]: i) the gNB RX-TX time difference at the UTSRP (clause 5.2.3); ii) the UE RX-TX time difference (clause 5.1.30); iii) the UE RX-TX subframe offset (clause 5.1.46); and iv) the downlink timing drift (clause 5.1.47). The 5GC can help by including accurate ephemeris information (position and velocity) at the time of the multi-RTT measurements, the epoch time, and the common TA parameters defined in Section 1.1.3 (ta-Common, ta-CommonDrift, and ta-CommonDriftVariation). It is worthwhile mentioning that TR 38.882, [111], indicates that:

- The UE location information is considered verified if the UE reported location is consistent with the network-based assessment within 5-10 km (which is similar to TN macro-cells). This would in fact allow country discrimination and the corresponding selection of an appropriate core network to support all regulatory services (e.g., lawful intercept and emergency calls).
- The network-based assessment shall neither impact significantly the latency of the targeted services nor infringe privacy requirements related to the UE location.

NTN-TN AND NTN-NTN MOBILITY AND SERVICE CONTINUITY

Several mobility and service continuity features were improved or introduced from scratch in Rel. 18 NTN.

NTN-TN mobility: for both NTN-to-TN and TN-to-NTN mobility, SIB-based adaptations were introduced.

- NTN-to-TN: the NTN can broadcast the cell coverage areas and cell information of both 5G and 4G terrestrial networks in the newly introduced SIB25. This information is characterised by a list of geographical TN areas with the corresponding frequency allocations. With such information, the UE can reduce the power consumption by avoiding TN measurements based on such coverage areas.

- TN-to-NTN mobility: the terrestrial cells can broadcast the ephemeris of NTN neighbouring satellites via SIB19, and the UE can use this information to connect to the NTN component.

The new SIB25 block is described in TS 38.331, [88]. It contains a list of TN coverage area's information to assist NTN UEs in RRC_IDLE and RRC_INACTIVE, in the field coverageAreaInfoList, which includes: the TN area ID, its reference location (as per Rel. 17)[5], and the distance from the TN coverage reference location in multiples of 1 meter (up to 65535).

RACH-less handover: as indicated in TS 32.321, [76], and TS 38.331, [88], RACH-less handover has been introduced. This is a Layer 3 (L3) procedure that avoids the initiation of a RACH procedure during handover (which can be either between gNBs or during a feeder/satellite switch-over). This ultimately reduces the RA congestion in the target cell. As described in clause 5.33 of TS 38.321, the initial uplink transmission of a RACH-less handover procedure can be performed either using a dynamic uplink grant or a configured uplink grant defined by the RRC.

Conditional handover: building on the enhancements introduced in Rel. 17, a new even (CondEventD2) has been introduced. In this case, a reference location and a distance threshold for the source and target cells are defined. In addition, in time-based CHO, the procedure can be combined with the RACH-less handover described above.

Satellite switch-over: upon a hard or soft satellite switch-over in the quasi-Earth fixed case, with the same SSB frequency and gNB, it is possible to implement a satellite switch-over with resynchronisation. This procedure avoids L3 mobility for the UEs in the cell, as it maintains the same PCI on the geographical area. CHO can also be configured to simultaneously operate with this procedure. When the satellite switch-over is soft, the capable UE can start synchronising with the target satellite

[5] TS 38.331 indicates that the location from Rel. 17 is provided as the Ellipsoid-Point parameter in LTE, defined in TS 37.355, [112]. This field provides the degrees of latitude and longitude with 23 and 24 bits, respectively, plus a bit to indicate the sign of the latitude.

before the source satellite ends to serve the area, without needing to connect simultaneously to both (it acquires the new synchronisation at t-ServieStart and applies it after t-Service) as long as the SSBs from the source satellite and target satellite are not overlapped in time. In the hard switch-over case, the UE can start synchronising with the target only after the satellite switch procedure has been initiated (the UE synchronises to the target after t-Service). With respect to mobility aspects, it shall be mentioned that, apart from the NTN enhancements discussed above and in Section 1.1.3 for Rel. 17, the same principles and procedures for TNs apply. In particular: i) for mobility in RRC_IDLE and RRC_INACTIVE, the specifications are in clause 9.2.1 of TS 38.300; and ii) for mobility in RRC_CONNECTED, the specifications are in clause 9.2.3.2 of TS 38.300 (CHO is reported in clause 9.2.3.4). Finally, we report below all the measurements' enhancements introduced for NTN in addition to those specified in clause 9.2.4 of TS 38.300, [91]:

- The network can configure: i) multiple SS/PBCH block Measurement Timing Configuration (SMTC) in parallel per carrier, for a given set of cells depending on the capabilities of the UE; ii) measurement gaps based on multiple SMTCs; and iii) assistance information for the UE (such as the ephemeris, the common TA parameters, and k_{mac}) provided in SIB19, in order to perform measurements in the neighbouring cells in idle, inactive, or connected state.

- The network-controlled adjustments of the SMTCs can exploit assistance from the UE in RRC_CONNECTED, while UEs in idle or inactive state can adjust the SMTCs based on their location information and the information in SIB19.
  - The UE assistance information includes the service link propagation delay differences between the serving cell and the neighbouring ones.
  - When the UE is idle or inactive, it depends on the UE implementation whether the NTN neighbouring cell measurements on a cell defined in SIB3/SIB4, but not included in SIB19, can be performed.
  - When the UE is connected, it is again up to implementation whether to perform the neighbouring NTN cell measurements on a cell included in the measurement configuration, but without the corresponding information on the measurement configuration or in SIB19.
  - The UEs can perform time-based and location-based measurements on neighbouring cells in idle or inactive states (for these measurements, please see the discussions above). Time-based measurements can be applicable for the feeder link switch-over in cell selection/re-selection.
  - To limit the required measurements, the rules for cell re-selection are reported in TS 38.304, clause 5.2.4.2, [93].

TABLE 4.5: Target frequency allocations and duplexing for Rel. 19 NTN

UL (UE-to-SAN)	DL (SAN-to-UE)	Duplexing
1668-1675 MHz	1518-1525 MHz	FDD
2000-2020 MHz	2180-2200 MHz	FDD NR & LTE
1626.5-1660.5 MHz	1518-1559 MHz	FDD
1668-1675 MHz	1518-1559 MHz	FDD
10.7-12.75 GHz	14-14.5 GHz	FDD

TABLE 4.6: Candidate frequency allocations and duplexing for Rel. 19 NTN

UL (UE-to-SAN)	DL (SAN-to-UE)	Duplexing
1616-1626.5 MHz	1616-1626.5 MHz	TDD
10.7-12.75 GHz	13.75-14.5 GHz	FDD
10.7-12.7 GHz	13.75-14.5 GHz	FDD
10.7-12.75 GHz	12.75-13.25 GHz	FDD
10.7-12.7 GHz	12.7-13.25 GHz	FDD

4.2.3 Future Release 19

The normative work for Rel. 19 is currently on-going and it will be completed by December 2025. Below, we report the most relevant activities that are being performed related to NTN.

4.2.3.1 Payload, User Equipment, and Frequency Allocations

The payload and user equipment characteristics being discussed are reported in Section 1.3.3, as they will be defined as part of the RAN enhancements. With respect to frequency allocations, the targeted bands are reported in Table 4.5, while Table 4.6 reports other potential candidates to be discussed. It can be noticed that these frequency allocations include the extension of L/S band, but also the addition of Ku-band NTN.

4.2.3.2 SA and CT

Within SA1, the study item on "Study on satellite access - Phase 3" (FS_5GSAT_Ph3) addressed three new use cases, and the related requirements, for 5GSs exploiting an NTN component. TR 22.865, [113], captures the set of specific use cases and the related service requirements for:

- Store and Forward (S&F) operations for delay-tolerant communications, specifying that this is an NTN operation mode where the 5GS can provide a certain level of service (storing and forwarding data) when the satellite connectivity is intermittent or temporarily unavailable.

- UE-Satellite-UE communications, in which some UEs might be allowed to communicate using the satellite access without the need to go down to the ground network (for the User Plane), thus avoiding large latencies and limited data rates, in addition to reducing the consumption of backhaul resources.
- GNSS-independent operation, to allow the provisioning of satellite access to UEs without GNSS capabilities or without access to such capabilities.
- Positioning enhancements for satellite access, for specific scenarios.

The set of consolidated requirements for the above use cases is provided in TR 22.865, clause 6. The corresponding normative work (stage 1) is captured in TS 22.261 within the WI on "Satellite access Phase 3" (5GSAT_Ph3): clause 6.46.7 for satellite and relay UEs, clause 6.46.8 for S&F, clause 6.46.9 for UE-Sat-UE communications, and clause 6.46.10 for positioning aspects. With respect to relays, it shall be mentioned that 3GPP discarded solutions based on Integrated Access and Backhaul (IAB) nodes, in favour of Wireless Access Backhaul (WAB) solutions for which the NTN component can only be used on the backhaul link. Within SA2, based on the above studies, the study item on "Study on Integration of satellite components in the 5G architecture Phase 3" (FS_5GSAT_ARCH_Ph3) is addressing the architecture enhancements, with the results reported in TR 23.700-29, [114]. In this document, three key issues, and the potential solutions, are identified: i) support of regenerative payloads (at least one eNB/gNB on-board); ii) support of S&F communications; and iii) support of UE-Sat-UE communications. The main assumptions for these considerations are reported hereafter:

- ISLs and feeder links are assumed to only act as transport layer links, which means that they will not be specified within 3GPP.
- S&F solutions lead to intermittent UE-Sat-ground connectivity and they shall work also without ISLs.
- In the UE-Sat-UE case, a connection to the ground is always assumed for the Control Plane. This use case refers mainly to IP Multimedia Subsystem (IMS) multimedia telephony service (MMTEL) and mission critical services. Priority is given to scenarios with the UEs under the same satellite coverage.
- In all scenarios, the objective is that of minimising the impact to UEs, network functions, and entities, which means reusing the existing procedures and functionalities as much as possible

The related management enhancements are discussed in the study item on "Study on Management Aspects of NTN Phase 2" (FS_NTN_OAM_Ph2) and will be reported in TR 28.874, [115], which clearly highlights that the reference management scenario for NTN is the one provided in TS 28.530, [116]. The following aspects are being considered: i) the management of connections and associations between the satellite and the ground systems (gNB/eNB/CN/management system); ii) management for mobility coordination (including NTN neighbour cell management and NTN Tracking area management); iii) management for the support of S&F and UE-Sat-UE communications; and iv) management of secure connections in NTN.

4.2.3.3 RAN

At RAN level, the work plan for the WI " NTN for IoT Phase 3" (NR_NTN_Ph3) aims at introducing further enhancements related to, [117]:

- Evaluate downlink coverage enhancements to support additional reference satellite payload parameters for both GSO and NGSO systems in FR1 and FR2:
  - definition of additional satellite payload parameters assuming power sharing among satellite beams or different satellite beam patterns/sizes over the satellite footprint, such that satellite beams may not all be simultaneously active or may be active below the nominal EIRP density per satellite beam due to limited power and limited feeder link bandwidth;
  - define the related power sharing assumptions and the link/system level evaluation methodology/KPIs;
  - study, if required, link level enhancements for FR1 and/or system level enhancements for FR1/FR2 for dynamic and flexible power sharing between beams or different satellite beam patterns/sizes. In this context, it shall be noticed that no SSB enhancements are foreseen other than the extension of its periodicity, which would only apply to the NTN component. NGSO systems, in particular LEO satellites at 600 km, are prioritised for this study and the UE antenna is assumed to have -5.5 dBi gain for handheld terminals (with two receivers).
- uplink capacity enhancements in FR1, in particular through the design of Orthogonal Cover Codes (OCC) for DFT-s-OFDM PUSCH, at least for the multiplexing of 2 or 4 UEs with PUSCH repetitions. It is explicitly reported that this enhancement is not targeting: MU-MIMO capabilities, PUSCH DMR, initial access, and PRACH;
- specification of the signalling of the intended service area for broadcast services via NTN (in particular SIB signalling and the required signalling between the 5GC and the NG-RAN);
- support of regenerative payloads, specifying all the required enhancements for intra-/inter-gNB mobility (Xn over feeder link or over ISL);
- support RedCap/eRedCap UEs in FR1.

4.3 NTN Release 18 Software implementation

This section reports all the software adaptations required to incorporate Release 18 NTN features into a Release 17 software-based gNB implementation. The software modifications are presented for each network element (gNB, UE, and CN) and include feature implementation for: uplink coverage enhancement, NTN to TN mobility, TN to NTN mobility, RACH-less handover, conditional handover, satellite switch with resynchronization, and discontinuous coverage.

4.3.1 gNB SW Modifications

4.3.1.1 General

The gNB shall be capable of implementing additional UE capability parameters received in the **UECapabilityInformation** message (RRC Protocol TS38.331 v18.2.0):

- **sib19-Support-r18:** Specifies whether the UE supports receiving Release 18 (Rel.18) fields within SIB19.

- **softSatelliteSwitchResyncNTN-r18:** Indicates if the UE supports a soft satellite switch with re-synchronization, as per TS 38.331[RD 8]. If the UE supports this feature, it must also indicate support for hardSatelliteSwitchResyncNTN-r18.

- **hardSatelliteSwitchResyncNTN-r18:** Specifies whether the UE supports a hard satellite switch with re-synchronization, according to TS 38.331. A UE with this capability must also indicate support for nonTerrestrialNetwork-r17. If the UE supports hardSatel-liteSwitchResyncNTN-r18 but not softSatelliteSwitchResyncNTN-r18, it can still perform a hard satellite switch with resynchronization in networks that support soft satellite switching, as defined in TS 38.331.

- **mt-SDT-NTN-r18:** Indicates whether the UE supports initiating an MT-SDT (Mobile Terminated-Satellite Direct Terminal) procedure in an NTN environment. This includes initi-ating a random access procedure with a 4-step RA type. If the UE also supports twoStepRACH-r16 for NTN, it can initiate a 2-step RA type in response to an MT-SDT indication received in a paging message, as specified in TS 38.331, [88].

- **ntn-VSAT-AntennaType-r18:** Indicates the type of antenna used by a VSAT (Very Small Aperture Terminal) UE, specifying whether it is electronically or mechanically steered. UEs that support this feature must also indicate support for nonTerrestrial- Network-r17.

- **ntn-VSAT-MobilityType-r18:** Specifies whether a VSAT UE is mobile or fixed. UEs supporting this capability must also indicate support for nonTerrestrialNetwork-r17.

- **fr2-Add-UE-NR-CapabilitiesNTN:** Defines NTN-specific capabilities that differ from those in Terrestrial Networks (TN). If absent, the capabilities specified in fr2-Add-UE-NR-Capabilities will apply to NTN as well.

- **AddlocationBasedCondHandoverEMC-r18:** Indicates whether the UE supports lo-cation based conditional handover for an NTN Earth-moving cell, i.e. condEventD2 as spec-ified in TS 38.331, [88]. A UE supporting this feature shall also indicate the support of condHandover-r16 for NTN bands and the support of nonTerrestrialNetwork-r17. UE shall set the capability value consistently for all FDD-FR1 NTN bands and all FDD-FR2 NTN bands respectively.

- **ntn-NeighbourCellInfoSupport-r18:** Indicates whether the UE supports configuration of ntn-NeighbourCellInfo-r18 in MeasObjectNR for dedicated ephemeris. A UE supporting this feature shall also indicate the support of nonTerrestrialNetwork-r17.

- **ntn-DMRS-BundlingNGSO-r18:** Indicates whether the UE supports DM-RS bundling for PUSCH over consecutive slots in NGSO scenarios and pre- compensation to keep phase rotation due to timing drift within the phase difference limit. A UE supporting this feature shall indicate support of uplinkPreCompensation-r17 and at least one of dmrs-BundlingPUSCH-RepTypeA-r17, dmrs-BundlingPUSCH-RepTypeB-r17 or dmrs-BundlingPUSCH-RepTypeC-r17.

4.3.1.2 Uplink Coverage Enhancement

- The gNB shall be capable of delivering two additional parameters in SIB19 (RRC Protocol TS38.331 v18.2.0):
  - **numberOfMsg4HARQ-ACK-Repetitions:** It indicates the number of repetition slots for PUCCH transmission with HARQ-ACK information for Msg4, see clause 9.2.6 in TS 38.213. The first/leftmost bit corresponds to the repetition factor 1, the second bit corresponds to repetition factor 2, the third bit corresponds to the repetition factor 4, and the last/rightmost bit corresponds to the repetition factor 8. The repetition factor 1 shall be indicated together with at least one other repetition factor.
  - **rsrp-ThresholdMsg4HARQ-ACK:** This threshold is used by the UE for determining the configuration of the MAC entity for PUCCH repetition for Msg4 HARQ-ACK, as specified in clause 6.2.1 in TS 38.321.

- The gNB shall be capable of setting the indication on the number of repetition slots for PUCCH trans-mission with HARQ-ACK information for Msg4 in the Downlink Assignment Index field of DCI format 1_0 with Cyclic Redundancy Check (CRC) scrambled by the Temporary Cell Radio Network Tempo-rary Identifier (TC-RNTI). (PHY Protocol TS38.212 v18.4.0)

4.3.1.3 NTN to TN mobility

The NTN gNB shall be capable of delivering TN coverage information to assist neighbor cell meas-urements for the UEs in an NTN cell in SIB25. SIB25 shall specify:

- **coverageAreaInfoList:** Contains a list of TN coverage area's information to assist skipping TN measurements for NTN UEs in RRC_IDLE and RRC_INACTIVE, as defined in TS 38.304, [93].
- **tn-DistanceRadius:** Distance from the TN coverage area reference location. It is used for skipping TN measurements in RRC_IDLE and RRC_INACTIVE, as defined in TS 38.304, [93]. Each step represents 1m.

4.3.1.4 TN to NTN mobility

The TN gNB shall be capable of delivering **ntn-NeighCellConfigList** field with NTN coverage information to assist neighbor cell measurements for the UEs in an TN cell in SIB19. [RRC Protocol TS38.331 v18.2.0] **ntn-NeighCellConfigList** provides a list of NTN neighbour cells including their ntn-Config, car-rier frequency and PhysCellId. This set includes all elements of ntn-NeighCellConfigList and all elements of ntn-NeighCellConfigListExt. If ntn-Config is absent for an entry in ntn-NeighCellConfigListExt, the ntn-Config provided in the entry at the same position in ntn-NeighCellConfigList applies. Network provides ntn-Config for the first entry of ntn-NeighCellConfigList. If the ntn-Config is absent for any other entry in ntn-NeighCellConfigList, the ntn-Config provided in the previous entry in ntn-NeighCellConfigList applies.

4.3.1.5 RACH-less handover

- The gNB shall be capable of configuring RACH-less handover for a UE. To this aim, it should be capable of setting the **RACH-lessHO** parameter of the **reconfigurationWithSync** information element within **CellGroupConfig** information element. **RACH-lessHO** shall indicate the following fields (RRC Protocol TS38.331 v18.3.0):
  - **targetNTA:** This field refers to the timing adjustment, see TS 38.213, [89] and TS 38.321, [76], indicating the NTA value which the UE shall use for the target PTAG of handover. The value zero corresponds to NTA=0, while the value source corresponds to the NTA value of the source PTAG indicated by the tag-Id. In this version of the specification, the network shall always configure this field if rach-LessHO is part of an RRCReconfiguration message.
  - **tci-StateID:** This field indicates a beam that the UE should use in the target cell to monitor PDCCH for initial uplink transmission and indicates the TCI state information to be used in the target cell. The network configures this field in case this cell is not a NTN cell.

  - **ssb-Index:** This field indicates a beam that the UE should use in the target cell to monitor PDCCH for initial uplink transmission, see TS 38.321, [76]. The network configures this field when cg-RRC-Configuration is not configured for the initial uplink transmission in RACH-less handover in NTN or in case this cell is not a mobile IAB cell.

- The gNB shall be capable of configuring an uplink grant of type 1 for RACH-less handover trough the **CG-RRC-Configuration** field of **ConfiguredGrant-Config** information element. **CG-RRC-Configuration** field can be present only if **rach-LessHO** is present in **reconfigurationWithSync** and shall indicate the following fields:

  - **cg-RRC-RSRP-ThresholdSSB:** An RSRP threshold configured for SSB selection for the CG as specified in TS 38.321, [76]. This field is absent in cg-LTM-Configuration.
  - **cg-RRC-RetransmissionTimer:** Indicates the initial value of the configured grant retransmission timer used for the initial transmission of CG with CCCH (for CG-SDT) or DCCH message (see TS 38.321, [76]) in multiples of periodicity.
  - **rrc-DMRS-Ports:** Indicates the set of DMRS ports for SSB to PUSCH mapping (see TS 38.213, [89]). The first (left-most / most significant) bit corresponds to DMRS port 0, the second most significant bit corresponds to DMRS port 1, and so on. A bit set to 1 indicates that this DMRS port is used for mapping.
  - **rrc-NrofDMRS-Sequences:** Indicates the number of DMRS sequences for SSB to PUSCH mapping (see TS 38.213, [89]).
  - **rrc-SSB-Subset:** Indicates SSB subset for SSB to CG PUSCH mapping within one CG configuration. The first/leftmost bit corresponds to SS/PBCH block index 0, the second bit corresponds to SS/PBCH block index 1, and so on. Value 0 in the bitmap indicates that the corresponding SS/PBCH block is not included in the SSB subset for SSB to CG PUSCH mapping while value 1 indicates that the corresponding SS/PBCH block is included in SSB subset for SSB to CG PUSCH mapping. If this field is absent, UE assumes the SSB set includes all actually transmitted SSBs.
  - **rrc-SSB-PerCG-PUSCH:** The number of SSBs per CG PUSCH (see TS 38.213, [89]). Value one corresponds to 1 SSBs per CG PUSCH, value two corresponds to 2 SSBs per CG PUSCH and so on.
  - **rrc-P0-PUSCH:** Indicates P0 value for PUSCH in steps of 1dB (see TS 38.213, [89]). When this field is configured, the UE ignores the p0-PUSCH-Alpha. This field is absent in cg-LTM-Configuration.
  - **rrc-Alpha:** Indicates alpha value for PUSCH. alpha0 indicates value 0 is used, alpha04 indicates value 4 is used and so on (see TS 38.213, [89]).

When this field is configured, the UE ignores the p0-PUSCH-Alpha. This field is absent in cg-LTM-Configuration.

- The target gNB of a RACH-less handover procedure shall be capable of dynamical signalling of an UL grant.

4.3.1.6 Conditional handover

- The NTN gNB shall be capable of delivering the **movingReferenceLocation** field of SIB19 specifying the reference location of the serving cell of an NTN Earth-moving cell at a time reference. It is used in the evaluation of eventD2 and condEventD2 criteria for the serving cell in RRC_CONNECTED, and location-based measurement initiation in RRC_IDLE and RRC_INACTIVE when **distanceThresh** is also configured, as defined in TS 38.304, [93]. The time reference of this field is indicated by **epochTime** in **ntn-Config** of the serving cell. This field is only present in an NTN cell.
- The gNB should be capable of specifying criteria for triggering of event D2 conditional handover by setting **eventD2** and **condEventD2** parameters in **ReportConfigNR** information element. It should be able to set:
  - **distanceThreshFromReference1, distanceThreshFromReference2:** Distance from a moving reference location determined by the UE based on the serving cell movingReferenceLocation broadcast in SIB19 or referenceLocation and the corresponding epoch time and satellite ephemeris configured within the MeasObjectNR associated to the event for condEventD2. Each step represents 50m.
  - **timeToTrigger:** Time during which specific criteria for the event needs to be met in order to execute the conditional reconfiguration evaluation.
  - **hysteresisLocation** is a parameter used within entry and leave condition of a location based event triggered reporting condition. The actual value is field value * 10 meters.
  - **timeToTrigger** specifies the value range used for time to trigger parameter, which concerns the time during which specific criteria for the event needs to be met in order to trigger a measurement report.
  - **reportOnLeave:** Indicates whether or not the UE shall initiate the measurement reporting procedure when the leaving condition is met if configured in eventD1, eventD2, eventH1, eventH2.

4.3.1.7 Satellite switch with resynchronization

The TN gNB shall be capable of delivering **SatSwitchWithReSync** field in SIB19, setting the **ntn-Config** parameter from Rel.17 and the two following additional parameters:

- **t-ServiceStart:** Indicates the time information on when the target satellite is going to start serving the area currently covered by the serving satellite. The field indicates a time in multiples of 10 ms after 00:00:00 on Gregorian calendar date 1st January 1900 (midnight between Sunday, December 31, 1899, and Monday, January 1, 1900). The exact start time is between the time indicated by the value of this field minus 1 and the time indicated by the value of this field. The reference point for t-ServiceStart is the uplink time synchronization reference point of the serving satellite.
- **ssb-TimeOffset:** Indicates the time offset of the SSB from target satellite at its uplink time synchronization reference point with respect to the SSB from source satellite at its uplink time synchronization reference point. It is given in number of subframes.

4.3.2 UE SW Modifications

4.3.2.1 General

The UE shall convey the following additional capability parameters in the **UECapabilityInformation** message, which the gNB shall be capable of interpreting: **sib19-Support-r18, softSatelliteSwitchResyncNTN-r18, hardSatelliteSwitchResyncNTN-r18, mt-SDT-NTN-r18, ntn-VSAT-AntennaType-r18, ntn-VSAT-MobilityType-r18, fr2-Add-UE-NR-CapabilitiesNTN, Add locationBasedCondHandoverEMC-r18, ntn-NeighbourCellInfoSupport-r18**, and **ntn-NeighbourCellInfoSupport-r18**. The capability information parameters are described in section 4.3.1.1.

4.3.2.2 Uplink Coverage Enhancement

- The UE shall be capable of receiving **numberOfMsg4HARQ-ACK-Repetitions** and **rsrp-ThresholdMsg4HARQ-ACK** parameters from SIB19. (RRC Protocol TS38.331 v18.2.0)
- The UE shall be capable of receiving the indication on the number of repetition slots for PUCCH transmission with HARQ-ACK information for Msg4 from the Downlink Assignment Index field of DCI format 1_0 with Cyclic Redundancy Check (CRC) scrambled by the Temporary Cell Radio Network Temporary Identifier (TC-RNTI) in case more than one repetition option is indicated in the SIB. (PHY Protocol TS38.212 v18.4.0)
- The UE shall be capable of indicating its capability to apply PUCCH repetitions for Msg4 HARQ-ACK in Msg3 PUSCH, considering the content of **rsrp-ThresholdMsg4HARQ-ACK** parameter and its RSRP. (RRC Protocol TS38.331 v18.2.0)
- The UE shall be capable of transmitting the PUCCH repetitions for Msg4 HARQ-ACK in the dedicated PUCCH resources, if available, or it determines the

number of slots for repetitions of a PUCCH transmission with HARQ-ACK information based on an indication by **numberOfPUCCHforMsg4HARQACK-RepetitionsList** and the DCI information.

4.3.2.3 Network Verified UE Location

- Upon receiving a **NR-Multi-RTT-RequestCapabilities** information element request from the LMF, the UE shall be capable of indicating its compatibility with the network verified UE location function in the field **nr-NTN-MeasAnd-Report**, part of the information element N**R-Multi-RTT-MeasurementCapability**. The **nr-NTN-MeasAndReport** field, if present, indicates that the UE supports UE Rx-Tx Measurement and Report for Multi-RTT with single satellite in NTN with the following capabilities:
  - UE Rx-Tx time difference and UE Rx-Tx time difference offset measurement and report for Multi-RTT positioning;
  - Reporting DL timing drift due to Doppler over the service link associated with the UE Rx-Tx time difference measurement period.
- The UE shall be capable of measuring:
  - UE Rx - Tx time difference subframe offset;
  - DL timing drift.
- The UE shall be capable of providing the NTN related multi-RTT measures trough the information element **nr-NTN-UE-RxTxTimeDiff**. This field provides the offset of the UE Rx-Tx time difference measurement for NTN and comprises the following subfields:
  - **nr-NTN-UE-RxTxTimeDiffSubframeOffset** specifies the UE Rx - Tx time difference subframe offset in unit of subframe, as defined in TS 38.215, [110].
  - **nr-NTN-DL-TimingDrift** specifies the DL timing drift measurement, as defined in TS 38.215, [110]. The granularity of nr-NTN-DL-TimingDrift is 0.1 ppm. Values are given in unit of corresponding granularity.

4.3.2.4 NTN to TN mobility

- The UE shall be capable of receiving System Information Block 25 and its parameters **coverageAreaInfoList** and **tn-DistanceRadius**. The SIB25 parameters are described in section 4.3.1.3.
- The UE shall be capable of exploiting the TN coverage area's information to assist skipping TN measurements for NTN UEs in RRC_IDLE and RRC_INACTIVE. For UE camping on NTN cell, if the UE supports skipping TN measurement, and the UE has obtained its location information, and if **coverageAreaInfoList**

and **tn-AreaIdList** are broadcast in system information, the UE may not perform measurements of a TN frequency when UE is not in the coverage of that frequency provided via **tn-AreaIdList**, regardless of the frequency priority.

- **coverageAreaInfoList:** this indicates a list of TN coverage areas to assist skipping TN measurements for NTN UEs in RRC_IDLE and RRC_INACTIVE states.
- **tn-AreaIdList:** this indicates a list of TN area identities associated with each frequency to assist skipping TN measurements for NTN UEs in RRC_IDLE and RRC_INACTIVE states. Each TN area identity in the list identifies a TN coverage area.

4.3.2.5 TN to NTN mobility

- The UE shall be capable of receiving **ntn-NeighCellConfigList** field of SIB19 containing NTN coverage information
- The UE shall be capable of exploiting the NTN coverage area's information to assist neighbor cell measurements for the UEs in an TN.

4.3.2.6 RACH-less handover

- The UE shall be capable of receiving a **RACH-lessHO** indicaition parameter of the **reconfigurationWithSync** information element within **CellGroupConfig** information element.
- The UE shall be capable of receiving uplink grant of type 1 for RACH-less handover trough the **CG-RRC-Configuration** field of **ConfiguredGrantConfig** information element. **CG-RRC-Configuration field** can be present only if **rach-LessHO** is present in **reconfigurationWithSync**.
- The UE shall be capable of reacting to a rach-less handover initiation. The initial uplink transmission of a RACH-less handover procedure can be performed either using a dynamic uplink grant or a configured uplink grant Type 1 pre-allocated by RRC, if configured.
  1> if cg-RRC-Configuration is configured:
  2> select a configured uplink grant for initial uplink transmission according to TS38.321 clause 5.8.2;
  2> perform initial uplink transmission in the first available CG occasion for RACH-less handover according to TS38.321 clause 5.8.2;
  2> monitor the PDCCH as specified in clause 5.7 and TS 38.213.
  1> else:
  2> if tci-StateID is configured in rach-LessHO:
  3> indicate to lower layers the TCI state information included in tci-StateID.
  2> else if ssb-Index is configured in rach-LessHO:

3> indicate to lower layers the SSB index included in ssb-Index.
2> monitor the PDCCH as specified in TS 38.213.

4.3.2.7 Conditional handover

- The UE shall be capable of receiving a **movingReferenceLocation** field of SIB19 specifying the reference location of the serving cell of an NTN Earth-moving cell at a time reference.

- The UE shall be capable of receiving a receiving **eventD2** and **condEventD2** parameters from **ReportConfigNR** information element. The parameters of **eventD2** and **condEventD2** are described in section 4.3.1.6.

- The UE shall be capable to managing of **eventD2** (from 5.5.4.15a of 38.331) The variables in the formula are defined as follows:

  - **Ml1** is the distance between UE and a moving reference location for this event, not taking into account any offsets. The moving reference location is determined based on movingReferenceLocation and the corresponding epoch time and satellite ephemeris for the serving cell broadcast in SIB19.
  - **Ml2** is the distance between UE and a moving reference location for this event, not taking into account any offsets. The moving reference location is determined based on the parameter referenceLocation and the corresponding epoch time and satellite ephemeris configured within the MeasObjectNR associated to this event.
  - **Hys** is the hysteresis parameter for this event (i.e. hysteresisLocation as defined within reportConfigNR for this event).
  - **Thresh1** is the threshold for this event defined as a distance, configured with parameter distanceThreshFromReference1 in reportConfigNR for this event, from a moving reference location determined based on the parameter movingReferenceLocation and the corresponding epoch time and satellite ephemeris for the serving cell broadcast in SIB19.
  - **Thresh2** is the threshold for this event defined as a distance, configured with parameter distanceThreshFromReference2 in reportConfigNR for this event, from a moving reference location determined based on the parameter referenceLocation and the corresponding epoch time and satellite ephemeris configured within the MeasObjectNR associated to this event.
  - **Ml1** is expressed in meters.
  - **Ml2** is expressed in the same unit as **Ml1**.
  - **Hys** is expressed in the same unit as **Ml1**.
  - **Thresh1** is expressed in the same unit as **Ml1**.

- **Thresh2** is expressed in the same unit as **Ml1**.

1> consider the entering condition for this event to be satisfied when both condition D2-1 and condition D2-2, as specified below, are fulfilled;
Inequality D2-1 (Entering condition 1)
Ml1 – Hys > Thresh1
Inequality D2-2 (Entering condition 2)
Ml2 + Hys < Thresh2

1> consider the leaving condition for this event to be satisfied when condition D2-3 or condition D2-4, i.e. at least one of the two, as specified below, are fulfilled;
Inequality D2-3 (Leaving condition 1)
Ml1+Hys<Thresh1
Inequality D2-4 (Leaving condition 2)
Ml2 – Hys > Thresh2

4.3.2.8 Satellite switch with resynchronization

- The UE shall be capable of receiving **t-ServiceStart** and **ssb-TimeOffset** parameters of **SatSwitchWithReSync** field from SIB19.
- The UE shall be capable of acquiring DL synchronization with the SpCell served by the satellite indicated by **ntn-Config** in **SatSwitchWithReSync** between the time indicated by **t-ServiceStart** and the time indicated by **t-Service** for the serving cell.
- The UE shall be capable of starting timer T430 with the timer value set to **ntn-UlSyncValidityDuration** from the subframe indicated by **epochTime** in **ntn-Config** in **SatSwitchWithReSync** from Rel. 17.

4.3.2.9 Discontinuous Coverage

- The UE shall be able to exploit the satellites coverage availability information and its current and expected future locations to determine the **Start of Unavailability Period** and/or the **Unavailability Period Duration** for when it expects to be out of coverage, [83].
- The UE shall be able to include the **Start of Unavailability Period** and/or the **Unavailability Period Duration** and the **Unavailability-Type** fields within the **Unavailability-Information** information element of the Initial Registration request to the AMF.
- The UE shall be able to include the **Start of Unavailability Period** and/or the **Unavailability Period Duration** and the **Unavailability-Type** fields within the

Unavailability-Information information element of the Initial Registration request to the AMF.

4.3.3 CN SW Modifications

4.3.3.1 Network Verified UE Location

- The AMF shall be capable of initiating the location verification of a UE trough the **5GC-NI-LR** procedure and to include in indication in the request to the LMF that the request is for UE location verification.

- The LMF should be capable of sending an LPP Request Location Information message to the UE for location verification. This request includes indication of Multi-RTT measurements requested in **NR-Multi-RTT-RequestLocation-Information**, including any needed measurement configuration information, and required response time.

4.3.3.2 Discontinuous Coverage

- The AMF shall be able to determine, during the registration procedure, the values of the negotiated extended DRX parameters, the **timer T3324**, and the periodic registration update **timer T3512** to be provided to the UE. They are determined based on the discontinuous coverage maximum time offset, the unavailability period duration and the start of the unavailability period. The AMF should set the value of the mobile reachable timer and implicit de-registration timer based on the unavailability period duration and the start of the unavailability period.

- The AMF shall be able to determine (if not provided by the UE) or update the **Unavailability Period Duration** and/or the **Start of Unavailability Period** based on the **Unavailability Type** and other information as the Satellite Coverage Availability Information.

Chapter 5

O-RAN compliant NTN Disaggregated Architectures

This chapter reports the outcomes of the paper "RAN Functional Splits in NTN: Architectures and Challenges", [10], and it aims to explore the concept of RAN disaggregation, a fundamental pillar of the O-RAN architecture, which splits traditional base stations into different functional units.

The chapter introduces key NTN elements, including non-terrestrial nodes and communication links, and details three O-RAN-compliant system architectures: Full-gNB, Orbit-split and Feeder-split designs. These architectures address the unique challenges of NTN systems, such as dynamic link performance and interface mapping.

By detailing various functional split options and their implications for NTN integration, this chapter provides a comprehensive analysis of the challenges and opportunities presented by disaggregated architectures in dynamic and resource-constrained environments. Special attention is given to the unique requirements imposed by NTNs, including latency, bandwidth, and interface constraints, as well as the potential of AI-driven solutions to optimize these configurations within the evolving 6G ecosystem.

5.1 NTN System Architectures

In this section, we present two O-RAN-compliant NTN system architectures where the O-RAN framework is integrated into NTN elements: the Orbit-split and Feeder-split architectures. Part of this section reports the outcomes of the paper "O-RAN Based Non-Terrestrial Networks: Trends and Challenges" [5]. The focus is on two critical design aspects: *i*) the distribution of O-RAN components across NTN and terrestrial network entities; and *ii*) the mapping of O-RAN interfaces to the unique physical links of NTN systems, addressing challenges such as intermittent and unstable connections.

An NTN-based RAN architecture consists of NTN access segment, NTN Gateway (GW), and a terrestrial segment. The NTN access segment is constituted by different NTN payloads, i.e., network elements on-board a satellite, that orbit the

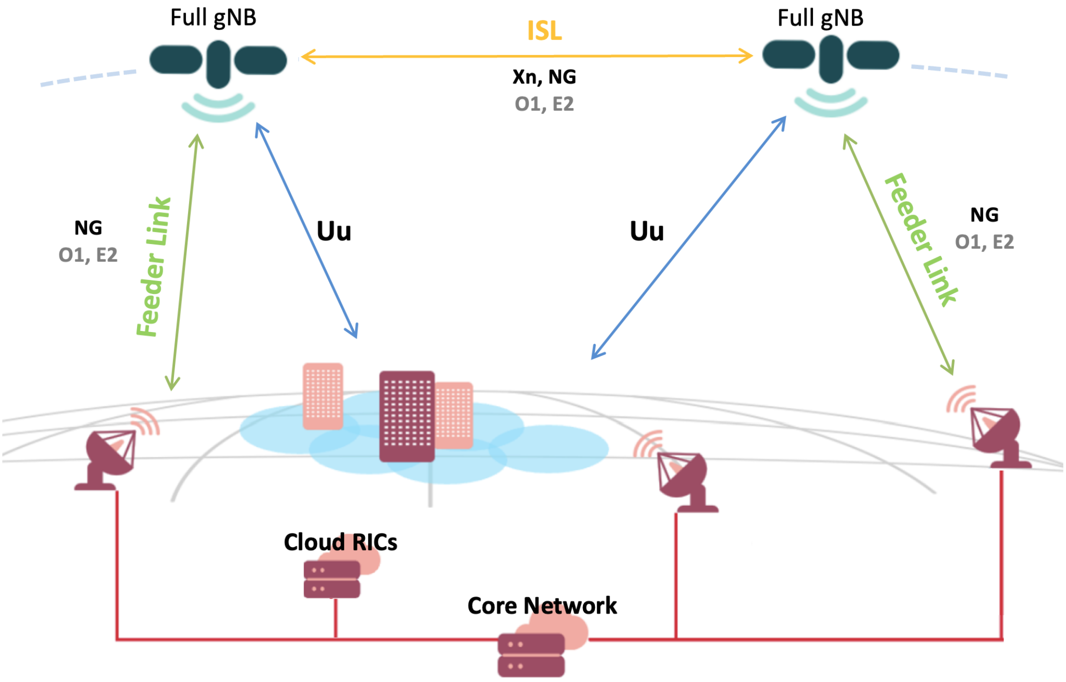


FIGURE 5.1: High-level system architecture of full gNB onboard.

Earth at different altitudes providing services to the on-ground users. The GW interconnects the payload to the terrestrial segment via a feeder link. The terrestrial segment is constituted of the CN. Finally, in the user segment, we consider users directly connected to the NTN elements. Rel. 19 of 3GPP foresees the standardization of features assuming regenerative payload [118]. This enables complete or partial implementation of the RAN protocol stack on the payload. In this framework, we identify three main architectural solutions that may exploit a functionally split RAN, described in the following subsections as Orbit-split, and Feeder-split architectures.

In the context of this work, the non-RT RIC and the near-RT RIC are assumed to be implemented in the cloud, interconnected to the ground network elements through the ground distribution network. Depending on the dimension of the network slice they have to serve, they can be deployed closer to the network edge or to the central cloud. For what concerns the open interfaces, their mapping on the physical network links of this NTN architecture is a non-trivial task. In a terrestrial implementation of the O-RAN architecture, the interfaces are built upon the ground distribution network, inter-connecting all the network elements. This Internet Protocol (IP) based network is often based on reliable optic fiber links, providing stable capacity and latency. On the other hand, in NTN networks some key open interfaces have to rely on intermittent links with unstable performances. The specific mapping of the O-RAN interfaces to the physical network links is addressed specifically for the three considered NTN system architectures in the following sections.

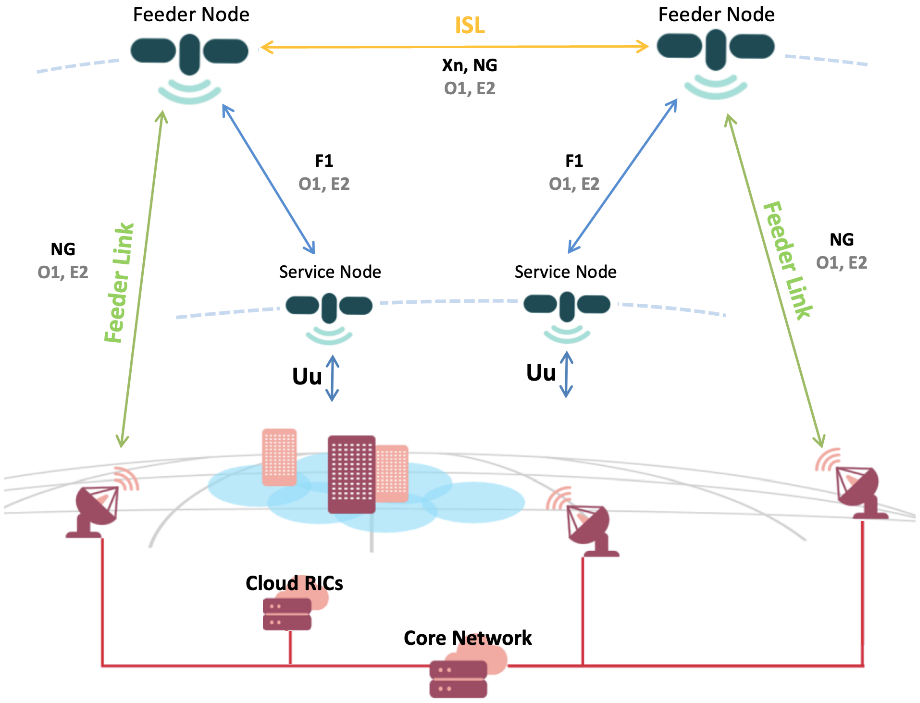


FIGURE 5.2: High-level system architecture of in-orbit RAN split.

5.1.1 Full-gNB Architecture

In the *conventional* architecture shown in Figure 5.1, the NTN access segment nodes are organized in a single constellation in which all the nodes serve the UEs implementing the full gNB. The feeder links connect each node to the terrestrial segment through the GW, implementing the Next Generation (NG) interface. The on-ground coverage is provided by all the nodes that serve the users through the Uu air interface. Furthermore, the implementation of the NG interface between the nodes enables the routing of the user-plane and control-plane packets in space.

Concerning 3GPP and O-RAN interfaces mapping to physical links, the NR-Uu Air Interface is implemented on the user access link of all nodes, while the ISLs implement the Xn and NG interfaces and transports the E2 and O1 O-RAN interfaces. In this configuration, the feeder link implements the NG interface to the CN and transports the E2 and O1 O-RAN interfaces.

5.1.2 Orbit-split Architecture

In the *Orbit-split* architecture, the NTN access segment nodes, which constitute a single or multiple constellations, are logically divided into two groups: the feeder node and the service node, Figure 5.2. A feeder node can be connected to one or multiple service nodes, while a service node can be connected to a single feeder node. Different RAN operations are performed by feeder nodes and service nodes depending on the selected functional split configuration. The feeder links connect each feeder node to the terrestrial segment through the GW, implementing the NG interface.

The Split Interface (SI) interface connecting service nodes with the serving feeder node is implemented through Inter Satellite Links (ISL), as well as the Xn interface interconnecting two feeder nodes. The SI, NG, and Xn interfaces can be transported by any optical link or Satellite Radio Interface (SRI), e.g., DVB-S2X, as long as specific signaling operations are guaranteed. The on-ground coverage is provided by the service nodes that serve the users through the Uu air interface. Furthermore, the implementation of the NG interface between the feeder nodes enables the routing of the user-plane and control-plane packets in space.

As shown in Figure 5.2, service satellites are mainly devoted to provide connectivity to the UEs but they don't have feeder links. Most of the available payload mass and power is thus devoted to maximize the service up- and downlink capacity, so these satellites will connect via ISLs to the feeder satellites but will have neither feeder links nor ISLs among them. On the other hand, feeder satellites do not have direct link to the UEs, but they implement the full transport network in space using ISLs and feeder links and providing additional processing capabilities in space to implement RAN and if needed Core Network and Edge Computing functionalities. Although from Figure 5.2 one might infer that service satellites are flying lower than feeder satellites, this is only a logical representation.

The advantages of this architectural solution are manifold, namely:

- it allows higher service link throughput, since no resources have to be provisioned for feeder link and ISL and all available power can be devoted to the service link.

- it offers better scalability and flexibility, since the feeder satellites are totally agnostic regarding which spectrum and bandwidth is used for the service links. As long as the ISL and feeder links capacity does not become the bottleneck, new service satellites (more powerful and/or operating in a different frequency bands) could be progressively and seamlessly added.

Concerning 3GPP and O-RAN interfaces mapping to physical links, the NR-Uu Air Interface is implemented on the user access link of service nodes, while on the ISL interconnecting feeder and service nodes, the interface depends on the type of split. Indeed, if the service node embarks only the RU, the ISL carries the FH and O1 interface, while if both the RU and DU are on the service node the implemented interfaces are the F1, E2 and O1. In this configuration, the feeder link implements the NG interface to the CN and transports the E2 and O1 O-RAN interfaces.

5.1.3 Feeder-split Architecture

In the *Feeder-split* architecture, the NTN access segment nodes serve as gNB-DUs while the gNB-CUs are implemented on-ground, Figure 5.3. The feeder links connect the CUs to the DUs through the GW, implementing the SI. The SI interfaces can still be transported by any optical link or SRI, as long as specific signaling operations are guaranteed. The NTN access segment nodes are interconnected through

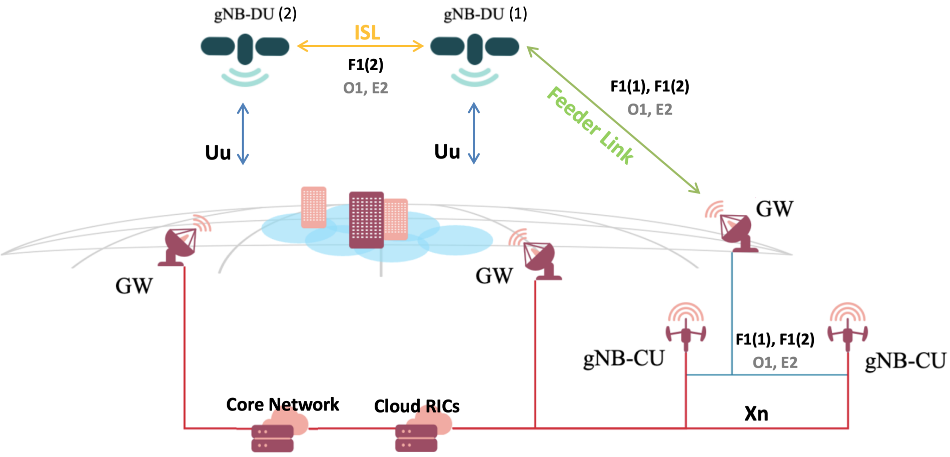


FIGURE 5.3: High-level system architecture of ground-orbit RAn split.

ISLs, enabling the distribution of the SI interface through ISL also to DUs that are not directly provided with a feeder link. The NG interface is implemented through the terrestrial network, as well as the Xn interface that interconnects the CUs on-ground. The on-ground coverage is provided by the DU-nodes that serve the users through the Uu air interface.

Concering 3GPP and O-RAN interfaces mapping to physical links, the NR-Uu Air Interface is implemented on the user access link, while on the feeder link, the interface depends on the type of split. Indeed, if the satellite embarks only the RU, the feeder link carries the FH and O1 interface, while if both the RU and DU are on-board the implemented Interfaces are the F1, E2 and O1. All the interfaces assigned to the feeder link are logical, *i.e.*, they can be implemented by means of any Satellite Radio Interface (SRI), such as DVB-S2X. Actually, the unstable performance of the feeder link is not a show-stopper problem, since the service provided by O1, E2, and F1 interfaces can be adapted to meet the instantaneous feeder link performances. On the contrary, the current standardization of these interfaces, as per 3GPP TR 38.473 [119] and O-RAN WG1 [35], do not allow their implementation upon intermittent links. As a consequence, for NGSO nodes with no direct visibility of the GW, the logical link with the serving gNB on-ground is ensured by INL.

5.2 Functional Split Options

According to TR 38.801, [20], the split of the gNB can be performed in eight different ways shown in Fig. 5.4.

The functional split options are are described in the following subsections highlighting the benefits of implementing a specific functional split in NTN and providing the analysis of the maximum one-way latency, and the computation of the required interface bandwidth, [120]. A summary of this analysis is provided in Table

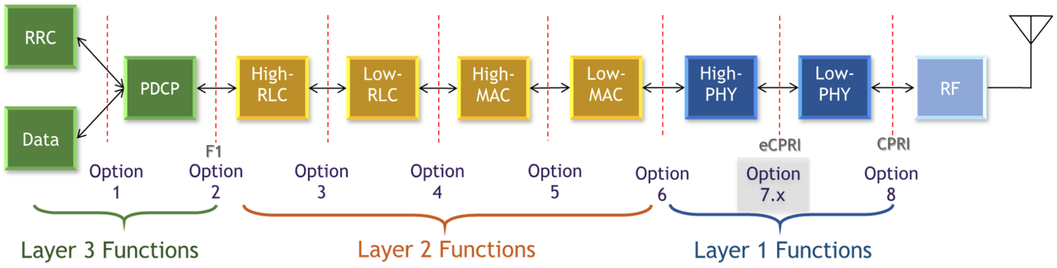


FIGURE 5.4: Available functional splits.

5.1 for options 1 to 8.

5.2.1 Option 8: RF/PHY

Option 8 splits the gNB between the PHY layer and the RF section. Indeed, only the RF sampler and the upconverter are left in the DU, resulting in a very simple DU, while all the remaining functions are centralized in the CU assuring the highest level of centralization. With this configuration, time In-phase - Quadrature (IQ samples are encapsulated in a protocol and delivered through the SI interface connecting the CU and DU. Thus, the data rate on the SI is maximum, constant, and scales with the number of DU and antennas. The DL fronthaul bitrate for split option 8 is defined by 3GPP in [121] as:

$$FH\ bitrate = SR * BTW * AP * 5$$

The UL fronthaul bitrate for split option 8 is defined by 3GPP in [121] as:

$$FH\ bitrate = SR * BTW * AP * 5$$

Where SR is the sample rate, BTW is the bitwidth, and AP is the number of antenna ports. For a scenario using 100 MHz bandwidth and 32 antenna ports the bitrate will be 157.3 Gbps [121] for both UL and DL.

Regarding the supported latency from a wireless requirement standpoint, fronthaul must not impact the functions or performance of existing devices. After thoroughly analyzing equipment specifications and implementation methods from multiple vendors, the latency for the functional split option between PHY and RF is determined in [122] to be 250 microseconds.

Split 8 offers several advantages and disadvantages. On the positive side, it enables the deployment of very small and cost-effective DUs while supporting uplink and downlink Coordinated Multi-Point (CoMP) with no performance degradation. It also facilitates the reuse of already deployed Common Public Radio Interface (CPRI) Remote Radio Heads (RRHs) or DUs, ensuring high levels of centralization and coordination across the protocol stack. The separation of RF and PHY enhances system modularity and allows operators to share RF components, while the DUs remain Radio Access Technology (RAT) agnostic. Furthermore, the largest processing resources are centralized in the CU pool, simplifying management and optimization.

However, Split 8 also presents challenges, such as requiring a high and constant bitrate on the fronthaul link that scales with the number of antennas. Additionally, it demands extra processing for CPRI compression at both DU and CU levels and imposes stringent latency requirements on the fronthaul, which can complicate deployment in some scenarios.

5.2.2 Option 7: Intra PHY

Option 7 splits the gNB at the PHY layer and is divided into options 7.1, 7.2, and 7.3 depending on the specific PHY functions that are centralized.

- **Option 7.1 (Low PHY)**:FFT is decentralized in the DU. This implies a reduction in the required interface data rate compared to option 8, even if it is still constant since the resource element mapping is performed in the CU. According to 3GPP in [121], the downlink fronthaul bitrate for split option 7–1 is specified. It is assumed here that split 7b matches split option 7–1, based on its description in [122]:

  $$\begin{aligned} FH\,bitrate =& SC * SY * AP * BTW * 2 * 1000 \\ &+ MAC\,info \end{aligned}$$

  The UL fronthaul bitrate for split option 8 is defined by 3GPP in [121] as:

  $$\begin{aligned} FH\,bitrate =& SC * SY * AP * BTW * 2 * 1000 \\ &+ MAC\,info \end{aligned}$$

  Where SC is the number of subcarriers, SY is the number of symbols, BTW is the bitwidth, and AP is the number of antenna ports. For a scenario using 100 MHz bandwidth and 32 antenna ports the DL bitrate will be 9.2 Gbps and the UL 60.4 Gbps [121]. In terms of supported latency, with the Hybrid Automatic Repeat reQuest (HARQ) process taking a maximum of 4ms and if we disregard air interface latency, the remaining time is allocated for process latency and the transmission between the DU and CU. Process latency consists of the CU processing delay, DU latency, and the time taken for transmission between the DU and CU. After a thorough evaluation of different vendors' equipment specifications and their implementation strategies, the observed maximum end-to-end latency is calculated in [122] to be 250μs. To ensure adequate processing time, the end-to-end latency must not exceed 250μs. Split 7.1 offers significant advantages, including reduced bitrate requirements on the fronthaul compared to Split 8, enabling more efficient deployments. It supports UL CoMP joint reception and DL CoMP coherent JT without performance loss and facilitates centralized scheduling and joint processing, enhancing network coordination and optimization. However, it comes with challenges, such as a high and constant fronthaul load, the need for very high UL bandwidth, and the complexity of managing subframe-level timing interactions between PHY components in the CU and DU, which demand precise synchronization and robust infrastructure.

- **Option 7.2 (Low PHY/High PHY)**: The precoding and resource element mapper is decentralized in the DU. This split option reduces the required data rate since the SI interface transports subframe symbols. Furthermore, the data rate becomes dependent on the required user traffic, implying a sensible data rate reduction in low traffic load contexts. According to 3GPP in [121], the downlink fronthaul bitrate for split option 7–1 is specified. It is assumed here that split 7a matches split option 7–2, based on its description in [122]:

  $$\begin{aligned} FH\ bitrate =& SC * SY * LA * BTW * 2 * 1000 \\ &+\ MAC\ info \end{aligned}$$

  The UL fronthaul bitrate for split option 7–2 is defined by 3GPP in [121] as:

  $$\begin{aligned} FH\ bitrate =& SC * SY * LA * BTW * 2 * 1000 \\ &+\ MAC\ info \end{aligned}$$

  Where SC is the number of subcarriers, SY is the number of symbols, BTW is the bitwidth, and LA is the number of layers. For a scenario using 100 MHz bandwidth and 32 antenna ports the DL bitrate will be 9.8 Gbps and the UL 15.2 Gbps [121]. In terms of supported latency, the considerations are similar to split 7.1. Split 7.2 features variable and moderate bitrate on the fronthaul, offering flexibility and potential multiplexing gains, [123]. It maintains support for UL CoMP joint reception and DL CoMP coherent JT without performance degradation, while enabling centralized scheduling and joint processing for enhanced coordination and optimization [20]. However, it requires an in-band protocol for Physical Resource Block (PRB) allocation, adding complexity, and still faces challenges with subframe-level timing interactions between PHY components in the CU and DU, necessitating precise synchronization and advanced management strategies [20].

- **Option 7.3 (High PHY)**: The scrambling, modulation, and layer mapping are decentralized in the DU. The signal delivered through the interface is modulated, sensibly reducing the required data rate proportionally to the modulation scheme used. Split 7.3 offers notable benefits, such as maintaining the close relationship between the FEC and MAC layer and modulating the signal in the DU, which significantly reduces the fronthaul bitrate. Additionally, the fronthaul load becomes cell load dependent, and pooling for the Turbo Codec is achievable compared to splits 1 through 6, [123]. Centralized scheduling is still possible [20], and JT and JR are supported, albeit with some limitations. However, Split 7.3 has drawbacks, including a more complex DU due to local modulation, no latency improvement for CoMP data path and CSI compared to the MAC-PHY interface, and the need for an in-band protocol for modulation, MIMO, and PRB allocation, [123]. Timing interactions at the subframe level between PHY components in the CU and DU remain a challenge, requiring precise synchronization [20].

5.2.3 Option 6: MAC/PHY

Option 6 splits the gNB between the MAC and PHY layers, resulting in all the physical processing being handled in the DU and the MAC being centralized in the CU. The SI interface delivers transport blocks and this leads to a considerable reduction in the required data rate compared to higher split options. Indeed, the load on the interface is proportional to the cell load. Relative to the higher split options, this option incurs greater overhead for scheduling control. According to 3GPP in [121], the DL fronthaul bitrate for split option 6 is specified as:

$$\begin{aligned} FH\ bitrate = &(PR+CR)*(BW/CBW)*(LA/CLA) \\ &\times\ (8/6) \end{aligned}$$

The UL fronthaul bitrate for split option 6 is defined by 3GPP in [121] as:

$$FH\ bitrate = (PR+CR)*(BW/CBW)*(LA/CLA)*(6/4)$$

Where PR is the peack rate, CR is the signaling rate, BTW is the bitwidth, BW is the bandwidth, LA is the number of layers, and CLS is the number of layers for control signaling. In a setup utilizing 100 MHz bandwidth, 8 layers, and 256 QAM modulation, the bitrate is DL 5.6 Gbps and UL 7.1 Gbps [121].

The latency supported for signalling and data in functional splits between MAC and PHY is stringent, with the HARQ process capped at 4ms. After accounting for process delays in PHY and RRH, including HARQ scheduling, and transmission time from RRH to UE, around hundreds of microseconds remain for transmission from CU to DU. Considering and analyzing multiple vendor equipment specifications and implementation strategies, the calculated maximum end-to-end latency in [122] is 250μs for the functional split between MAC and PHY. If the RLC timing process is cloud-implemented, processing delay is reduced beyond what the CU currently experiences, thereby increasing the maximum end-to-end latency to over 250μs.

Overall, split 6 offers benefits such as a low and cell load-dependent bitrate on the fronthaul, the feasibility of JT, and centralized scheduling. It also enhances processing for layers from MAC and above within the CU-pool, improving overall network efficiency [20]. However, the split separates the close relationship between FEC and MAC, introducing potential inefficiencies. Fronthaul delays may cause issues in 5G, particularly with shorter subframes [124]. Baseband pooling is limited to L2/L3 layers, covering only about 20% of the overall baseband processing requirements. Additionally, there are increased latencies for data paths and CSI feedback over the fronthaul, and the need for an in-band protocol for modulation, multi-antenna processing, and PRB allocation adds complexity, [123].

5.2.4 Option 5: Intra MAC

Option 5 splits the gNB inside the MAC layer, centralizing the scheduling task in the CU and leaving the MAC time-critical processing in the DU. The HARQ procedures and the functions where performances are proportional to the latency are instantiated in the DU, meaning a drastic reduction of the interface latency constraints. However, many of the computationally critical tasks are left in the DU, reducing the benefits of shared processing

The MAC's role involves data transfer and managing radio resources, so the bandwidth is similar to option 6, with a DL bandwidth of 5.6 Gbps and UL 7.1 Gbps [121]. The timing requirement for signaling and data within functional splits in intra-MAC is stringent, with the HARQ process lasting no more than 4ms. After accounting for the process delay of PHY and RRH, HARQ scheduling, and transmission time from RRH to UE, the remaining transmission time from CU to DU is about a few hundred microseconds, [122].

Split 5 ensures that the DU is limited to functions requiring real-time communication, reducing complexity [124]. It supports low and cell load-dependent bitrate on the fronthaul and enables efficient interference management across multiple cells, facilitated by enhanced scheduling technologies. The fronthaul latency requirements are flexible, depending on the realization and interaction of scheduling functions between the CU and DU. However, Split 5 introduces a complex interface between the CU and DU, making it challenging to define scheduling operations across them. Additionally, there are limitations in supporting certain CoMP schemes, restricting its adaptability, [20].

5.2.5 Option 4: RLC/MAC

Option 4 splits the gNB between the RLC and MAC layers. In this configuration, the scheduler is decentralized in the DU and it is distant from the closely related RLC functions leading to performance degradation. The SI data rate is lower and cell load dependent.

According to 3GPP in [121], the DL fronthaul bitrate for split option 4 is specified as:

$$FH\ bitrate = PR * (BW/CBW) * (LA/CLA) * (8/6)$$

According to 3GPP in [121], the UL fronthaul bitrate for split option 4 is specified as:

$$FH\ bitrate = PR * (BW/CBW) * (LA/CLA) * (6/4)$$

Where PR is the peack rate, CBW is the bandwidth for control signals, BW is the bandwidth, LA is the number of layers, and CLA is the number of layers for control signaling. In a setup utilizing 100 MHz bandwidth, 8 layers, and 256 QAM modulation, the bitrate is DL 5.2 Gbps and UL 4.5 Gbps [121]. Regarding supported latency, the scheduling interval between RLC and MAC operates per TTI. With MAC scheduling taking around 500 μs and RLC Protocol Data Units (PDU) segmentation or concatenation taking about 200 μs, the combined latency for RLC and MAC is approximately 100 μs.

Overall, split 4 provides the advantage of a low and cell load-dependent bitrate on the fronthaul link, enhancing flexibility and efficiency. However, it separates the closely related RLC and MAC layers, which can lead to operational inefficiencies [124]. Furthermore, it offers no significant benefits for LTE [20] and is impractical for 5G scenarios with shorter subframe sizes due to latency and timing challenges [124].

5.2.6 Option 3: intra RLC

Option 3 splits the gNB between high and low RLC layer, instantiating the RLC segmentation functions in the DU and the Automatic Repeat Request Automatic Repeat Request (ARQ) in the CU. This configuration facilitates the interconnection of multiple MAC entities with a common RLC entity. It also reduces the latency constraints since the scheduling computation is decentralized in the DU.

Concerning fronthaul bitrate, the transmitting side incorporates an RLC buffer and real-time flow control; on the receiving side, due to both the reduced number and frequency of UL packets compared to DL packets, along with reordering windows, the bandwidth requirement is less than that of option 2. Regarding supported latency, the end-to-end transmission delay of function split within intra RLC is not constrained by HARQ. Due to the high latency tolerance of the transmission buffer and the reordering windows on the receiving side, the latency for option 3 is about 10ms, [122].

5.2.7 Option 2: RLC/PDCP

Option 2 foresees the PDCP and RRC functions centralized in the CU while all the other gNB functions are performed in the DU. In this configuration all real-time functions are located in the DU, meaning very relaxed SI interface latency requirements. Further, the centralized functions still enable the dual connectivity feature. However, the vast majority of computational complexity is kept in the DU, implying a marginal computational multiplexing gain in the CU.

According to 3GPP in [121], the DL fronthaul bitrate for split option 2 is specified as:

$$\begin{aligned} FH\ bitrate = & PR * (BW/CBW) * (LA/CLA) * (8/6) \\ & +\ signaling \end{aligned}$$

According to 3GPP in [121], the UL fronthaul bitrate for split option 4 is specified as:

$$\begin{aligned} FH\ bitrate = & PR * (BW/CBW) * (LA/CLA) * (6/4) \\ & +\ signaling \end{aligned}$$

Where PR is the peack rate, CBW is the bandwidth for control signals, BW is the bandwidth, LA is the number of layers, and CLA is the number of layers for control signaling. In a setup utilizing 100 MHz bandwidth, 8 layers, and 256 QAM modulation, the bitrate is DL 4 Gbps and UL 3 Gbps [121]. The maximal end-to-end transmission latency for a functional split between PDCP/RLC is not constrained by HARQ, allowing this option to tolerate high latency with buffer presence. Considering end-to-end service latency, the latency requirement for the interface between NR eNB and CN can meet fronthaul latency requirements between PDCP/RLC. Preliminary evaluations in [122] suggest this functional split option supports a latency of about 10 ms.

5.2.8 Option 1: PDCP/RRC

Option 1 implements all user plane functions in the DU, leaving only the Radio Resource Control RRC function inside the CU. This split offers marginal performance gains compared to the monolithic gNB implementation and imposes a limitation on the number of DUs connected to the same CU. According to 3GPP in [121], the DL fronthaul bitrate for split option 2 is specified as:

$$FH\ bitrate = PR * (BW/CBW) * (LA/CLA) * (8/6)$$

According to 3GPP in [121], the UL fronthaul bitrate for split option 4 is specified as:

$$FH\ bitrate = PR * (BW/CBW) * (LA/CLA) * (6/4)$$

In a setup utilizing 100 MHz bandwidth, 8 layers, and 256 QAM modulation, the bitrate is DL 4 Gbps and UL 3 Gbps [121]. The end-to-end transmission delay for the functional split between RRC/PDCP is not constrained by HARQ. The interaction between RRC and PDCP primarily results from PDCP configuration signaling, which demands low latency. The latency for option 1 matches that required for the interface between gNB and NG-Core, and preliminary assessments in [122] suggest it is about 10 ms.

5.3 NTN-imposed Split Interface Challenges

As presented in the previous section,3GPP identified 8 possible functional splits and standardized split option 2. However, other functional splits can be considered to tackle the NTN architecture complexity, thus allowing more degrees of freedom in terms of: i) which parts of the stack are implemented in each unit according to the 8 possible splits, and ii) where each set of network functions is executed. Indeed, most network functions can be virtualized and moved to different nodes as long as latency and interface throughput requirements are met. However, selecting the gNB functional split is a challenging task since each split is characterized by different delay and connectivity requirements for the mid-haul, and different workload on the CU and DU.

In this section, we leverage the discussion of the current 5G NTN development in terms of architectural solutions to thoroughly analyze the impact of the typical NTN channel impairments on the available functional splits to provide design criteria. The analysis is supported by a link budget and latency computation for the feeder-link interface exploited in the feeder-split architecture and by a latency computation for the Optical Inter Satellite Links (OISL) link between feeder and service nodes in the orbit-split architecture.

Even if 3GPP defines a strict set of RAN functions to be implemented in the CU and RU, in the scope of this work we assume that there is no limitation in the number and type of RAN functions that can be implemented in the CU and DU. This assumption enables the implementation of all 8 functional splits between the two units. Accordingly, we define the interface connecting the CU and DU as the SI, able to handle the signalling and PDU delivery of all the functional splits.

5.3.1 Interface Capacity

The capacity supported by the SI interface is a fundamental aspect to analyze in the design of a functionally split RAN. In terrestrial networks, this interface is usually implemented through high-capacity Ethernet networks or dedicated fiber optic links, guaranteeing adequate performance and reliability. In order to introduce the functional split concept in NTN networks, the SI interface must be implemented through the ISLs or NTN feeder links, depending on the considered split architecture. These type of links are less reliable and usually has less capacity than the terrestrial network links for which the functional split has been initially designed, especially when considering the feeder link. As investigated in the previous section, the required capacity changes notably in the different split options. The third column in Table 5.1 reports the maximum capacity needed per split option assuming a gNB with 100 MHz channel bandwidth, 256 QAM modulation with the highest spectral efficiency, corresponding to index 27 of the available ModCod, and 8 MIMO layers. Options 1 to 5 require a maximum data rate of 4 Gbps in DL and 3 Gbps in UL, with option 6 requiring an increased UL data rate of 5 Gbps. In these cases, the maximum data rate is reported even if the actual data rate needed is a function of the traffic required by the users. Increasing the split option to 7 and 8, the requirement increases drastically reaching 157.3 Gbps in UL and DL for option 8. Thus, the available data rate on the CU-DU interface becomes a critical parameter in selecting the optimal functional split option of the specific NTN system.

Considering the *orbit-split* architecture, the SI can be delivered through OISL interconnecting the feeder and service nodes. In the considered constellation design performed in 6G-NTN project deliverable 3.4 "Report on VLEO space segment (1st version)", [6], the distance between feeder and service nodes varies along the orbit between approximately 250km and 800km. State-of-the-art OISL can support a throughput up to 300 Gbps at 2000 Km distance considering a 7cm aperture and 2W of transmit power, [7]. In this case, the provided throughput does not pose a limit to the kind of split that can be implemented, supporting split 8 as well. The OISL still implies lower reliability compared to terrestrial fixed networks due to possible pointing mismatches.

On the other hand, considering the *feeder-split* architecture, the SI is delivered through the feeder link, posing more limitation on the supported functional splits. Table 5.2 shows the feeder link bandwidth and the transmission power necessary to meet the requirements of the different splits in terms of data rate, where the DVB-S2X is used as a radio air interface for the feeder-link and the highest spectral efficiency ModCods are selected. Since the air interface on the SI has not been defined for NTN, we choose the DVB-S2X as it is optimized specifically for satellite data transmissions. With the aim to define the required transmission power to reach the spectral efficiency, a link budget computation has been performed considering the losses defined in [125] for the considered frequency bands. As shown in Table 5.2, a very high transmission power and bandwidths are required in W and Q/V bands to provide a sufficient throughput for split option 7 and 8. Splits 1-6 can be satisfied in Ku and Ka bands, lowering the transmission power and the exploited bandwidth. It is worth highlighting that requirements in Table 5.1 are valid for one gNB. As per 3GPP specification [126], a maximum number of 64 beams for 5G NR is available. However, a satellite might generate an higher number of beams. Thus, if a one to one mapping between the 5G NR beams and the satellite beams is followed, then several DUs should be instantiated onboard the NTN node. Therefore, assuming the maximum traffic load on each DU, the aggregate traffic delivered through the feeder link increases with the number of DUs. Consequently, given an available feeder link data rate, a trade-off is imposed between the functional split option and the number of DUs that can be instantiated onboard.

5.3.2 Interface Latency

RAN functional blocks require tight collaboration among them in order to deliver good quality service. In fact, the communication latency limit between some RAN functions is very stringent, because of functional timers. The total latency comprises the internal computational time of the RAN functions and the PDU transport latency. Deploying the functions at different points of the network means increasing the PDU transport latency component of the delay, making the distance between the network nodes a parameter to be strictly controlled.

The maximum channel delays in each functional split option, as reported for information by 3GPP in TR 38.801 [119], are listed in the second column of table 5.1. Options 1 to 3 have relaxed latency limits since they are upper-bounded only by the maximum RAN-5GC interface delay. In higher functional splits the latency limit is more stringent. Option 4 is limited by the scheduling interval between RLC and MAC. Considering the time for MAC scheduling and for RLC PDU segment concatenation, the maximum latency of option 4 is approximately 100 μs. In options 5 to 8 the most challenging timer to be met refers to the HARQ process. Indeed, it must be completed in maximum $5ms$. Taking into account the processing delays and the transmission time to the UE, the time left for the SI transport delay is approximately 250 μs. In the *orbit-split* architecture, the distance between service and feeder satellites varies considerably in the orbit reaching a maximum of approximately 800km in the designed LEO constellation. A 800km link implies a one way PDU transport latency over the OISL of about 2.5ms that is supported only by splits options 1, 2, and 3.

In the *feeder-split* architecture, the RAN functions are split between the on-ground component and the onboard one. Thus, the SI interface must be implemented through the SRI on the feeder link. In this case, the PDU transport latency coincides with the feeder link channel delay. Figure 5.5 shows the maximum channel delay on the user link varying the elevation angle of the beam center, while Fig. 5.6 provides the one-way channel delay on the feeder link. The minimum elevation angles considered in the user and feeder links are respectively $30°$ and $10°$, as specified in [127]. The

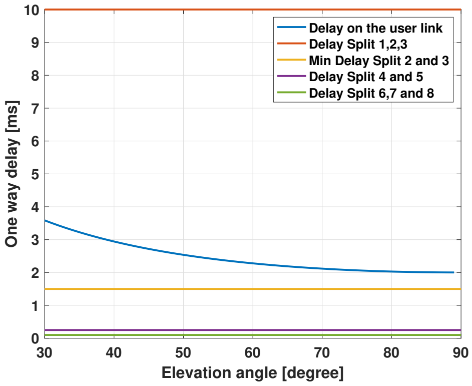


FIGURE 5.5: Maximum one-way delay over the user link varying the elevation angle of the beam center vs the delay of each split. LEO at 600 km of altitude and 50 km of beam diameter.

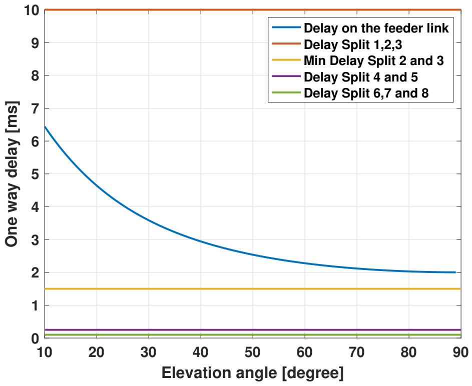


FIGURE 5.6: One-way DELAY on the feeder link varying the elevation angle vs the delay of each split. LEO at 600 km of altitude and 50 km of beam diameter.

graph refers to a scenario with LEO satellites orbiting at 600 Km of altitude and generating beams of 50 km diameter. Comparing the channel delays of our reference scenario with the maximum SI interface delays, it is clear that NTNs impose a real challenge to the implementation of the functional split paradigm. Indeed, options 1 to 4 can support a delay compliant with LEO NTN systems, while higher-level splits impose a latency constraint that can not be met in current systems. In order to enable an efficient implementation of high-level splits, next-generation systems should: *i*) drastically decrease the required computational time to leave more time for the PDU transport latency, or *ii*) relax the latency constraint on the HARQ function. As suggested by 3GPP in [127], the latency constraint on the HARQ function can be relaxed extending the maximum number of parallel HARQ processes from 16 to 32. To further relax the timing requirement, the already existing HARQ processes could be reused. In this case, the same HARQ process is connected to multiple data transmissions. Thus, the HARQ feedback becomes unnecessary and can be disabled.

5.4 AI-enabled Adaptive Functional Split Optimization

In the previous section, we have thoroughly described the main challenges that need to be addressed to complete the first step of the realization of the function split into the NTN component. However, in order to enable the full integration of TN-NTN components, a further level of optimization is needed, which combines resource optimization and Artificial Intelligence in the RAN.

In a static system, the optimal functional split can be already determined in the system design phase. On the contrary, the NTN architecture changes morphology in a highly dynamic way, thus, a dynamic optimization of its functional split could be beneficial. Considering the Feeder-split architectural option, the NTN architecture shall enable a system-aware and proactive functional split optimization. Indeed, the RIC will be in charge of computing the optimal functional split based on status data collected from the network and redeploying the network functions in the CU and DU according to it. One of the main optimization objectives to be investigated is the minimization of the on-board payload energy consumption. Indeed, the satellite has limited available power, this implies that every single watt from batteries and the photovoltaic unit shall be wisely exploited. Moreover, not only the communication payload power is a scarce resource, but also it is not constant in time since it is a function of the satellite's position in the orbit and it depends on the instantaneous power required by the other satellite subsystems. In this framework, an AI application deployed in the RIC will be able to select and implement the optimal functional split while guaranteeing an appropriate QoS. Precisely, shifting RAN functions from the on-board DU to the on-ground CU frees up resources on the payload but increases the required feeder link performances and the latency of the CU-implemented RAN functions. This application operates by collecting data from the network about: *i*) type and volume of requested user traffic; *ii*) payloads computational power capabilities; *iii*) payloads instantaneous available power, and *iv*) the feeder link instantaneous data rate and transport latency.

Additionally, in system configurations with lower restrictions on power consumption, an additional optimization objective is the maximization of the exploitation of the feeder link, following the time-varying behavior of its performances. In this regard, the RIC will exploit the same KPIs of the previous case to proactively select the best-fitting functional split. The most challenging aspect of the dynamic functional split implementation is the high functional flexibility required on the payload. Indeed, that grade of flexibility that can be met by: *i*) relying on general-purpose computing processors, or *ii*) implementing the single RAN functions on specialized and isolated hardware that can be individually activated. The general-purpose computing technology currently available requires a precisely optimized softwarized gNB implementing interfaces for each considered split, [128]. Additionally, the latter case implies high complexity design and poorly exploited payload hardware.

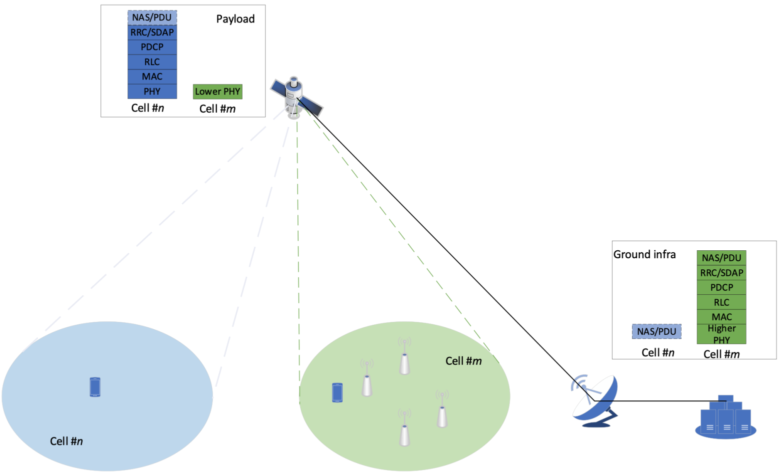


FIGURE 5.7: Cell/Area-Specific AFS

Additionally, it is possible to tailor the Adaptive Functional Split (AFS) implementation to situations where different areas, different UEs, different services, or different services require a different functional split in the same gNB:

- Cell/area-specific AFS: Different functional split options for different cells/areas
- Scenario-specific AFS: Different functional split options in different scenarios
- UE-specific AFS: Different functional split options for different UEs
- Service-specific AFS: Different functional split options for different services

Please note, to illustrate the different AFS options in the rest of this section, a lower layer functional split option, which splits the PHY layer to a lower PHY sublayer and a higher higher-PHY sub-layer, is used as an example, while a higher layer functional split option, which contains the entire gNB protocol layers, is used as another example. However, these options are only used for illustration purposes, and they should not be interpreted as the only options for supporting the proposed AFS.

5.4.1 Cell/Area-Specific AFS

Figure 5.7 shows an example for the cell/area-specific AFS scheme. In this scheme, a satellite may serve different cells or different areas by using different functional split options at the same time. For example, TN and NTN NW may coexist in the area covered by cell m, e.g. along a seashore, which may prefer to deploy a lower layer function split such that more AS protocol layers can be centrally located on the ground, which enables to apply a central scheduling for handling TN-NTN coexistence and lower layer mobility solutions for TN-NTN mobility. In contrast, cell n may cover an area without TN coverage, e.g. in the deep sea. In this case, cell n may benefit from using a higher layer function split, which can help to achieve a lower

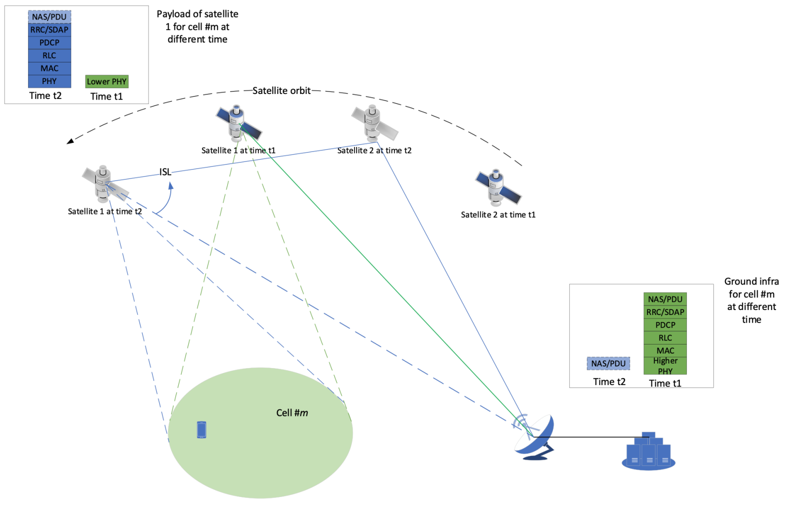


FIGURE 5.8: Scenario-Specific AFS

latency in the AS layer and support onboard MEC in 6G NTN. Please note, cell n and cell m may use the same physical lower PHY entity onboard the satellite, and they are logically separated in Figure 5.7 for illustration purpose only. Please also note, the same note applies for the rest of the figures in this section.

5.4.2 Scenario-Specific AFS

Figure 5.8 illustrates an example for the scenario-specific AFS, where the satellite may determine to adapt its functional split based on the real time scenario, e.g. if an ISL is needed. As shown in this figure, satellite 1 at time t1 may have a direct feeder link connection to the gateway and ground network, and it may apply a lower layer split function. However, afterwards, satellite 1 may move away from the gateway. And at time t2, satellite 1 has to establish an ISL towards another intermediate satellite (e.g. satellite 2) for its connection towards the ground network, since satellite 1 has moved out of the gateway's reachability. In this case, in order to reduce the load posed by the data of satellite 1 on the ISL and/or the feeder link of satellite 2, it may be preferred for satellite 1 to switch from lower layer split to higher layer split.

5.4.3 UE-Specific AFS

Due to the complexity/resource/power constraint at satellite, a satellite may only be able to support some of the UEs with additional onboard protocol layers and computing resource, but the other UEs may only be supported with lower protocol layers onboard the satellite. Thus, in Figure 5.9, a UE-specific AFS scheme is shown, where different function split options may be applied for serving different UEs. For example, UE1 may be consuming a low latency service, but not UE2. Thus, in this example, the satellite may apply a higher layer split for UE1 but not UE2, which allows to use the precious onboard resource in a smart manner by taking account of each UE's specific requirement.

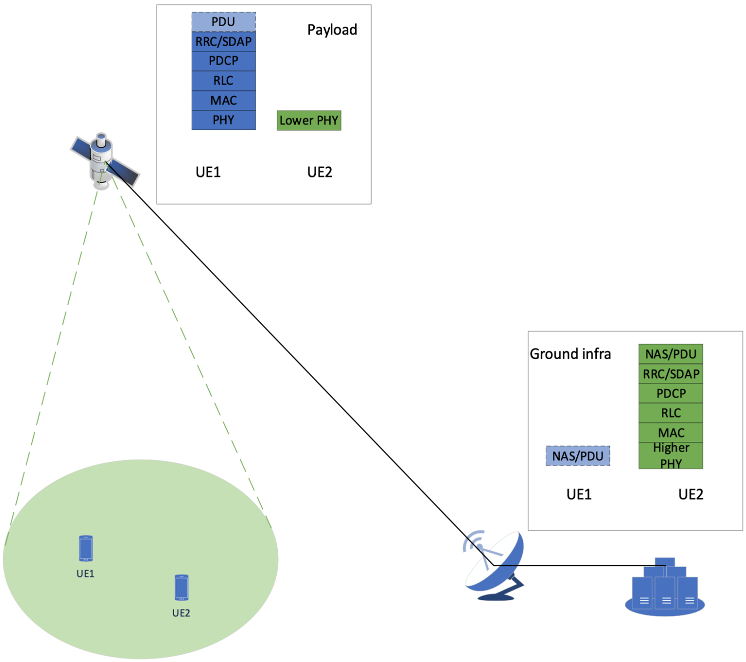


FIGURE 5.9: UE-Specific AFS

5.4.4 Service-Specific AFS

In Figure 5.10, an example for supporting the service-based AFS is illustrated. In this example, the different services or PDU sessions of the considered UE may be served by using different split options. For example, the UE's session requiring an onboard MEC or low latency (e.g. UE1's PDU session m) may be served by a high layer split, but lower layer split may be preferred for serving another session (e.g. UE1's PDU session n) and the control plane of the UE to save the precious onboard resources. This scheme provides a finest granularity level for NTN NW to adapt its functional split function, based on the UE's service-specific requirements.

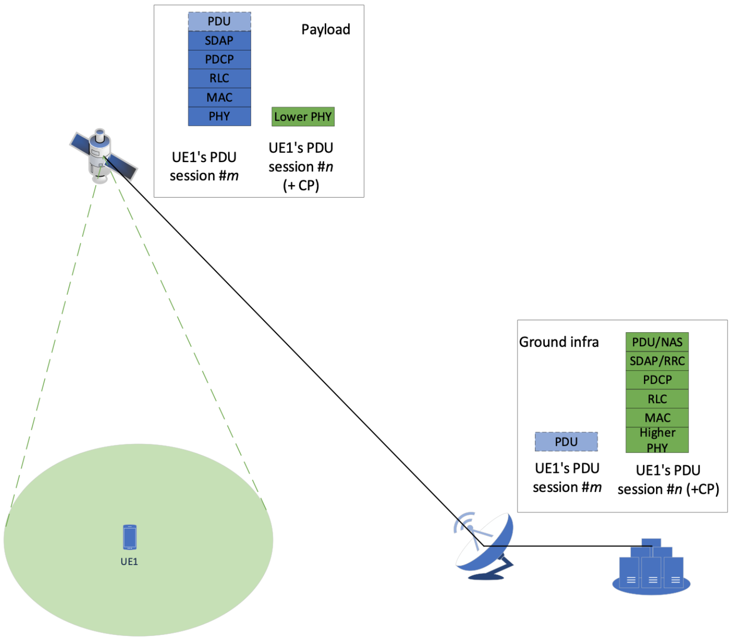


FIGURE 5.10: Service-Specific AFS

TABLE 5.1: Overview of delay and bandwidth requirements, pros and cons of functional splits, informative data from [20].

Split option	Required interface BW	Interface bandwidth scaling	Maximum interface latency	Latency constraint motivation	Split benefit
1: RRC/PDCP	4/3 Gbps	Traffic dependent	10ms	Max latency between gNB and CN	Lowest interface requirement.
2: PDCP/RLC		Traffic dependent. Scaling: MIMO layers	1.5 - 10ms	Depends on the buffers size	Data closer to the user.
3: Intra RLC	<option 2				
4: RLC/MAC	4/3 Gbps		$\sim 100\mu s$	Depends on MAC sched. and RLC PDU comp. time	No specific benefit.
5: Intra MAC			Hundreds of μs	Intra MAC signaling time requirement	Coordinated multi-point (CoMP) schemes.
6: MAC/PHY	4.1/5.6 Gbps		250 μs	Limited by HARQ time	Resource pooling for layers including and above MAC.
7c: Intra PHY	22.2/21.6 Gbps				Transmit and receive joint processing is possible. CoMP schemes.
7b: Intra PHY	81,6/81,6 Gbps				
7a: Intra PHY	22,2/86,1 Gbps	Traffic independent. Scaling: antenna ports			
8: PHY/RF	157,3/157,3 Gbps				Small, cost-effective RU. Highest level of centralization. Inter-RAT RUs exploitation.

TABLE 5.2: Link budget and capacity analyses

	W	Q/V	Ku	Ka
	83.5 GHz	47 GHz	28.75 GHz	17 GHz
Target Modulation and Coding Scheme (ModCod)	256 APSK 3/4			
Spectral efficiency	5.9 bit/s/Hz			
Target C/N	24.02 dB			
Bandwidth	22 GHz		1 GHz	
Target C/N0	112.4 dBHz			
Required Ptx	59.8 dBW	34.8 dBW	6.07 dBW	0.17 dBW
Capacity	156.4 Gbit/s		7.81 Gbit/s	

Chapter 6

O-RAN based NTN RAN Optimization

This chapter aims to explore the motivations and contributions of RIC-based RAN optimization, focusing on its critical role in NTNs and their integration with terrestrial networks. The motivations arise from the growing complexity of NTNs, characterized by dynamic topologies, resource limitations, and variable latency, which require intelligent, adaptive approaches to ensure efficient and reliable network performance. The chapter explores TN-NTN traffic offloading as another cornerstone of RIC-based optimization. By leveraging Non-RT control functionalities, networks can analyze long-term traffic patterns and optimize resource allocation across terrestrial and non-terrestrial domains. This approach ensures balanced load distribution and maximizes the utilization of available resources, addressing the unique challenges posed by NTNs. In addition to traffic offloading, the chapter emphasizes the significance of real-time control loops and underline their importance in the context of channel prediction for 6G-NTN Next Generation Multiple Access (NGMA). These ML-based control loops can predict channel conditions dynamically, enabling the network to adapt promptly to environmental changes. This capability enhances spectrum efficiency and ensures robust communication under the stringent demands of 6G NTNs. Finally, this chapter explores how Cloud-Native Network Function (CNF) orchestration enhances control loop functionality, leveraging AI to enable proactive resource management and scalable, resilient network operations tailored for 6G.

6.1 Control Loops

6.1.1 non-RT and near-RT Control Loops

In the ever-evolving landscape of telecommunications, the transition toward 6G and the incorporation of NTNs are reshaping the boundaries of connectivity. NTNs, which encompass satellite networks, unmanned aerial systems, and high-altitude platforms, address critical challenges in providing global, ubiquitous coverage. These networks extend connectivity to remote and underserved areas, enable disaster recovery, and support a wide range of applications requiring seamless communication beyond the reach of terrestrial infrastructure. However, NTNs introduce significant complexities due to their dynamic topologies, resource constraints, and variable latency conditions. Against this backdrop, the optimization of RAN becomes an indispensable element in ensuring robust performance and efficient resource utilization.

RIC-based RAN optimization has emerged as a transformative approach in addressing the unique challenges posed by NTNs. By introducing programmable intelligence into the RAN through the Near-RT and Non-RT RIC architecture, networks gain the ability to adapt dynamically to fluctuating conditions. RIC not only provides a platform for deploying advanced optimization algorithms but also facilitates the integration of artificial AI to enable proactive and efficient network management. This approach aligns seamlessly with the demands of NTNs, where maintaining consistent service quality amidst shifting operational parameters requires intelligence-driven decision-making.

RIC empowers NTNs to tackle some of their most pressing challenges, such as load balancing and traffic steering. The dynamic nature of NTN traffic patterns necessitates intelligent mechanisms to alleviate congestion and ensure equitable resource distribution. By leveraging AI-driven models and advanced learning frameworks, RIC enables networks to predict traffic fluctuations and redirect users to underutilized cells or frequency bands, maintaining consistent performance. Techniques such as those discussed in [129] and [130] highlight the effectiveness of programmable RIC-enabled solutions in dynamically managing traffic demands.

Another critical aspect of RAN optimization is ensuring QoS, which becomes particularly complex in high-latency NTN environments. RIC-based solutions integrate AI models capable of analyzing real-time quality metrics and predicting service demands. These insights allow the network to allocate resources dynamically, prioritizing essential services such as emergency communications and high-definition video streaming. The relevance of such techniques is emphasized in studies like [131].

Interference management presents yet another challenge in NTNs, where overlapping signals can degrade communication quality. Through RIC, advanced algorithms for interference mitigation are deployed, enabling the network to dynamically identify and address sources of interference. Techniques such as graph-based optimization and frequency-domain coordination, as outlined in [132] and [133], enhance signal integrity and reduce service disruptions, ensuring a seamless user experience.

Similarly, the allocation of radio resources is optimized through RIC by leveraging machine learning models that adapt to real-time network conditions. These AI-driven schedulers maximize spectral efficiency, an essential requirement in NTNs where bandwidth is often limited and must be shared across diverse applications. Studies such as [134] and [135] demonstrate the potential of RIC-based frameworks to optimize spectral efficiency.

Mobility management in NTNs also benefits significantly from RIC-based optimization. In high-mobility scenarios, such as those involving satellites or drones, seamless connectivity relies on efficient handover mechanisms. Predictive machine learning models embedded within the RIC framework anticipate mobility patterns, enabling the network to proactively adjust handover parameters. This reduces connection drops and ensures uninterrupted service, a critical factor for applications demanding high reliability. Studies like [136] and [137] provide a comprehensive overview of these advancements.

Beyond individual optimization functions, RIC facilitates dynamic placement of network functions, optimizing the use of onboard resources in NTNs. By deploying computational capabilities closer to demand hotspots, RIC minimizes latency and improves network responsiveness. This dynamic placement strategy is particularly advantageous in satellite and aerial networks, where resource constraints require careful management to balance performance and efficiency. Research such as [138] exemplifies how these methods enhance network performance.

Furthermore, RIC supports advanced network slicing mechanisms, allowing NTNs to create virtualized partitions tailored to specific applications. This capability ensures that resources are allocated according to the varying demands of services such as IoT, ultra Reliable Low Latency Communication (uRLLC), and enhanced mobile broadband. The adaptability of RIC also extends to sustainability, as energy-efficient solutions are integrated into its optimization framework. AI-driven power management functions, such as adaptive sleep modes and real-time power allocation, reduce energy consumption, aligning NTN operations with global sustainability goals, as demonstrated in [139].

Dynamic spectrum sharing is another critical function enabled by RIC. NTNs often operate in environments where spectrum resources are limited and must coexist with terrestrial networks. RIC-based solutions employ AI for real-time spectrum sensing and dynamic allocation, facilitating efficient use of shared frequency bands. This capability not only enhances spectral efficiency but also ensures harmonious integration between NTNs and terrestrial infrastructure, as highlighted in [140].

In summary, RIC-based RAN optimization represents a paradigm shift in the management of NTNs, enabling intelligent, flexible, and efficient solutions to overcome the inherent challenges of these networks. By embedding AI-driven capabilities within the RAN, RIC provides the foundation for resilient and adaptable communication systems that meet the evolving demands of modern connectivity. This chapter delves into the potential of RIC-based RAN optimization, presenting two novel algorithms designed to harness the power of RIC in transforming NTN operations.

6.1.2 Real Time Control Loops

The rapid evolution of NTNs, driven by escalating demands for ultra-low latency, global coverage, and seamless integration with terrestrial infrastructure, has necessitated profound changes in the way control loops operate within these networks. These control loops are pivotal for maintaining network efficiency and include Near-RT and Non-RT mechanisms, which enable decision-making and optimization tasks across varying timescales, particularly crucial for NTNs where delays caused by satellite communication links must be minimized.Near-RT control loops typically operate within the 10-millisecond range, addressing tasks such as resource allocation and traffic management, while Non-RT loops handle longer-term planning and strategy—a critical consideration for NTNs where dynamic adjustments are needed to manage satellite orbits, handovers, and variable link conditions. However, the current framework struggles to address scenarios requiring real-time decision-making at sub-10-millisecond intervals, exposing critical gaps in achieving uRLLC, especially in NTNs where the inherent latency of satellite links complicates achieving stringent timing requirements. This limitation is particularly evident in NTN applications such as dynamic beam steering, inter-satellite link management, and real-time traffic routing, underscoring the inadequacy of existing near-RT and non-real-time non-RT control loops for these tasks.

In practical scenarios, network functionalities such as user scheduling and beam management necessitate precise and rapid control. For instance, user scheduling for uRLLC requires sub-millisecond responsiveness to ensure packet preemption and timely data delivery, [47]. Similarly, beam management relies on processes such as beam sweeping and reference signal transmission, often requiring timescales well below 10 milliseconds. Despite their potential, xApps and rApps in Near-RT RICs face challenges in executing these operations. The latency associated with data transmission to and from the RIC, combined with restricted access to granular, low-layer data such as I/Q samples and transmission queues, imposes fundamental constraints on their suitability for real-time control. Additionally, privacy concerns and the risk of excessive overhead when transferring sensitive user-plane data further complicate their deployment.

To address these challenges, the introduction of distributed applications (dApps) within the O-RAN ecosystem emerges as a groundbreaking development. As proposed in the paper [141], dApps represent a specific implementation of real-time control loops tailored to operate directly at the CUs and DUs. This design significantly reduces latency, bypasses the limitations of the Near-RT RIC, and enables immediate access to critical network information. By leveraging edge-based intelligence, dApps exemplify how real-time control loop principles can be practically applied within the O-RAN framework, bridging the gap between existing capabilities and the demands of real-time operations.

The advantages of dApps are multifaceted. First and foremost, they address latency and overhead concerns by eliminating the need for data transfer over the E2 interface. Operations traditionally executed at the RIC can now be performed locally at the DU or CU, ensuring faster response times and reduced control-plane traffic. Moreover, advancements in edge computing and AI have made it feasible to deploy lightweight, resource-efficient machine learning models at the edge, enabling real-time inference and decision-making. This capability is crucial for implementing sophisticated algorithms for tasks like beamforming, modulation recognition, and RAN slicing, which require granular control and rapid adaptability to changing network conditions.

Another critical benefit of dApps lies in their ability to access and process data that is otherwise unavailable or impractical to handle in a centralized manner. By executing directly at the DUs and CUs, dApps can utilize real-time I/Q samples, mobility data, and other low-layer metrics to fine-tune network performance. This granular access allows for highly tailored solutions that align with specific user requirements and instantaneous network states, enhancing overall QoS and user experience. Additionally, dApps enable extensibility and reconfigurability within the network, supporting seamless updates and integration of new functionalities through containerized deployments.

Use cases and applications of dApps span a wide array of scenarios, each benefiting from the unique attributes of real-time control and edge intelligence. In beam management, for example, dApps can implement advanced algorithms for beam selection and optimization, such as DeepBeam, which leverages deep learning models to analyze I/Q samples for precise angle-of-arrival and beam classification [142]. This capability is essential for reducing latency and overhead while maintaining privacy and security standards. Similarly, in RAN slicing and scheduling, dApps can dynamically allocate resources and prioritize traffic based on real-time conditions and forecasted demands, ensuring optimal performance for uRLLC, eMBB, and mMTC traffic types.

While the adoption of dApps promises significant advancements, it also presents challenges that must be addressed for successful deployment. Resource management at the edge is a critical consideration, as dApps require sufficient computational power to handle concurrent operations without compromising performance, [143]. Additionally, the development of standardized interfaces for seamless interaction between dApps, DUs, CUs, and other O-RAN components is essential to ensure interoperability and platform independence. Furthermore, orchestration mechanisms are needed to determine the optimal distribution of intelligence between dApps and xApps, balancing factors such as data availability, control timescales, and network workload.

In conclusion, the introduction of dApps marks a pivotal step in the evolution of the O-RAN architecture, enabling real-time inference and control that was previously unattainable with xApps and rApps alone. By extending intelligence to the edge, dApps unlock new possibilities for network optimization, delivering unparalleled performance and adaptability in an increasingly complex and demanding connectivity landscape. This innovation not only addresses the limitations of current O-RAN implementations but also sets the stage for a future where cellular networks achieve true self-organization and optimization.

6.2 TN-NTN Traffic Offloading

This section reports the outcomes of the 6G-NTN project deliverable 4.2 "Report on 6G-NTN radio controller (1st version)", [11], introducing innovative strategies to enhance network performance by distributing traffic intelligently between terrestrial networks and non-terrestrial networks. This represents a perfect example of optimization algorithm that is enabled by the closed-loop control of Near-RT RICs. Indeed, by leveraging reinforcement learning techniques, the approach aims to balance network load and proactively manage traffic, ensuring efficient resource allocation and improved service quality. This integrated architecture, supported by simulation-based insights, demonstrates how advanced optimization frameworks can maximize throughput and resource utilization while addressing the unique challenges of TN-NTN environments.

6.2.1 State-of-the-art

The optimization of traffic off-loading is crucial for enhancing the performance of cellular networks. The traffic off-loading task is often addressed with a load balancing approach that involves the optimal allocation of resources, management of handovers, interference, and balancing the load between multiple cells and carriers, [130]. This is especially challenging in the context of 5G and 6G networks, where the efficient allocation of radio resources is essential. To address these challenges, various innovative approaches have been proposed. One approach involves the use of ML and AI based load balancing algorithms running on Open RAN RIC (RAN Intelligent Controller). In this context, [144] discusses the implementation of the traffic steering xApp in a hierarchical and modular fashion. The results presented emphasize the advantages of this modular approach, illustrating how AI/ML tools can intelligently manage the functioning of the xApp and consequently enhance overall system performance. The authors in [145] leverage DRL and Graph Neural Networks (GNN) to achieve up to 10% gain in throughput, 45-140% gain in cell coverage and 20-45% gain in load balancing compared to baseline greedy techniques. Additionally, proactive approaches using multi-agent reinforcement learning and deep deterministic reinforcement learning algorithms have been explored to maximize UE throughput and improve handover decisions based on local channel measurements. Furthermore, the use of GNN and deep Q-learning approaches have been proposed to learn Q-functions from cell and UE deployment instances, providing scalable and effective solutions for connection management. In [146], an algorithm for RAT allocation predicated on Federated Meta-Learning (FML) is introduced, thereby facilitating the expeditious adaptation of RICs to dynamically evolving environments. A simulation environment has been devised, featuring LTE and 5G NR service technologies. Within this simulation framework, the primary aim is to satisfy UE demands within prescribed transmission deadlines, thereby enhancing the delivery of heightened QoS values. To enhance the collection of realistic data for xApp training, in [129], ns-O-RAN is introduced. It is a software framework that integrates a real-world, production-grade Near-RT RIC with a 3GPP-based simulated environment on ns-3, enabling at the same time the development of xApps and automated large-scale data collection and testing of DRL-driven control policies for the optimization at the user-level. Several works focus on the optimization of the load in small-cells deployments. In [147], a load balancing technique based on network segmentation and adaptive sleep scheduling for 5G-IoT networks is proposed. This technique involves the formation and grouping of sub-segments for each network segment to process IoT applications with different QoS requirements. Additionally, adaptive dynamic sleep scheduling is executed by each small cell base station based on its load level. The load balancing policy involves the transfer of overloaded traffic from small cell base stations to the macro cell when the average load of any small cell base station surpasses that of the macro cell. Simulation results validate the proposed technique, demonstrating higher success probability, power efficiency, reduced energy consumption, and packet drops of small cell base stations compared to existing techniques. In addition, the research by Addali, [148], [149], focuses on addressing the challenges of unbalanced load distribution in 5G small-cell networks through the introduction of a Utility-based Mobility Load Balancing algorithm Utility-based Mobility Load Balancing algorithm (UMLB) and a Load Balancing Efficiency Factor Load Balancing Efficiency Factor (LBEF). The UMLB algorithm considers both operator and user utility, aiming to achieve efficient load transfer from overloaded cells to under-loaded neighboring cells. The study emphasizes the importance of accurately adjusting handover parameters to prevent inefficient resource usage and degradation of QoS. Additionally, the research discusses the impact of user mobility on the algorithm and evaluates the performance through computer simulations. The proposed algorithm demonstrates promising results in improving network performance and user satisfaction by effectively balancing the load distribution in small-cell networks. In [150], an extreme Swap-based Load Balancing (SLB) algorithm between APs, which minimizes the load imbalance at cell edges is proposed. The experimental setup uses a dataset contributed by Irish mobile operators. Recent literature has explored the application of controller and machine learning algorithms to assist self-optimizing and proactive schemes in making load balancing decisions. However, the authors in [151] argue that these algorithms often lack the ability to forecast upcoming high traffic demands, particularly during popular events, leading to cold-start problems and low convergence speed in hotspots with skewed load distribution. To address these challenges, three contributions are proposed. Firstly, it introduces urban event detection using Twitter data to forecast changes in cellular hotspots, enabling context-awareness. Secondly, it simulates a proactive 5G load balancing strategy considering the prediction of skewed-distributed hotspots in urban areas. Finally, it optimizes this context-aware proactive load balancing strategy by forecasting the best activation time. Research work has timidly focused also on load balancing for non-terrestrial networks. In [152], the authors propose a novel load balancing algorithm designed specifically for a multi-RAT (radio access technology) network that encompasses both NTN and TN. To address this gap, the authors introduce the concept of a Radio Resource Utilization Ratio (RRUR) as a common load metric to assess the cell load of each RAT. This metric is utilized in conjunction with an adaptive threshold to identify overloaded cells. The proposed algorithm comprises two key steps: intra-RAT load balancing and inter-RAT load balancing. In the first step, the algorithm redistributes the load by offloading edge UEs from overloaded cells to neighboring underutilized cells based on the RRUR of each cell. If a cell's RRUR remains above a predefined threshold, the algorithm proceeds to the second step, which involves offloading delay-tolerant data flows of UEs to a satellite link as a form of inter-RAT load balancing. Additionally, the algorithm incorporates an estimation of the impact of load redistribution on the target cell load to minimize unnecessary load balancing actions. The effectiveness of the proposed algorithm is demonstrated through simulation results, which indicate that it not only achieves more even load distribution across terrestrial cells but also enhances network throughput and the number of quality-of-service satisfied UEs compared to previous load balancing algorithms. Additionally, in [153] the challenge of efficiently utilizing limited beam resources to serve users in NGSO communication systems are addressed. To enhance spectrum utilization, the authors propose a multi-satellite beam hopping algorithm that integrates load balancing and interference avoidance. The algorithm decomposes the multi-satellite beam hopping problem into three subproblems: multi-satellite load balancing, single-satellite beam hopping pattern design, and multi-satellite interference avoidance. Simulation results demonstrate that the proposed method significantly reduces the load gap among satellites and improves the average traffic satisfaction rate. Additionally, it exhibits superior performance in terms of unmet capacity compared to other benchmarks, thereby achieving better alignment between offered and requested data. Finally, [154] introduces a decentralized Load Balancing Satellite Handover (LBSH) strategy based on multi-agent reinforcement Q-learning, implemented within the Network Simulator 3 (NS-3) software. The LBSH strategy aims to minimize the total number of HOs and the blocking rate while ensuring a balanced distribution of load among satellites. The results demonstrate the superiority of the proposed LBSH method over existing approaches, with a significant reduction in the average number of HOs per user and the blocking rate, thereby showcasing its potential to optimize satellite handover management and enhance network performance in LEO satellite constellations within the context of 5G and beyond technologies.

6.2.2 Architecture

Modern mobile cellular networks rely on a dense deployment of cells to accommodate a growing and demanding user base. However, the mobility of UE often results in uneven load distribution among network cells, leading to degraded QoS and sub-optimal resource utilization. This challenge is further exacerbated by the increasing demand for high data rates and the non-uniform spatial distribution of UEs, which causes resource overloading in specific cells. Addressing these issues necessitates effective load-balancing strategies to optimize resource allocation and maintain QoS [152]. TN employ various cell-load balancing techniques to distribute traffic evenly across cells, ensuring efficient operations and enhancing the user experience. However, in scenarios where UEs cannot transition to neighboring cells due to resource constraints or limited coverage, these mechanisms are less effective. Integrating NTN introduces a complementary dimension for load balancing, enabling traffic redistribution not only within TN but also between TN and NTN. This dual-layer optimization enhances TN QoS by offloading traffic to NTN, thereby improving overall network throughput and resource efficiency.

Optimally assigning UEs to either TN or NTN is a complex task that must be managed by a centralized entity. This process requires real-time network status information gathered from UEs and cells. In rural TN deployments, traffic patterns fluctuate significantly throughout the day, leading to overburdened base stations during peak times and under-utilization during off-peak hours. Incorporating NTN in such environments allows traffic offloading, mitigating overloads during peak periods and reducing energy consumption during low-demand periods. Indeed, TNs in rural areas often face fluctuating user traffic throughout the day, leading to high base station loads during peak hours and under-utilization during off-peak periods. In scenarios where TNs are overloaded, an AI/ML component can analyze the network's status and determine the optimal amount of traffic to offload to NTNs. The AI/ML's primary objective is to maximize the overall Signal-to-Interference plus Noise Ratio (SINR) of the TN while considering the capacity constraints and high variability of NTNs. Given NTNs' limited throughput and higher latency, not all users are suitable for offloading. Therefore, the AI must predict future traffic patterns and assess NTN capabilities to ensure only users with traffic requirements compatible with NTN performance are handed over.

The proposed architecture for traffic off-loading consists of terrestrial TN cells and an NTN node providing overlapping service coverage. Both TN and NTN share a common AI/ML server and core network, allowing seamless integration. Figure 6.1 (a) illustrates this architecture. The majority of data required for AI/ML training and inference originates from the ground network, making on-ground AI/ML server deployment the most efficient choice. This reduces latency and transport network load while ensuring resource availability for computationally intensive tasks.

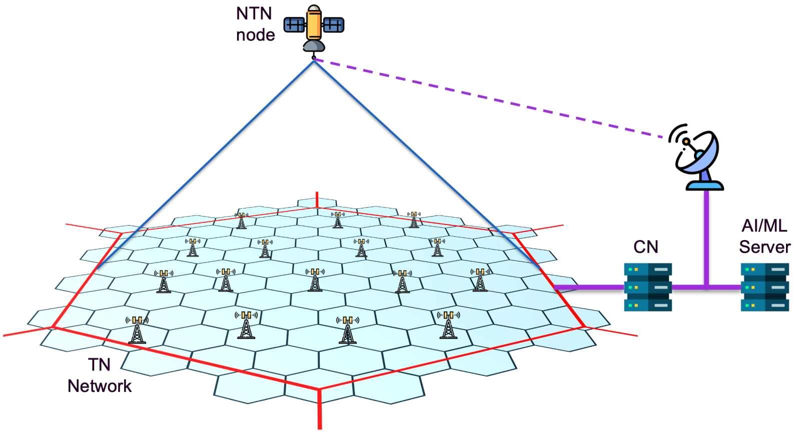


FIGURE 6.1: Reference architecture for Traffic Off-Loading

The traffic off-loading manages AI/ML processes as follows:

6.2.2.1 Data Collection

AI/ML servers collect and process diverse datasets required for training and inference to support load balancing operations. Table 6.1 provides an overview of the input data sources, their usage, and the network that is used to distribute the data. When determining the location of the AI/ML server, it is important to consider the load on specific parts of the distribution network. This ensures that the server placement aligns with the network's operational demands, enabling efficient processing and minimizing potential bottlenecks.

TABLE 6.1: INPUT DATA TO PERFORM MODEL TRAINING AND INFERENCE IN TRAFFIC OFF-LOADING

Collected Data	Source	Distribution network	Usage
UEs SINR meas.	UE	Ground	Training, inference
UE location	UE	Ground	Training, inference
UE service	UE	Ground	Training, inference
TN cell res. allocation	TN	Ground	Training, inference
NTN cell res. allocation	NTN	FL	Training, inference
Sat. ephemeris	NTN	FL	Training, inference
UE allocation to NTN	AI-ML	FL	Output from inference

6.2.2.2 Model Training

Model training is suitable to perform on an on-ground AI/ML server to leverage localized datasets and optimize the model for specific deployment scenarios. The training process includes:

- Location-Specific Training: Models may be trained using data from their respective regions to adapt to local traffic patterns and environmental conditions.
- Federated Learning: A global model can be achieved by aggregating insights from multiple localized models, ensuring broader applicability while maintaining regional relevance.
- Online Updates: Dynamic scenarios require models to adapt to evolving conditions. Initial models, trained offline, can be incrementally updated in real-time to accommodate changes in traffic and network status.
- The AI/ML server can perform offline training to develop an initial AI/ML model using data collected from the network.

6.2.2.3 Model Storage

AI/ML models may store in geographically appropriate servers:

- Location-specific models reside in servers near the deployment region.
- Global models, obtained through federated learning, are stored in centralized repositories for broader applicability.

6.2.2.4 Model Inference

Model inference is suitable to execute in on-ground AI/ML server, aligning with the ground-based origin of most input data and the computational requirements of inference tasks. By centralizing inference operations, latency can be minimized, and resource utilization can be optimized.

6.2.3 System Description

The system model considered for the traffic off-loading optimization technique is presented in this section. We consider a rural area served by a uniform deployment of terrestrial 6G Base Station (BS). As recommended in 3GPP TR 38.901, for the considered Rural Macro (RMa) scenario we assume a hexagonal grid layout of $N_{BS} = 19$ macro sites deployed at an Inter-Site Distance (ISD) of 5000 m. Each macro site has a height h_{BS} of 35 m and three sectors per site. All BSs reuse the same terrestrial frequency band B_{TN}. The same rural area is assumed to be served by a 6G NTN beam on an orthogonal band B_{NTN} of the same width.

The users are uniformly distributed in the reference area and generate uplink traffic assuming a full buffer model, meaning that every user in the network is continuously active, generating data traffic at the maximum possible rate. The number of UEs in the coverage area is configured to ensure that the requested user traffic exceeds the terrestrial system capacity, meaning that, at a specific instant of time, only a subset UE_S of users is served while users UE_W are waiting to be served. Additionally, all the radio resources of each BS are continuously assigned to the served users. In this network function, the radio resources are considered to have already been scheduled to the users assuming that: i) in the radio resource scheduling phase the served UEs are assigned the same amount of radio resources, and ii) each UE is allocated to the BS that has the strongest received power. The allocation of a user to the corresponding BS is represented by the set $U_A = u_A^1, \ldots, u_A^j, \ldots, u_A^(N_u)$, N_u being the cardinality of $UE_W \cup UE_S$, and $u_A^j = 1, \ldots, N_BS$.

In this context, part of the exceeding load of the TN is assumed to be offloaded to the NTN. The traffic offloading can happen in two ways:

1. Part of the users UE_W, that are waiting to be served by the TN, are served by the NTN on the orthogonal band B_{NTN}. In this case, the NTN serves users that are not assigned to any TN radio resource and, thus, the TN scheduling process is not impacted by the NTN service. The system throughput T_{TOT} (TN+NTN) is expressed as:

$$T_{TOT} = \sum_{BS} B_{TN}\; log_2(1 + SINR_{BS}) + T_{NTN}$$

   With $SINR_{BS}$ being the Signal to Interference plus Noise Ratio of the specific BS, and T_{NTN} being the throughput of the NTN service on the band B_{NTN}.

2. The network function is free to select the UEs to be served by the NTN not only among the subset UE_W, but also among the subset UE_S that is already served by the TN. If a UE connected to BS_k is offloaded to the NTN, then, a UE that belongs to UE_W and that is in the coverage area of BS_k is connected to BS_k in the same radio resources of the previous UE.

6.2.3.1 Problem Formulation

Leveraging on a broad and centralized view on the network, the traffic off-loading network function aims at increasing the system throughput compared to the legacy UE allocation described in case 1. To this regard, the network function can optimize the system throughput by selecting which UEs to offload from the TN to the NTN, aiming to maximize the SINR at the base stations. We define the optimization problem as the selection of W_A that satisfies the objective function P, where $W_A = \{w_A^1, \ldots, w_A^j, \ldots, w_A^{N_u}\}, w_A^j \in \{0, 1\}$ is a set that identifies the allocation of the user j to the NTN.

$$P = \max_{W_A} \sum_{BS} B_{TN} log_2\left(1 + SINR_{BS}\right) +\; T_{NTN}$$

$$C1: \sum_{j=1}^{N_u} w_A^j \leq C_{ntn}$$

$$C2: \sum_j w_A^j = \sum_j s_U^j, \; j|UE_j \in BS_k$$

The constraint in C1 denotes that the number of UEs that are offloaded must be lower or equal to the capacity C_{ntn} of the NTN. The constraint in C2 imposes that the number of UEs that are offloaded from the base station BS_k to NTN must be equal to the number of new UEs that connect to BS_k after the offloading operation. Indeed, $S_U = \{s_U^1, \ldots, s_U^j, \ldots, s_U^{N_u}\}, \; s_U^j \in \{0,1\}$ is a set that identifies the connection of a user UE_j to the TN after the offloading of the set W_A.

6.2.3.2 Optimization Framework

The optimization technique selected to implement this network function is based on Deep Q-Learning (DQL). DQL is an advanced Reinforcement Learning algorithm that leverages the power of deep neural networks to solve problems with high-dimensional state spaces, such as the traffic off-loading network function that is being evaluated. This approach extends the classic Q-Learning algorithm by using a neural network to approximate the Q-value function, which predicts the expected cumulative reward for taking a given action in a given state and following the optimal policy thereafter. DQL has been selected for this network function because its learning process involves an agent interacting with an environment and does not require large datasets. The agent perceives the environment's state and takes actions that affect the state, receiving rewards as feedback. The state space $\mathcal{S}$ represents all possible situations the agent can encounter. The action space $\mathcal{A}$ includes all possible actions the agent can take. These actions might be discrete or continuous. The reward function R(s,a) provides feedback from the environment, signaling how good or bad the action taken by the agent was in a specific state. This reward is used to guide the learning process, pushing the agent towards actions that maximize cumulative rewards. The Q-function Q(s,a) estimates the expected cumulative reward of taking action a in state s and following the optimal policy thereafter. The policy π is the strategy the agent uses to decide its actions based on the current state. The DQL framework is leveraged in this optimization function to address the high dimensionality of the considered environment. Specifically, in this first implementation, the DQL has a state space $\mathcal{S}$ composed by: i) the allocation of the UEs to the base stations U_A, ii) the signal strength of each UE, received from the base station they are allocated to ($SS = \{ss^1|_{BS_k}, \ldots, ss^j|_{BS_k}, \ldots, ss^{N_u}|_{BS_k}\}$, with $k = 1, 2, ..., N_{BS}$). As a second step, it is foreseen to evolve from the full-buffer traffic model and to consider a more detailed one. In this case, also the traffic state of each UE could be considered in the state space. The action space $\mathcal{A}$ is the set W_A that identifies the allocation of the UEs to the NTN. The reward is computed to reflect the optimization objective of maximizing the system throughput, as the throughput gain obtained by applying the selected action:

$$R(s_n, a) = \sum_{BS} B_{TN} log_2(1 + SINR_{BS}^{S_n}) - \sum_{BS} B_{TN} log_2(1 + SINR_{BS}^{S_{n+1}|_a})$$

6.2.4 Numerical Results

The simulation reflects the system description implementing the optimization framework and defining the scenario with which the optimizer interacts. As described in section Section 6.2.3, for the considered RMa scenario we assume a hexagonal grid layout of 19 macro sites with three sectors per site. In the complete scenario, the network function would need to optimize the off-loading of all the UEs in the satellite coverage area considering the radio resources they are allocated to. Since full-buffer traffic model is assumed and since the UEs are assigned with the same amount of radio resources, the problem is symmetric in the radio resources space. This means that it is possible to focus the optimization problem on a single radio resource without losing generality. For this reason, for the group UE_S, only N_{BS} users are considered, one for each base station. Similarly, other N_{BS} users are selected within UE_W such that each UE is associated to a different base station. The first group is represented by the red markers in Figure 6.2, while the second group is represented by the blue markers. At each iteration of the simulation the two groups of users are dropped in a random position that reflects the UE-BS allocation. Taking as example the base station 1 in the center of the area, the served UE (red) and the waiting UE (blue) are randomly deployed in the coverage area of BS1 marked by the red polygon. As mentioned above, the DQL decides the optimal UE to be offloaded based on the state space $\mathcal{S}$. The radio resource of the offloaded UE is allocated to the waiting UE present in the same BS coverage area.

The simulation reflects the configuration parameters from 3GPP TR38.901, [155], regarding the scenario and the path loss model. The relevant information about the base station antenna model is reported in Table 6.2. 3GPP BS ANTENNA MODEL from 3GPP TR 38.921 [156]. For the UE antenna a gain of 0 dB is considered in line with Rep. ITU-R M.2412-0. Additionally, a BS receiver noise figure of 5 dB and a thermal noise level of -174 dBm/Hz are considered.

The path loss model for the rural macro area selected from TR 38.901 is expressed as:

$$PL_{\text{RMa-LOS}} = \begin{cases} PL_1, & 10m \leq d_{2D} \leq d_{\text{BP}} \\ PL_2, & d_{\text{BP}} \leq d_{2D} \leq 10km \end{cases}$$

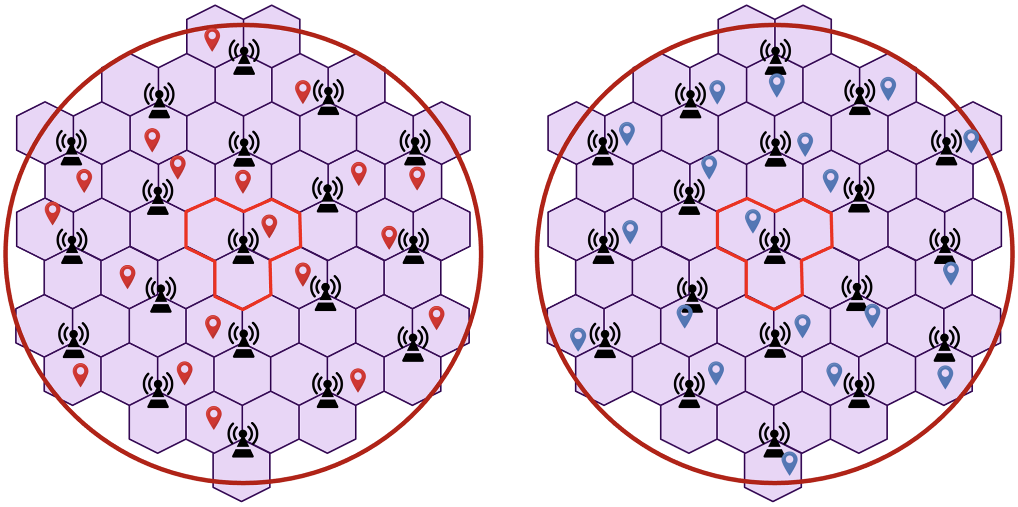


FIGURE 6.2: Allocation of served (red) and waiting (blue) user equpment

A_m (dB)	**30**
SLA_V (dB)	**30**
ϕ_{3dB} (deg.)	90
θ_{3dB} (deg.)	90
$G_{E,\max}$ (dBi)	5.5
L_E (dB)	2.0
(M, N)	(16, 8)
Number of supported polarizations, P	2
d_h (m)	0.5λ
d_v (m)	0.5λ
Horizontal coverage range (deg.)	± 60
Vertical coverage range (deg.)	90 to 120

TABLE 6.2: 3GPP BS antenna model, [156]

$$PL_1 = 20\log_{10}(40\pi d_{3D} f_c/3) + \min(0.03h^{1.72}, 10)\log_{10}(d_{3D}) \\ - \min(0.044h^{1.72}, 14.77) + 0.002\log_{10}(h)d_{3D}$$

$$PL_2 = PL_1(d_{\text{BP}}) + 40\log_{10}(d_{3D}/d_{\text{BP}})$$

Where $h = 5m$, $W = 20m$, and the UE are considered to be always in line of sight.

Initial results are reported in Figure 6.3, Figure 6.4 and Figure 6.5. Figure 6.3 shows the cumulative density function of the system spectral efficiency after the optimal allocation of the user to the NTN. The system spectral efficiency is computed

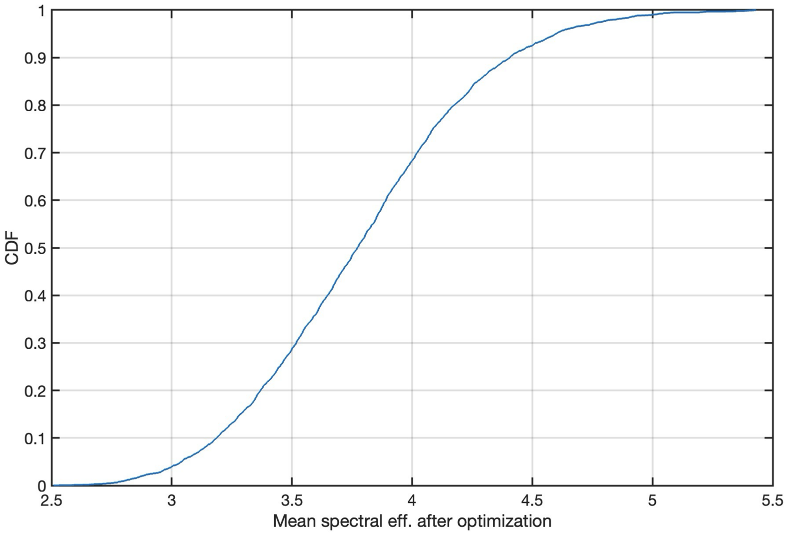


FIGURE 6.3: CDF of the mean system spectral efficiency after optimization

as the mean of the spectral efficiency of each BS. Figure 6.4 shows the performance increase in terms of percentual spectral efficiency after the optimization. Figure 6.5 shows the cumulative density function of the spectral efficiency of each BS in the scenario after the optimization. It provides additional insights on the local BS performances showing that, although the system spectral efficiency increase is promising, there are cases in which the spectral efficiency of a BS could be not acceptable.

6.3 Channel prediction for 6G-NTN Next Generation Multiple Access

The goal of this section is to address the evolution of NGMA considering Non Orthogonal Multiple Access (NOMA) and to provide a solution to the critical challenges the CSI ageing at the receiver. Indeed, NGMA aims at facilitating the efficient and flexible connection of many users/devices to the network, utilizing the available wireless radio resources. The section explores the current NGMA advancements and their potential integration into NTN systems. The contributions of this section to the existing literature are the following:

- A comprehensive discussion of different types of multiple access schemes for 6G NTN along with their pros and cons, specifically in NTN.
- A thorough analysis of the issues of CSI ageing.
- The design and evaluation, via numerical simulation, of a fully connected neural network capable of predicting CSI.
- Assessment of the predicted CSI in a code-based Non-Orthogonal Multiple Access (NOMA) technique, namely SCMA. The predicted CSI have been applied

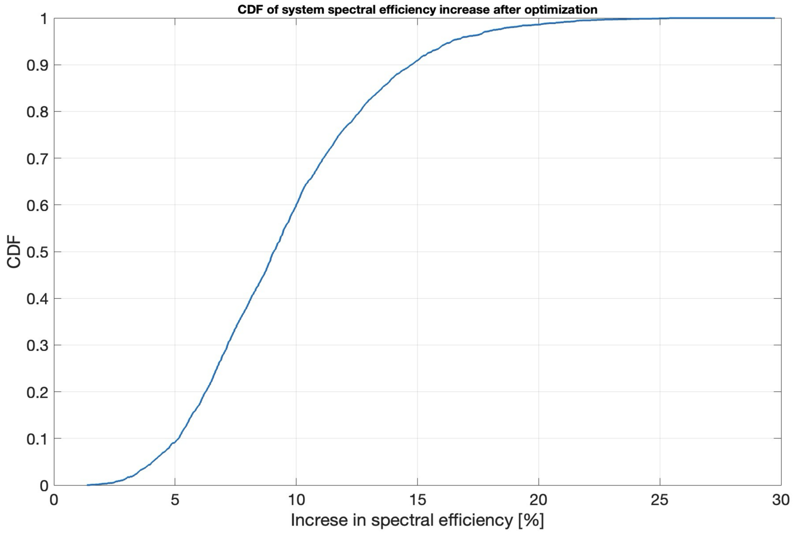


FIGURE 6.4: CDF of system spectral efficiency percentual increase after optimization

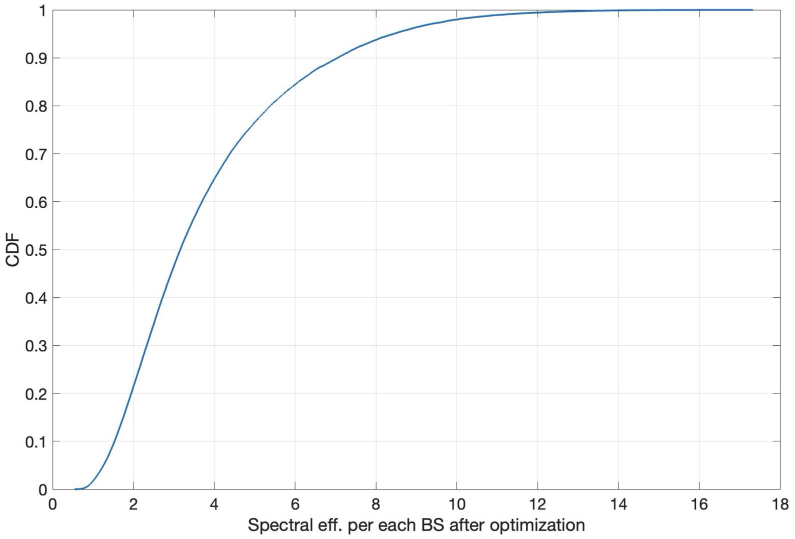


FIGURE 6.5: CDF of the spectral efficiency of the base stations after optimization

within the Message Passing Algorithm (MPA) at the receiver to jointly decode the superimposed symbols.

The CSI prediction algorithm represents a perfect example of ML-based algorithm that is enabled by real time control loops implemented at the network edge, and, in this case, directly in the base station.

6.3.1 Multiple Access Techniques

Multiple access schemes play pivotal roles in wireless communications. In this section, we propose a review and comparison of the major multiple access techniques along with their pros and cons for NTN.

6.3.1.1 Orthogonal Multiple Access

Owing to their benefits of minimal complexity and interference mitigation, orthogonal transmission strategies have seen widespread adoption in wireless communication systems. InOrthogonal Multiple Access (OMA), resources are strategically partitioned into separate, non-overlapping frequency bands, time slots, or codes, with each carefully assigned to an individual user.

Traditionally, satellite communication systems have relied on Frequency Division Multiple Access (FDMA), Time Division Multiple Access (TDMA), andCode Division Multiple Access (CDMA) as multiple access techniques. Notably, FDMA allows continuous access to the satellite in a given frequency band and its main advantage is simplicity. However, it is characterized by the lack of flexibility in case of reconfiguration and the loss of capacity when the number of accesses increases.

TDMA foresees the access to the channel during a time slot. This implies that the TDMA efficiency remains high for a large number of accesses. On the other hand, TDMA implies the need for synchronization through complex procedures. Finally, CDMA operates on the principle of spread spectrum transmission. The code sequence that spreads the spectrum constitutes the 'signature' of the transmitter. The receiver recovers the useful information by reducing the carrier's spectrum transmitted in its original bandwidth. This operation simultaneously spreads the spectrum of other users in such a way that these appear as noise of low spectral density. Besides the simplicity, the advantage of the CDMA is that it does not require any transmission synchronization between the transmitters. Moreover, it offers useful protection properties against interference from other systems and interference due to multiple paths; this makes it attractive for satellite communication with mobiles. In multibeam satellites, CDMA offers the potential of 100% frequency re-use between beams. The main disadvantage is poor efficiency, as a large satellite bandwidth is used for a low total network capacity with respect to the throughput of a single unspread carrier, in the case of a single beam network. The possibility of reusing frequency between adjacent beams greatly improves the overall efficiency. Another

limitation consists in the limited number of codes (and therefore the number of simultaneous users) offering the required performance in terms of inter-correlation properties.
In terrestrial communications, FDMA was prevalent in the first generation (1G), TDMA was prominent in the second generation (2G), and the third generation (3G) made use of CDMA. Similarly, Orthogonal Frequency Division Multiple Access (OFDMA) was prominent in the 4G era, and it continues in 5G, where users are assigned orthogonal resource blocks in terms of frequency, time, or code.
For instance, in Rel. 17, NTN-based 5G networks leverage OFDMA. The reason lies in the fact that, during the process of integration of Terrestrial and Non-Terrestrial communications, only minimal but strictly necessary changes to the 5G NR protocol stack have been made to employ it in a satellite system. This choice brings two main advantages: i) a simple transceiver design, implementation, and management; and ii) interference-free transmissions and, thus, enhanced overall quality and reliability of communications. On the other hand,OMA is characterized by the inefficient use of the spectrum, especially under the paradigm of UMTC massive access. Indeed, the broad coverage of the satellite beam leads to a large number of users contending for the limited time/frequency resources, resulting in severe congestion. As UMTC services may be offered through incomplete constellations, transmissions may be greatly delayed, exacerbating the effects of congestion. Moreover, procedures such as Random Access and user scheduling, are characterized by a long signaling phase, which can severely impair protocol and system performance [157].

6.3.1.2 Non Orthogonal Multiple access

In this section, NOMA refers to a scenario wherein multiple users concurrently utilize identical time/frequency resources. Different NOMA schemes, utilizing methods such as spreading, scrambling, interleaving, or exploiting multiple domains for user separation, have been considered for 5G. Although the specific approaches may vary, the fundamental concept remains consistent: accommodating multiple users within the same time-frequency resource blocks by utilizing distinguishing parameters.

1. **Space Division Multiple Access (SDMA)**: To overcome the limitations ofOMA, a new resource dimension, i.e., space, has been introduced. SDMA operates by directing communication signals toward geographically distinct paths. With the swift advancement of multi-antenna technologies, MIMO communication has emerged as a cornerstone of the current 5G standard and the prospective future wireless networks. By deploying multiple antennas at transmitters and/or receivers, MIMO systems leverage additional spatial degrees of freedom (DoFs) compared to their single-antenna counterparts. This principle can be applied to serve multiple spatially separated users using the same

time/frequency resources in a Multi-User MIMO (MU-MIMO) system. Consequently, SDMA techniques can serve multiple users/devices simultaneously in the same time/frequency/code domain while distinguishing them in the spatial domain. Linear Precoding (LP) is the predominant method for SDMA, given its low complexity. Notably, inter-user interference can be effectively suppressed by leveraging spatial DoFs to design appropriate transmit and/or receive beamformers. The exploitation of antenna arrays through beamforming, precoding, and MIMO techniques has been extensively addressed also in NTN due to their ability to provide high throughput in Full Frequency Reuse (FFR) scenarios [158]. These techniques can effectively exploit spectrum, especially when paired with smart antenna terminals, e.g.,VSAT terminals, with the ability to steer their main radiating lobe toward the satellite. However, less sophisticated terminals such as handheld and IoT devices do not benefit from SDMA in NTN due to their low antenna gain. Moreover, SDMA techniques require the perfect knowledge of the channel state information. This is a challenging task due to the high-speed movement of the satellite, which leads to the so-called issue of information ageing. Besides, to properly work, SDMA does not only require channel estimation but also scheduling, and interference management, which leads to substantial signaling overhead, particularly in dynamic networks such as NTN. It is also worth mentioning that SDMA is sensitive to user deployment. Since users in the same beam experience similar channel conditions, sophisticated scheduling algorithms based on users' channel coefficients correlation [159] or inter-users distances [160] need to be considered. Finally, SDMA demonstrates efficiency in networks with low loads but faces performance degradation in overloaded systems due to constraints imposed by limited spatial resources. This is especially true in NTN, where the scan angle resolution is low.

2. **Rate Splitting Multiple Access (RSMA)**: This multiple access technique has recently emerged as a promising transmission strategy for multi-antenna wireless networks. The key concept of Rate Splitting Multiple Access (RSMA) is splitting each user message into sub-messages at the transmitter. When applied in Downlink (DL), the transmitter partitions a portion of each user's message (referred to as *private message*) into a common message intended for all users served. The remaining private messages, along with the generated common message, are transmitted through beamformers. The amount of transmission power assigned to the common messages and the private messages is regulated by the scaling factor, the larger the scaling factor the more power is assigned to the private parts. Upon reception, the receivers treat the private messages of all users as interference while decoding the common message. Subsequently, they subtract the common message from the received signal, i.e., with Successive Interference Cancelation (SIC). Following this, the intended

private message is decoded and integrated with a portion of the decoded common message. In the uplink, the sub-messages from a given transmitter can be decoded at the receiver in a non-consecutive manner. This message-splitting capability provides RSMA with a flexible interference management strategy, allowing for partial decoding of interference and treating the residual signal as noise. RSMA offers a variety of transmission models depending on the specific approaches used for message splitting and combining, including linearly precoded 1-layer RS, 2-layer hierarchical RS (HRS), and generalized RS (GRS). The main advantage of RSMA is the flexibility [161], as the scheme is highly adaptable to varying network loads and user deployments characterized by varying channel directions and strengths. Furthermore, although the instantaneous CSI at the transmitter is quite difficult to obtain in practice, RSMA has been recognized to be more robust to imperfect CSIT compared to SDMA thanks to its DoF optimality in the multi-antenna broadcast channels. In the NTN scenario, the estimation of CSIT is even more challenging since the CSI is usually out-of-date due to the fast time-varying channel and long propagation delay. This outdated CSIT information causes a degradation in the network performances due to sub-optimal computation of the scaling factor and degraded interference cancellation performances. It is important to note that, in NTN, the performance of RSMA tends to decrease as the distance between users decreases, [162]. Given that NTNs are typically overloaded, the limited distance between users presents a challenge to RSMA, which can only be mitigated through appropriate user scheduling strategies. Furthermore, the substantial number of users that need to be served in an overloaded NTN system introduces a notable increase in complexity for calculating the optimal power scaling factor, which is essential for achieving optimal RSMA performances.

3. **Power domain NOMA**: The primary concept behind Power domain (PD)-NOMA is to serve multiple users utilizing identical time/frequency/code resources while distinguishing them based on power allocation. At the receiver side, the SIC algorithm is used to decode the superimposed packets. The main advantage of the PD-NOMA technique is user fairness [163], as during the power allocation, it prioritizes users with weaker channel strength to facilitate effective interference cancellation. On the other hand, PD-NOMA is sensitive to CSI inaccuracy. Besides, considering that the satellite channel can be assumed Rayleigh distributed, the users within the beams experience similar Signal-to-Noise Ratio (SNR); therefore, it is not easy to distinguish them in the power domain.

4. **Spreading-based NOMA**: Drawing inspiration from CDMA, Spreading-based NOMA aims at serving multiple users within the same time/frequency resources through spreading sequences. These schemes are further categorized into low-density spreading (LDS)-based and non-LDS-based approaches. In

LDS-based schemes, the spreading sequences consist of sparse or non-orthogonal, low-cross-correlation sequences achieved by deactivating a significant number of spreading signature chips in conventional CDMA. In the receiver of LDS-based schemes, iterative algorithms, such as the MPA, can be employed for joint detection of multiple data streams, approaching near maximum-likelihood performance. For non-LDS schemes, SIC or Parallel Interference Cancellation (PIC) techniques are applied at the receiver. Notably, LDS and non-LDS schemes include LDS-CDMA [164], LDS orthogonal frequency-division multiplexing (LDS-OFDM, [165]), SCMA, [166]), Pattern Division Multiple Access (PDMA), [167]), LDS Signature Vector Extension LDS-SVE, [168]), Multi-User Shared Access (MUSA), [169]),Non-Orthogonal Coded Multiple Access (NCMA), [170]),Non-Orthogonal Coded Access (NOCA), [171]),Low Code Rate Spreading (LCRS),Frequency Domain Spreading (FDS), and Group Orthogonal Coded Access (GOCA). Spreading-based NOMA techniques have the advantage of enhancing the spectrum efficiency and reducing signaling overhead, since users can choose their spreading sequences autonomously to reduce the signaling overhead and latency. However, spreading-based NOMA requires complex multi-user detection techniques and the design of the optimal spreading sequence, which is still a difficult problem, particularly when the number of users is large. In addition, most of the existing spreading-based NOMA schemes assume perfect CSI knowledge. As previously mentioned, this is a challenging task in a highly dynamic network such as NTN. Moreover, most of the existing spreading-based NOMA schemes assume perfect synchronization at the receiver. This can be achieved through a GNSS receiver. However, the user needs to autonomously and continuously synchronize, which implies the interruption of data transmission. A delayed signal from a user could disturb the message exchange in the iterative multi-user detection, leading to performance loss.

5. **Interleaving-based NOMA**: As implied by the name, Interleaving-based NOMA schemes use interleavers to differentiate between superimposed users. Additionally, low-code rate Forward Error Correction (FEC) can be incorporated alongside interleaving. For Multi-User Detection (MUD), an Elementary Signal Estimator (ESE) can be employed at the receiver. Notable interleaving-based schemes include Interleaved Division Multiple Access (IDMA) [172] and Interleave-Grid Multiple Access (IGMA) [173]. Similarly to other NOMA schemes, these schemes require an accurate knowledge of CSI at the receiver. In addition, as information bits are interleaved in time, the channel is also assumed to be constant within a time window. Due to the inherent dynamic nature of the NTN channel, such an assumption is typically not valid for NGSO systems. Moreover, the NTN channel is often characterized by slow fading due to atmospheric events, e.g., rain and clouds. Therefore, the effectiveness of chip-by-chip detectors would be reduced, as such algorithms rely on the reduced

adjacent bits correlation that is typically provided by interleavers.

6.3.2 Technical issues associated to NGMA in NTN

In contrast to OMA, precise CSI acquisition holds greater significance for NOMA. Indeed, the accuracy of CSI profoundly influences the performance of the receiving algorithms, as CSI is essential for reconstructing decoded signals for subtraction or to minimize the detection probability to jointly decode the signal. However, unlike TN channels, NTN channels exhibit two distinct characteristics:

- The significant propagation latency between satellites and on-ground devices, due to the long transmission distances, which can result in outdated CSI when it is acquired.
- The high mobility LEO satellites which leads to rapidly changing channels and severe Doppler frequency shifts.

To consistently track the variation of the channel, an effective approach is to design a linear predictor that leverages the relationship between outdated and instantaneous CSI or the estimation of the channel coefficients through outdated pilots. This relationship can be determined by analyzing the spatial and temporal correlation within the NTN channel models. The coefficients of the linear predictor can then be computed using various criteria, among which the Least Squares (LS) and Minimum Mean Square Error (MMSE). However, this approach has a twofold drawback. On the one hand, the use of pilots reduces the achievable throughput; on the other hand, the CSI acquisition procedure requires an exchange of control information whose duration is extended by the long round trip time. A second approach relies on the use of AI, [174]–[176]. Indeed, AI-based algorithms have the potential to learn from real data and perform more accurate predictions, despite typically requiring larger computational capabilities. A DL-based CSI prediction scheme is proposed in [174] to address the channel aging problem for downlink massive MIMO. Specifically, a satellite channel predictor composed of LSTM units is proposed. A more sophisticated solution for the CSI prediction in downlink massive MIMO is addressed in [175], where two modules, namely the Convolutional Neural Network (CNN) and LSTM, are designed to transform uplink CSI into downlink CSI, aiming at removing the need of a feedback message. The CNN was shown to be able to provide more accurate predictions for satellites at lower orbits and for smaller distances between the uplink and downlink bands. The Gated Recurrent Unit (GRU) architecture implemented in [176] was able to learn the temporal correlations of a LEO satellite channel based on historical time series, showing the impact of the elevation angle and the LEO altitude, among the others, on the symbol error rate.

6.3.3 Proposed solutions

6.3.3.1 System Architecture

the considered system architecture is the Orbit-split model, where the CSI prediction algorithm is implemented in the service nodes along with NGMA.

Generally, signal propagation between the user and the satellite is unshaded due to the high operation altitude of UEs and satellites. In this case, considering that the terminals are in the LoS of the satellite, the received signal can be expressed as:

$$y(t) = \sum_{j=1}^{J} h_j \cdot x(t - \tau_j(t)) e^{j2\pi f_{d_j}(t)} \tag{6.1}$$

where $\tau_j(t)$ is the delay due to the distance between the j-th user and the satellite, $x(t)$ is the transmitted signal, $f_{d_j}(t)$ is the Doppler shift, which can be written as:

$$f_{d_j}(t) = \frac{f_c v_s R_E \cos \varepsilon_j(t)}{c(R_E + h_{sat})} \tag{6.2}$$

with R_E indicating the Earth radius, f_c the frequency carrier, and h_{sat} the satellite altitude. Finally, h_j refers to the amplitude of the channel coefficients, which can be written as:

$$h_j = \frac{g_{i,n,s}^{(TX,t)} g_{i,n,s}^{(RX,t)}}{4\pi \frac{d_{i,s}^{(t)}}{\lambda} \sqrt{L_{i,s}^{(t)}}} e^{-\jmath \frac{2\pi}{\lambda} d_{i,s}^{(t)}} \tag{6.3}$$

where: i) $d_{i,s}^{(t)}$ is the slant range between the i-th user and the s-th node, which is assumed to be the same for all the co-located radiating elements on-board the node; ii) λ is the signal wavelength; iii) $\kappa B T_i$ denotes the thermal noise power, with B being the user bandwidth (for simplicity assumed to be the same for all users) and T_i the equivalent noise temperature of the i-th receiver; iv) $L_{i,s}^{(t)}$ represents the additional losses between the s-th node and the i-th user, assumed to be the same for all the co-located radiating elements on-board the node; v) $g_{i,n,s}^{(TX,t)}$ and $g_{i,n,s}^{(RX,t)}$ represent the transmitting and receiving complex antenna patterns between the i-th user and the n-th radiating element on-board the s-th node, respectively; and vi) $\varphi_{i,s}^{(t)}$ is the phase misalignment that might be present between different nodes due to non-ideal swarm synchronisation, modelled as a uniform random variable (r.v.) $\mathcal{U}(0, 2\pi)$. The additional losses are computed based on TR 38.811, [30]:

$$L_{i,s}^{(t)} = L_{i,s}^{(SHA,t)} + L_{i,s}^{(ATM,t)} + L_{i,s}^{(SCI,t)} + L_{i,s}^{(CL,t)} \tag{6.4}$$

in which: i) $L_{i,s}^{(SHA,t)}$ denotes the log-normal shadowing loss with standard deviation σ_{SHA}; ii) $L_{i,s}^{(ATM,t)}$ includes the atmospheric loss due to gaseous absorption; iii) $L_{i,s}^{(SCI,t)}$ is the scintillation loss; and iv) $L_{i,s}^{(CL,t)}$ is the Clutter Loss (CL), to be included for UEs in Non-Line-of-Sight (NLOS) conditions. Referring to the 3GPP channel model, the UE is defined to be in LoS or NLOS conditions with a probability that is a function of the elevation angle and the propagation environment (sub-urban, urban, dense-urban). In this context, we assume that a UE that is LoS (NLOS) conditions during the estimation phase is still in LoS (NLOS) conditions in the transmission phase. This assumption is motivated by observing that the probability that the propagation conditions of the UE will change from LoS (NLOS) to NLOS (LoS) after a few ms is negligible, considering that the probabilities of LoS or NLOS conditions are provided with a 10° granularity in TR 38.811, [30]. This assumption implies that the UE has the same CL and σ_{SHA} in both the estimation and transmission phases, but the realisations of the log-normal r.v. modelling the shadowing are different.

6.3.3.2 CSI Prediction

Taking into account the impact of transmission delay between the LEO satellite and the UEs, the CSI acquired by the LEO satellite may not accurately represent the true channel characteristics of the UEs. Consequently, utilizing outdated CSI in the demodulator could result in a degradation of system performance. In this section, we introduce an Neural Network (NN)-powered CSI predictor integrated within the LEO satellite receiver. The CSI predictor task is to predict the future CSI of a specific UE in the coverage area, relying on information about the current satellite's and UE's status, addressing the issue of outdated information. The module takes as input the satellite position and trajectory, and the UE position and velocity vector at time $T0$, and it outputs the UE's CSI at time $T0 + \Delta t$ ms.

The general prediction framework is shown in Fig. 6.6. The CSI prediction module performs offline training and online prediction. The purpose of the offline training task is i) to perform the pre-deployment training of the network and ii) to ensure that the neural network is updated with changes in channel characteristics by training it with the latest measured data, allowing it to adapt to recent developments. The purpose of the online prediction task is to respond to CSI requests from the receiver by providing real-time predictions. Offline neural network training would be enabled by historical measures of the satellite channel. Given the unavailability of a real channel dataset, to evaluate the proposed prediction technique, we developed a synthetic dataset based on a numerical satellite channel simulator. The structure of the offline pre-deployment training is represented in Figure 6.7. Although the synthetic data set enables rapid prototyping and provides strict control over channel parameters, it unavoidably reflects the modeling assumptions embedded in the simulator, thus introducing potential bias. In particular, phenomena such as atmospheric scintillation, residual hardware impairments, and co-channel interference, each difficult to model with high fidelity, may be underrepresented. Hence, the performance achieved in simulation should be regarded as an optimistic upper bound, and some mismatch is anticipated when the predictor is first confronted with on-orbit measurements.

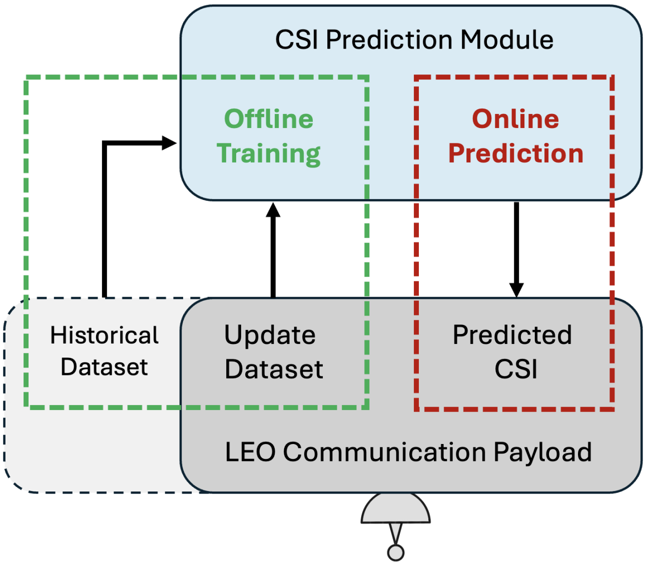


FIGURE 6.6: Framework of CSI prediction scheme.

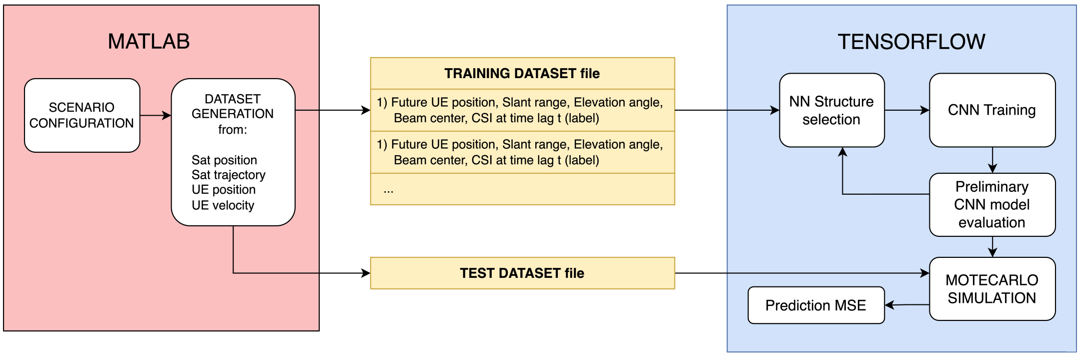


FIGURE 6.7: High-level flowchart of the AI-based CSI prediction simulation.

6.3.3.3 CSI Dataset Generation

The dataset file is generated by the satellite channel simulator, knowing the position and trajectory of the satellite and the UE. The satellite channel simulator has been configured with a LEO satellite that orbits at 600 km, compliant with 3GPP 38.811 Set-2 satellite parameters, [20]. A single moving beam is considered. The simulated channel is in UL and considers a sub-urban environment in NLOS. Users are uniformly distributed in the coverage area with a density of 1 UE per square kilometer and move at a pedestrian, vehicular, or railway velocity of 3, 100, and 250 km/h, respectively. The simulator computes the CSI of the UEs in the coverage area every Δt ms, taking as input the instantaneous satellite ephemeris and the specific UE position and velocity vector. At every simulation iteration, the position of the satellite and the users is updated according to their velocity vector and the new CSIs are computed. In this framework, the training dataset file is built by accumulating the information generated in each iteration. Specifically, each row of the dataset file is a list of data reporting the satellite ephemeris, the UE position, the UE velocity, and

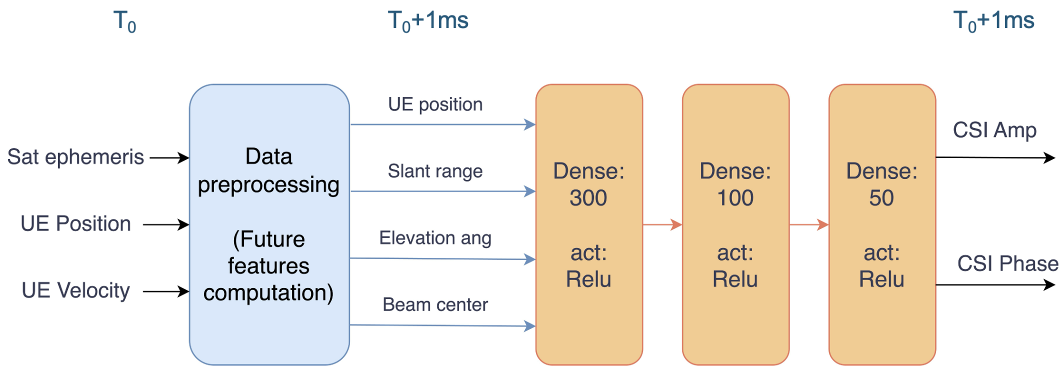


FIGURE 6.8: NN structure.

the information about the instantaneous CSI of the UE, used as a label for the NN training.

6.3.3.4 NN Architecture and Training

The NN architecture is shown in Figure 6.8. Before entering the neural network, the input data is elaborated by a data preprocessing block. Specifically, the satellite ephemeris, UE position, and UE velocity at time T0 are utilized to calculate the neural network's input features, representing the link's future parameters at time $T0 + \Delta t$:

- Future UE position,
- Future slant range between the UE and the satellite,
- Future elevation angle related to the specific UE,
- Satellite future beam center.

The NN consists of a three-layer feed-forward neural network using the Rectified Linear Unit (ReLU) activation function. The architecture includes 300 neurons in the input layer, 100 in the hidden layer, and 50 in the output layer. The choice of ReLU introduces non-linearity for capturing complex data patterns in the first two layers, while the output layer is kept linear to perform CSI regression. The network processes the link's future parameters through the first input layer, extracts abstract features in the hidden layer, and estimates the CSI amplitude and phase in the output layer. Given that this model must run in the satellite receiver for each served user, it is crucial to minimize the computational complexity. The adoption of a small feed-forward configuration balances model performance and computational efficiency, ensuring effective functionality while keeping resource demands at a minimum. Existing literature on mobility-driven channel prediction shows that shallow fully connected networks can approximate the non-linear mapping between kinematic features and future CSI with satisfactory accuracy once the salient geometry-based inputs are provided, [177], [178]. Deeper or recurrent alternatives, although more expressive in principle, would multiply both parameter count and inference latency, conflicting with the real-time and power-limited constraints of an on-board receiver. Guided by these insights and by the need to preserve a lean computational footprint, a compact three-layer feed-forward topology was therefore selected as the most pragmatic compromise between model expressiveness and resource efficiency.

The NN is trained to minimize the Mean Square Error (MSE) between the output CSI prediction and the actual CSI value by exploiting the Adaptive Moment Estimation (ADAM) algorithm. The MSE of the CSI amplitude MSE_{Amp} and of the CSI phase MSE_{Pha} can be expressed as:

$$MSE_{Amp} = \frac{1}{N}\sum_{n=1}^{N}\left(\hat{h}_n^{Amp} - h_n^{Amp}\right)^2 \tag{6.5}$$

$$MSE_{Pha} = \frac{1}{N}\sum_{n=1}^{N}\left(\hat{h}_n^{Pha} - h_n^{Pha}\right)^2 \tag{6.6}$$

Where $\hat{\mathbf{h}}_n^{Amp}$ and $\mathbf{h}_n^{Amp}$ represent the predicted amplitude of the CSI and its actual value, $\hat{\mathbf{h}}_n^{Pha}$ and $\mathbf{h}_n^{Pha}$ represent the predicted phase of the CSI and its actual value, and N is the total number of samples in the training set. The loss function used to train the NN is the aggregated MSE of the CSI amplitude and phase, denoted as MSE_{CSI}.

$$MSE_{CSI} = \frac{1}{N}\sum_{n=1}^{N}\left[\left(\hat{h}_n^{Amp} - h_n^{Amp}\right)^2 + \left(\hat{h}_n^{Pha} - h_n^{Pha}\right)^2\right] \tag{6.7}$$

$$\text{Loss}\left(\Theta\right) = MSE_{CSI} \tag{6.8}$$

6.3.3.5 NN Complexity Analysis

The complexity of the CSI prediction NN in terms of number of additions and multiplications can be assessed starting from the complexity of a single neuron. Given neuron n in the fully-connected layer i, its output $y_{n,i}$ can be expressed as:

$$y_{n,i} = ReLU\left(b_{n,i} + \sum_{K=1}^{N_{i-1}} y_{k,i-1} \cdot W_{k,i}\right), \tag{6.9}$$

where N_{i-1} is the number of neurons in layer $i-1$ and $b_{n,i}$ and $W_{k,i}$ are the trained parameters for layer i. Consequently, it is possible to compute the number of additions and multiplications performed in the fully connected layer i as $(N_{i-1} + 1) \cdot N_i$. Hence, being L the number of layers in the NN, the number of additions and multiplications C_{NN} of the complete fully-connected NN can be expressed as:

$$C_{NN} = \sum_{i=1}^{L-1}(N_{i-1} + 1) \cdot N_i, \tag{6.10}$$

Specifically, the total number of additions of the proposed NN is 36300, equal to the number of multiplications. It is worth mentioning that, since the NN computes the predicted CSI of a single user at a time, the total number of operations C_Q needed to predict the CSI of Q users in the coverage area becomes $C_{NN} \cdot Q$.

6.3.4 Numerical results

6.3.4.1 Performance of the NN based channel predictor

In the following, the performances of the CSI predictor are detailed. The metric to be evaluated is the mean squared error between the reference CSIs contained in the test dataset and the CSIs predicted by the NN module. Since the NN has two outputs: i) the amplitude; and ii) the phase of the complex CSI, the NN is evaluated with three metrics:

- the MSE of the amplitude of the complex CSIs;
- the MSE of the phase of the complex CSIs;
- the aggregated MSE of amplitude and phase.

The metrics obtained after the NN model training are reported in Table 6.3. The results show a better performance in the prediction of the CSI amplitude compared to the prediction of the CSI phase. This behaviour reflects the statistics of the dataset. Indeed, the amplitude of a UE's CSI at time t0 is highly correlated to the amplitude of a UE's CSI at the subsequent time instant t1, thus, the NN can easily learn the CSI amplitude behavior after the UE's movement in the time interval t0-t1. Regarding the phase, we can observe that the CSI accuracy is lower than that of the amplitude. This is because the phase of the CSI phase is less correlated in time.
To investigate how the CSI prediction performance varies along the orbit and inside the beam, in Figure 6.9(a) is shown how the MSE of the CSI amplitude varies at different elevation angles, while in Figure 6.9(b) the same analysis is shown for the MSE of the CSI phase. Figures show quite constant CSI prediction performance at different elevation angles, meaning that the NN can be proficiently exploited in extended coverage areas without losing prediction precision. The MSE of the CSI amplitude slightly increases at higher elevation angles, this is motivated by the distribution of the dataset examples. Since the training is performed considering an earth-fixed beam, the coverage area remains constant and the beam is steered at the satellite. This configuration produces a higher number of dataset examples at a lower elevation angle compared to the number of examples at an elevation angle closer to 90 degrees. Having more data at lower elevation angles, the NN can learn the CSI behavior slightly better than at higher angles. To obtain completely constant performance along the orbit we could artificially increase the portion of dataset closer to 90 degrees, but this will not be possible when training the network with a real dataset collected from a real NTN system.

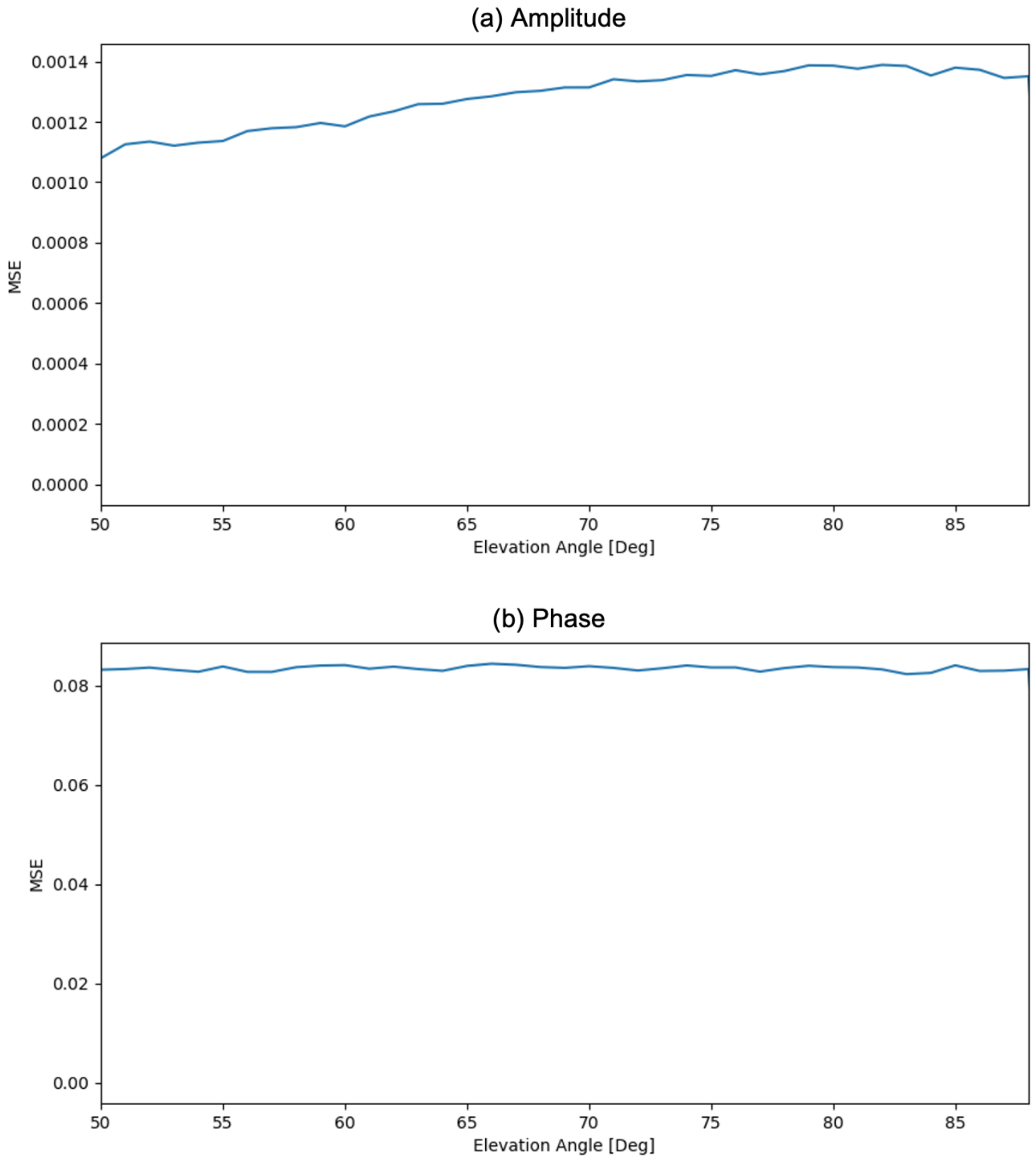


FIGURE 6.9: MSE of CSI amplitude (a) and phase (b) at different elevation angles.

TABLE 6.3: MSE metrics of CSI prediction NN

Metric	Value
Amplitude MSE	0,0012
Phase MSE	0,0843
Aggregate MSE	0,0836

6.3.4.2 Performance of the SCMA with the NN based predictor

To verify the goodness of the proposed prediction, we apply the CSI prediction to NOMA.

In particular, we consider a SCMA(6,4) system where multiple accesses are achieved

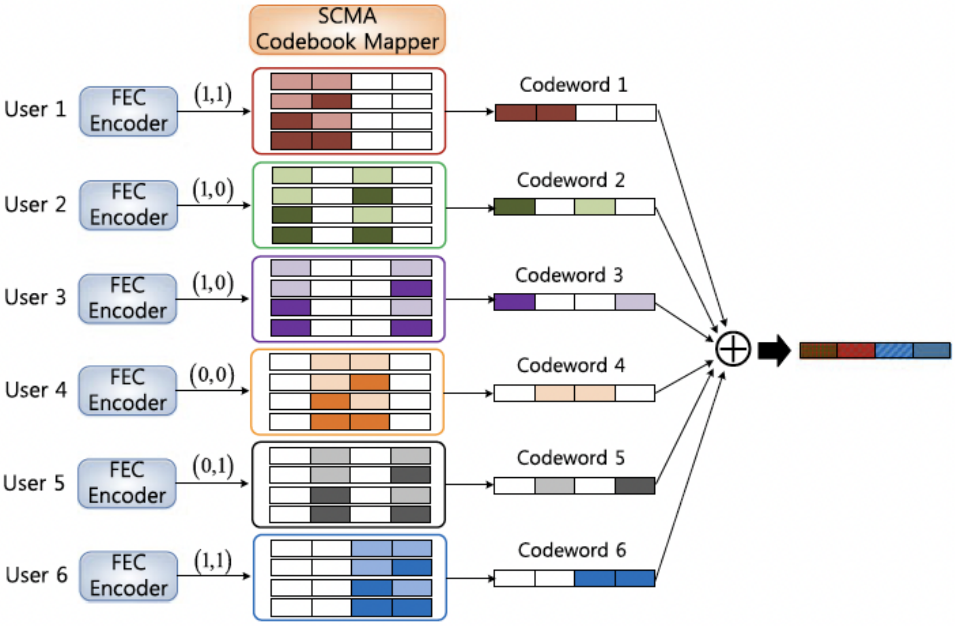


FIGURE 6.10: SCMA encoder with K = 4, N = 2, and J = 6.

through the sharing of the same time-frequency resources among users. The users are assumed to be in a connected mode, i.e., they performed the random access and the data transmissions are scheduled.

The considered uplink SCMA is shown in Figure 6.10. Data bits of a user are mapped to a K dimensional codeword from the codebook and transmitted on K subcarriers sequentially. The number of codebooks is six and the number of codewords is four as well as the number of sub-carriers.

In terms of channel estimation, the gNB only knows the channel coefficient of the first OFDM symbol of each user and will use it for all the symbols in the MPA receiver. However, the channel coefficients change with time, due to the satellite movement. The performance is evaluated through the average Block Error Rate (BLER) of the six users.

Figure 6.11 reports the BLER achieved using the predicted CSI compared to the BLER obtained with ideal estimation. As observed, the estimation derived from the NN closely approximates the BLER under ideal conditions, where the gNB has perfect knowledge of the CSI for each user in every OFDM symbol. This similarity is attributed to the statistics of the phase error, which remains within the robustness threshold of the SCMA technique, enabling accurate decoding of the six codewords. It is worth mentioning that without the CSI prediction and the ideal channel estimation, the BLER is 1 for the entire range of simulated values of EbN0. Notably, employing CSI prediction can reduce the number of pilots required for channel estimation, potentially increasing the overall throughput.

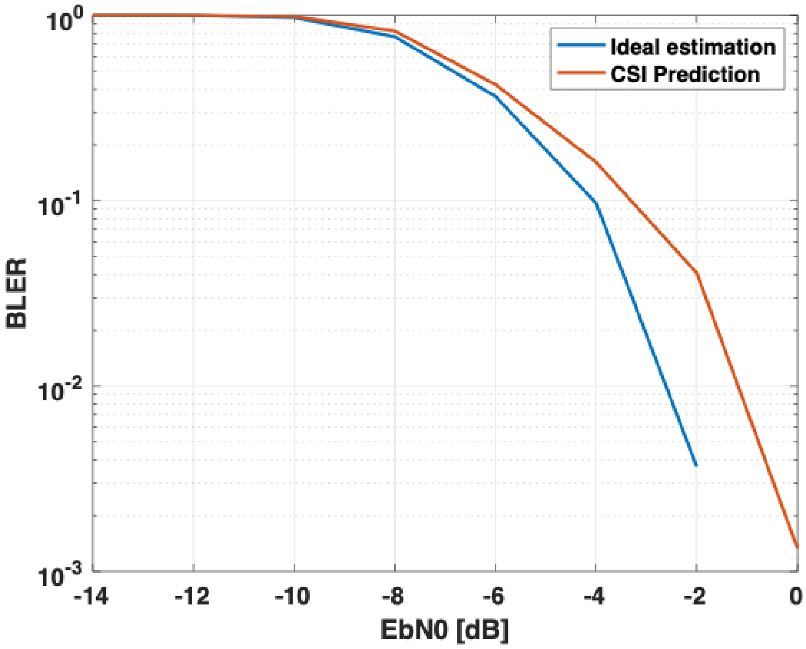


FIGURE 6.11: BLER vs EbN0. Code rate = 193/1024

6.3.5 Discussion and open issues

This article investigates the advantages and disadvantages of different multiple access techniques in NTN. The findings reveal that each access technique offers distinct benefits, however all of them necessitate accurate CSI. Consequently, an NN-empowered CSI prediction has been proposed and tested with a SCMA. The numerical results demonstrate that the proposed solution can achieve high prediction accuracy compared with adopting the outdated CSI directly. For future research, we have identified the following open issues. One critical aspect is the CSI acquisition, and, therefore the design of pilots in the chosen access schemes. Once the CSI is acquired, NN algorithms can be employed to predict future CSI, reducing the need for excessive pilot signals and consequently increasing the throughput. To optimize power consumption on satellites, the NN architecture should be compact and efficient. A smaller neural network reduces computational load and energy requirements, making it more suitable for the limited power resources available on-board satellites. A second important open issue will be the absence of a GNSS receiver at the UE, raising synchronization challenges. It is crucial to determine which network elements will interact to provide accurate UE positioning and how to manage synchronization without GNSS. Furthermore, decisions must be made regarding the placement of the core network to optimize performance and efficiency. Finally, the 3D architecture can efficiently serve massive UEs with different Quality of Service requirements. In particular, the concept of swarm introduces possibilities for distributing access or implementing distributed beamforming in the uplink. This requires mechanisms for master satellite identification and coordination among satellites to enhance connectivity and service reliability.

6.4 CNF Orchestration in NTN

This section reports the outcomes of the paper submitted for publication "A Novel Framework for Proactive CNF Orchestration in 6G NTN", prepared in collaboration with Alice Piemonti, Vito Cianchini, Carla Amatetti, Massimo Neri, and Alessandro Vanelli-Coralli.

Control loops, implemented through O-RAN RICs, form the cornerstone of intelligent network management in the transition to 6G. These loops enable real-time and non-real-time optimization of network functions by embedding AI-driven decision-making at multiple levels of the network architecture. As the industry embraces CNFs, within O-RAN RIC frameworks, the orchestration of these functions becomes crucial for delivering seamless and efficient RAN optimization. Indeed, the optimization algorithms implemented in Near-RT control loops often exhibit fluctuating computational demands, driven by dynamic changes in network conditions or the specific optimization tasks required at any given moment. These variations can lead to suboptimal utilization of computational resources when not effectively managed. CNFs, which extend the virtualization concepts of VNFs into containerized environments, align perfectly with the dynamic and distributed nature of O-RAN. Their integration empowers networks to adapt rapidly to fluctuating resource demands, particularly in the complex and variable conditions of NTNs.

In the framework of CNF-based NTN RAN optimization, this section introduces a novel architecture for proactive CNF orchestration, leveraging AI to dynamically forecast resource demand. Using ML models, our framework can predict time-series data for critical metrics such as CPU usage, memory consumption, and bandwidth requirements. These predictions enable the orchestrator to allocate resources in advance, reducing cold start delays[179] and ensuring a smoother scaling process. A proactive approach not only enhances performance by preemptively provisioning resources before bottlenecks occur but also optimizes energy consumption and resource allocation, particularly in environments with fluctuating workloads [180].

The contributions of our work are:

1. **A novel framework for CNF orchestration in NTN**, embedding AI as a native component to enable proactive network management. This framework also includes an architecture for Machine Learning operations (MLOps) designed to train, re-train, save, and deploy machine learning models within the network environment. Additionally, the integration of the CNF orchestrator in the NTN architecture ensures dynamic resource allocation and management, providing enhanced flexibility and scalability, essential for the complex NTN ecosystem.

2. **AI/ML-based solutions for time-series resource forecasting**, enabling anticipatory scaling decisions to efficiently manage both overprovisioning and underprovisioning scenarios.

6.4.1 Related Work

The concept of VNFs has become fundamental in 5G networks, making optimal orchestration essential to manage the increasing complexity of these systems. Recent works highlight the necessity of AI for efficient VNF placement. For example, [181] emphasises the importance of AI-driven solutions to optimise VNF placement and orchestration, laying the groundwork for future networks. Similarly, [182] proposes an Orchestration Framework that leverages machine learning across the entire network function vistrualization environment. The framework includes a monitoring module providing key metrics such as topology, host machine specifications, and resource utilisation, including CPU, memory, and bandwidth (RX/TX). However, despite being tested in a real-world environment, the paper does not provide concrete results regarding the performance of its AI modules. In [180], a taxonomy is provided for ML-based orchestration, proposing a reference architecture to control overall resource utilisation, energy efficiency, and application performance. The authors stress the critical role of AI in improving VNF/CNF placement, achieving higher accuracy, and reducing computation delays.

Many existing studies focus on VNF placement, often treating it as a mathematical optimization problem. Some of them aim to optimize different aspects of VNF placement through AI algorithms. For instance, [183] uses deep reinforcement learning to automatically deploy Service Function Chains (SFCs) consisting of VNFs. In [184], multiple machine learning techniques are used to cluster data into accurate VNF profiles, aiming to optimise the QoS of MANO systems while minimising runtime monitoring loads.

However, these solutions primarily offer reactive orchestration, where VNF placement adjusts only after resource thresholds are exceeded. We believe a proactive approach, especially in containerized environments, would be more efficient, as it can reduce cold start delays. [185] proposes an ML-centric platform called RC utilising two approaches: Reinforcement Learning (RL) for direct resource management actions like scaling and migration, and predictive ML models for workload forecasting. The platform's independent ML framework enables seamless adaptation to production workloads, with servers learning optimal thresholds for resource management. RC has proven that ML techniques outperform traditional methods, offering more accurate predictions and better resource allocation. Similarly, [186] employs machine learning models like ARIMA and LSTM to predict incoming traffic in a 5G O-RAN environment, aiming to reduce operational expenditure (OPEX) by optimising VNF placement and scaling of virtual resources. This approach allows proactive orchestration, helping to avoid delays that could violate Service Level Agreements (SLAs). However, the authors acknowledge the challenge of model generalisation, where conventional supervised learning models struggle with dynamic environments and limited training data, often resulting in poor performance when handling new scenarios.

Only a few studies specifically address CNF placement. [187] presents a mathematical model for CNF placement in O-RAN, improving end-to-end delay in data plane operations under resource-constrained environments. Another notable example, [179], uses a model-free Q-Learning RL agent to optimise container instances in serverless environments. The agent monitors the environment and dynamically scales up or down in advance the number of active containers to reduce cold start latency. This approach outperforms the default Kubeless auto-scaler, which only reacts after resource spikes. By proactively adjusting resources, the RL agent mitigates cold start delays and efficiently manages workloads, offering better performance and resource utilisation than traditional serverless scaling methods. Additionally, [188] explores dynamic resource allocation for Kubernetes clusters through ML, including more advanced Deep Neural Networks.

Even though many studies focus on dynamic approaches for VNF allocation, often leveraging advanced algorithms, they still react to events only after they occur. To address this limitation, our framework centers on proactive orchestration, using AI/ML to anticipate network function resource needs in a containerized context. This proactive approach effectively manages both overprovisioning and underprovisioning scenarios, ensuring CNFs are allocated with the exact resources they require. This capability is essential in NTNs, where onboard resources are limited, making it critical to reduce energy consumption while maintaining network QoS.

6.4.2 A Novel Framework for CNF Orchestration

In this section, we present the key components of our architecture for Cloud-native Network Function orchestration (Fig.6.12), where AI is envisioned as a native component (described in detail in Fig. 6.13). The CNF orchestrator is designed to seamlessly integrate and manage both terrestrial and non-terrestrial networks and is incorporated into the network segment.

6.4.2.1 CNF Orchestrator

The CNF Orchestrator (Fig. 6.12) is the core component responsible for managing the deployment and orchestration of VNFs in containers. In the context of NTNs, the CNF Orchestrator is designed to address the unique challenges posed by satellite-based communication systems, such as high latency, dynamic link conditions, and limited computational resources onboard satellites. In our vision of 6G NTNs, VNFs are deployed in containers and managed by Kubernetes. The dynamic orchestration enabled by Kubernetes is crucial for maintaining performance not only in terrestrial networks but also in highly dynamic satellite environments, where fluctuating network conditions, variable latency, and limited onboard resources are common. Kubernetes' ability to rapidly adapt to changing traffic patterns and network conditions is vital for ensuring optimal service delivery in NTNs, particularly in LEO constellations where constant handovers and network shifts occur.

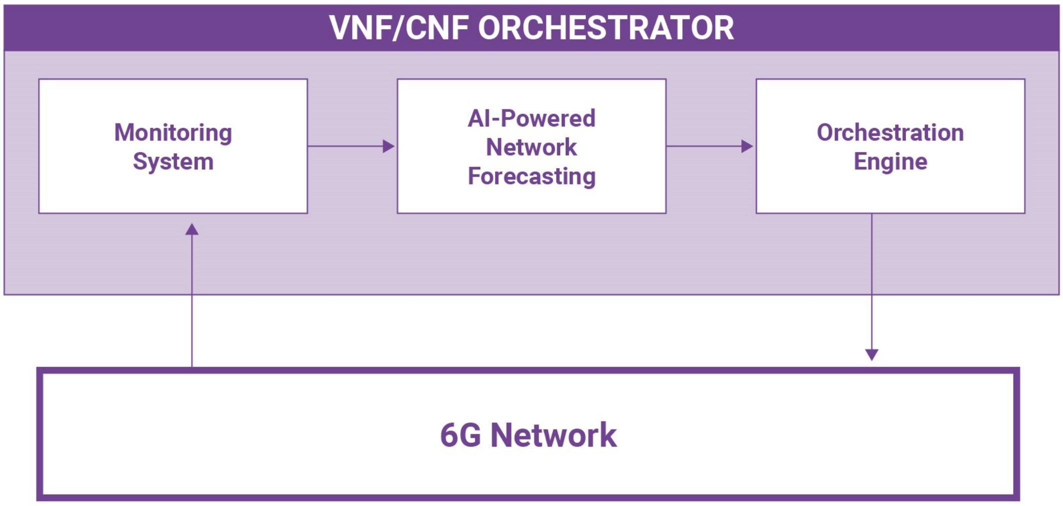


FIGURE 6.12: A novel framework for CNF orchestration, with AI as a native component of the architecture.

Key components of the CNF Orchestrator include:

- **Monitoring System**: Continuous monitoring is essential for optimizing resource management and enhancing orchestration decisions. A monitoring tool like Prometheus tracks real-time metrics from CNFs deployed across both terrestrial and satellite nodes. It collects data such as CPU usage, memory consumption, and network bandwidth, which are crucial not only for operational management but also for providing the foundational data necessary for training machine learning models.

- **AI-Powered Network Forecasting**: It manages the full lifecycle of AI/ML operations to enable proactive resource management. It includes services for data collection and storage, database for historical and predicted metrics, and a Machine Learning Platform for executing ML pipelines to forecast resource utilization across terrestrial and satellite environments. This module is shown in detail in Figure 6.13, and each subcomponent is explained in the following section.

- **Orchestration Engine**: This customised orchestrator powered by AI, overrides Kubernetes' default autoscaler to implement a proactive scaling strategy. By utilising forecasting data produced by AI, it facilitates faster container deployments, reduces cold start delays, and prevents both overprovisioning and underprovisioning of resources. This proactive approach ensures that both terrestrial and non-terrestrial resources are allocated efficiently. This is crucial in satellite networks, where rapid resource allocation helps maintain seamless communication despite varying conditions in orbit, and minimizing waste is vital due to the limited computational resources onboard.

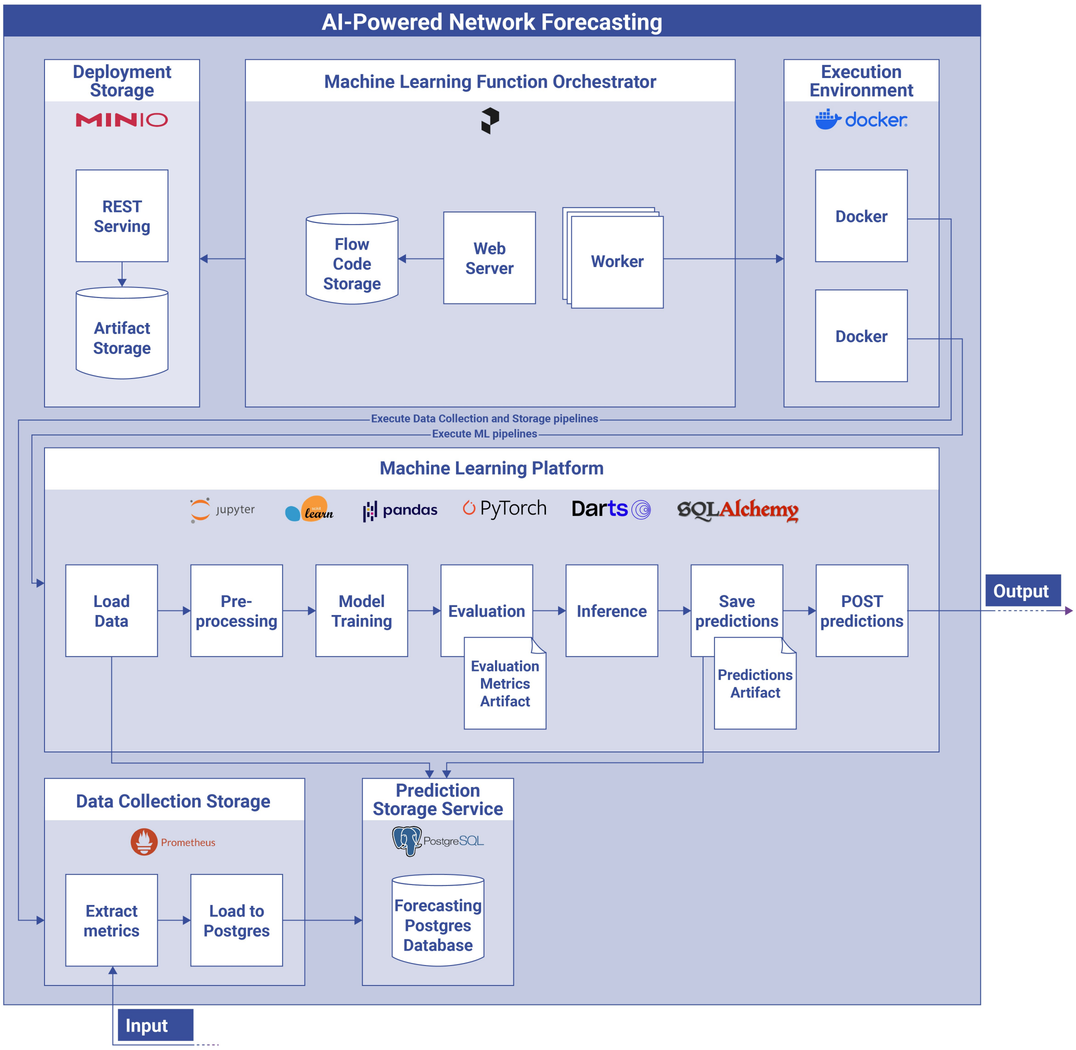


FIGURE 6.13: Detailed architecture of the proposed AI-Powered Network Forecasting component.

6.4.2.2 The AI-powered Network Forecasting

The AI-powered Network Forecasting module (Fig. 6.13) is designed to handle the full lifecycle of AI/ML operations, consisting of several key services:

- **Data Collection Storage**: This service extracts time series of the real-time metrics captured by the Monitoring System and stores them in the Prediction Storage Service for further processing. It ensures that relevant performance metrics are continuously captured from both terrestrial and non terrestrial nodes, and readily available for future analysis.
- **Prediction Storage Service**: This service is responsible for storing both the historical input data (time series) used for Machine Learning and the outputs of the Machine Learning pipelines. A time series database is preferred for this service due to its optimization for handling time-based data, enabling efficient storage and retrieval of both input data and forecasting results.

- **Machine Learning Platform**: This platform manages the execution of ML pipelines, specifically designed to forecast container resource utilization, including CPU, memory, and bandwidth requirements across terrestrial and satellite environments. It adheres to established standards for ML operations [189], encompassing the entire lifecycle, from data loading and preprocessing to model training and inference, ensuring accurate and timely predictions.
- **Machine Learning Function Orchestrator (MLFO)**: The MLFO manages ML pipelines, accessing to saved ML models and artifacts from the Deployment Storage, and executing these pipelines in a distributed execution environment powered by Docker containers. It automates the scheduling and tracking of machine learning experiments, simplifying the management of complex ML systems. By running ML models in isolated execution environments, the MLFO ensures consistency and allows for parallel execution of pipelines.

6.4.3 AI/ML for CNF Resource Forecasting

In this section, we present and analyze several AI/ML models implemented and tested for time series forecasting. These models were selected based on their ability to handle historical data patterns and their applicability to our specific use case of forecasting container resource utilization in network environments.

6.4.3.1 Models

- **ARIMA:** ARIMA is a well-established statistical method for forecasting time series data based on its own historical values and lagged dependencies. It does not involve a traditional training phase like other machine learning models, making it computationally efficient. ARIMA works by combining three key components: autoregression (AR), differencing (to make the series stationary), and moving average (MA). This allows it to model the temporal structure of the data and produce reasonable forecasts. However, while ARIMA is faster to deploy and often effective for simple patterns, its predictive performance tends to be limited when compared to advanced machine learning models, particularly for more complex or nonlinear time series. In our experiments, we used ARIMA as a baseline model to benchmark the performance of more sophisticated ML models.
- **LSTM for Univariate Forecasting:** LSTM, a type of recurrent neural network (RNN), is particularly suited for sequential data, such as time series, as it excels at capturing both long-term dependencies and short-term fluctuations. It uses a gated mechanism to control the flow of information, allowing it to retain or forget specific patterns as needed. For our univariate forecasting task, where the model predicts future values based on the history of a single variable, we trained the LSTM using only past data from that variable. In this experiment,

the LSTM achieved strong predictive performance, demonstrating its adaptability to dynamic time series with shifting patterns.

- **LSTM for Multivariate Forecasting with Covariates:** To further improve forecasting accuracy, we extended the LSTM model to handle multivariate forecasting, where the goal is to predict multiple related time series simultaneously. This approach leverages the relationships between multiple variables, allowing the model to learn from the dependencies across different data streams. Additionally, we introduced covariates, variables that are not the primary targets of the forecast but are closely related and influence the dependent variables. These covariates, identified through data analysis, help the model understand external factors that impact the target variables. The multivariate LSTM with covariates demonstrated significant improvements over the univariate approach, providing more accurate forecasts by considering a wider range of influencing variables.

6.4.3.2 Performance Evaluation

In this section, we evaluate and compare the performances of the different models. The experiments are designed to emulate real-world scenarios, providing a comprehensive assessment of each model's predictive capabilities. To promote transparency and reproducibility, all experiments are open source and available in this GitHub repository [190].

- **Dataset Overview:** All models were trained and evaluated on the same dataset to ensure a fair comparison of performance metrics. The dataset, publicly available [191], contains collected metrics from three types of applications running in a cloud-native environment. Among these, the collection of metrics from the OpenAirInterface 5G Core network function AMF (Access and Mobility Management Function), was found to be particularly relevant for our experiments, as it represents a good example of a network function that can be deployed in containers, analyzed, and predicted. To prepare the dataset for training, we applied a few preprocessing steps. First, we downsampled the data to reduce the number of null values. Following that, we applied interpolation to handle any remaining missing data points in the time series. Once the dataset was cleansed, it was split into two sets: 70% for training and 30% for testing.

- **Experimental Setup:** For each model class introduced above, the experiments simulate a real-world scenario, where historical resource usage metrics are available, and the objective is to predict the resource usage for the next time period. During the testing phase, we iterate through the test set, predicting one step ahead each time. After making each prediction, the true value from the test set is added to the training data, simulating a real-time forecasting scenario where new data continuously becomes available. This allows us to test how well each model adapts to new information.

The models are evaluated using three performance metrics:

- **MAPE**: Measures the accuracy of the predictions as a percentage error.
- **Mean Absolute Error (MAE)**: Provides the average magnitude of the errors in a dataset without considering their direction.
- **Root Mean Square Error (RMSE)**: Highlights larger errors by squaring the error terms, emphasising deviations that may be more impactful in practice.

6.4.4 Results

The results for CPU usage forecasting are shown in Table 6.4. The baseline model, ARIMA (Fig. 6.14), produced a MAPE of 0.32. As expected, ARIMA's performance, while reasonable, is not as strong as more advanced machine learning models. Next, we evaluated the LSTM model for univariate forecasting (Fig.6.15). Interestingly, despite its potential for learning temporal patterns, LSTM showed only a slight improvement over ARIMA when using a single variable (CPU usage) for prediction, rather than significantly outperforming it. This suggests that, in this case, univariate forecasting may not capture the full complexity of the system's behaviour. However, when moving to multivariate forecasting with covariates (Fig.6.16), where the LSTM model was provided with additional related variables alongside CPU usage, we observed a significant improvement in performance. This indicates that there is a clear relationship between CPU usage and the additional covariates. The multivariate model achieved a MAPE of 0.28, demonstrating the benefit of incorporating multiple variables to capture interdependencies, thereby reducing the error and improving the accuracy of the predictions. Since each error metric highlights a different aspect of forecasting performance, it is useful to clarify why their trajectories diverge. Although MAPE is convenient because it is scale-free, it expresses the error as a percentage of the true value. When CPU utilization occasionally drops to very low levels, the denominator in this ratio becomes small, and the percentage error inflates, masking improvements that occur at higher loads. In contrast, MAE and RMSE measure the absolute magnitude of the residuals (with RMSE further penalizing large deviations) and therefore react more strongly when the overall spread of the errors is reduced. Consequently, the multivariate LSTM registers only a modest relative gain in MAPE, yet exhibits a much larger drop in MAE and RMSE. In other words, most of the predictive improvement comes from shrinking the largest absolute errors, particularly during high-load periods, rather than from uniformly reducing percentage error across the full utilization range.

Model	MAPE	MAE	RMSE
ARIMA	0.32	0.24	0.29
LSTM univariate	0.31	0.23	0.30
LSTM multivariate	0.28	0.08	0.11

TABLE 6.4: Comparison of AI/ML Models

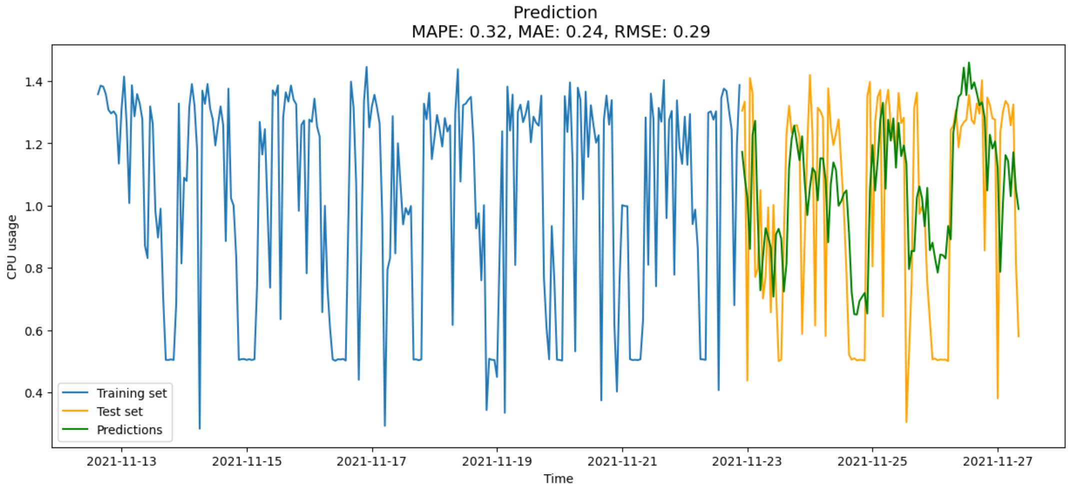


FIGURE 6.14: CPU usage univariate forecast with ARIMA

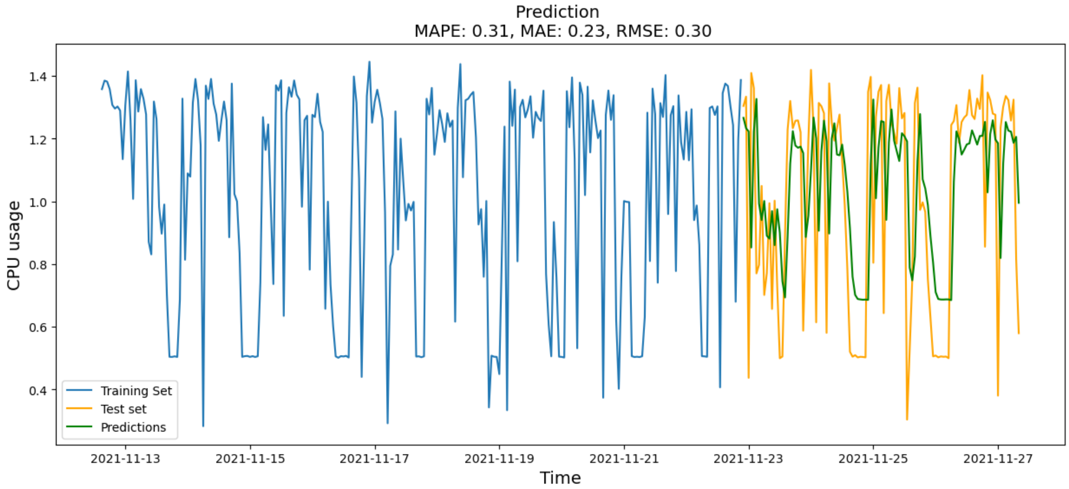


FIGURE 6.15: CPU usage univariate forecast with LSTM

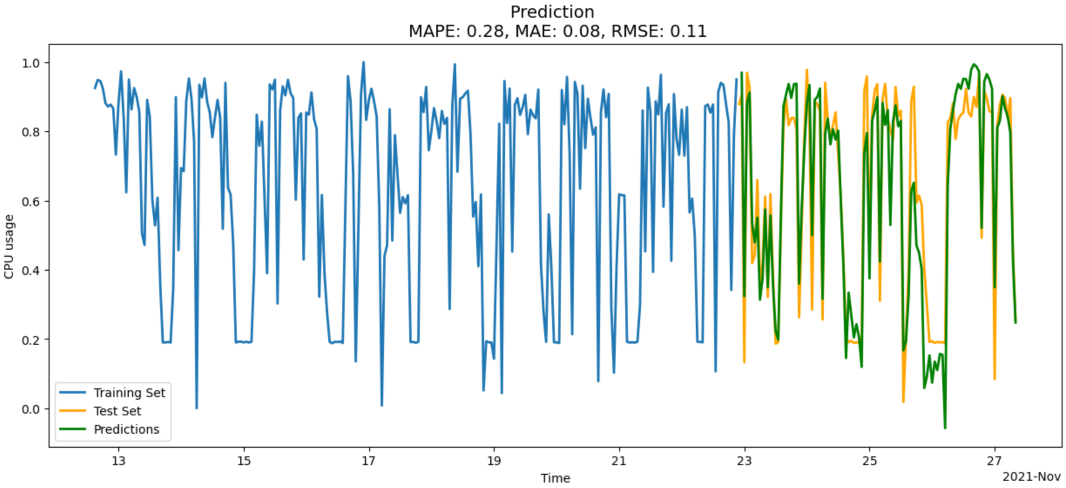


FIGURE 6.16: CPU usage multivariate forecast using covariates with LSTM

Chapter 7

Conclusions

The main objective of this thesis was to investigate the exploitation of the O-RAN paradigm and architecture in NTNs, assessing its potential to deliver flexible, interoperable, and scalable network solutions while addressing the distinct challenges of NTNs, such as dynamic behavior, resource limitations, and operational complexities. Consequently, the initial part of the thesis introduced the O-RAN standard and provided an overview of key SatCom system architectures, with a specific emphasis on NTN architectures as defined by the 3GPP standardization process. The research then focused on three core aspects of O-RAN integration in NTNs: i) architectural design and compatibility with O-RAN; ii) implementation of RAN disaggregation in NTNs; and iii) leveraging near-real-time and real-time control loops to enhance RAN performance.

Specifically, the research proposed three distinct architectural models for incorporating O-RAN principles into NTNs: the Full-gNB, Orbit-split, and Feeder-split architectures. The first foresee a full gNB implemented on a single satellite node, the second splits the gNB between two different satellite nodes, and the third splits the gNB between a ground and a satellite node. These models addressed the unique challenges of satellite-based networks, including high mobility and diverse latency demands.

The following step has been an in-depth exploration of functional splits for RAN disaggregation, analyzing their impact on NTN performance. Eight options, ranging from PHY splits to higher-layer protocol splits, were evaluated. The results revealed that lower-layer splits, such as Options 7 and 6, faced significant challenges due to their high throughput and stringent latency requirements. In contrast, higher-layer splits offered greater flexibility in addressing NTN-specific impairments, but provided less granularity in control. NTN-specific constraints on the split interface, such as interface capacity and latency, were thoroughly examined, and adaptive solutions were proposed to address these challenges.

The study underscored the advantages of virtualization in 3GPP NTN systems, showcasing how software-driven RAN and CN solutions simplify network upgrades to accommodate new 3GPP features. By leveraging the flexibility of virtualization, operators can deploy and update networks efficiently as standards evolve, avoiding the need for hardware-centric upgrades. The chapter examined the NTN standardization journey within 3GPP, focusing on the architectures defined in Releases 17 and 18, ongoing efforts, and the expected advancements in Release 19. Furthermore, it analyzed how virtualization supports the seamless adoption of new functionalities and detailed the software updates required to integrate Release 18 NTN features into a software-based gNB, UE, and CN originally designed for Release 17.

Finally, the study evaluated the integration ofNon-RT, Near-RT,RT control loops within NTN RICs. For Near-RT control loops, a novel AI-based framework for TN-NTN traffic offloading was developed, utilizing deep Q-learning to dynamically allocate resources between terrestrial and non-terrestrial layers. Numerical results showed significant improvements in system throughput, with a 7% increase in spectral efficiency in 70% of cases, achieving optimal load balancing under varying network conditions. Additionally, real-time control loops were used to develop CSI prediction techniques to mitigate the impact of channel aging caused by satellite mobility. A neural network-based CSI prediction model enhanced decoding accuracy and improved the performance of SCMA in NTN environments. Numerical results indicated that the proposed solution achieved high prediction accuracy compared to using outdated CSI directly. Ultimately, the control loop concept was advanced with a proactive CNF orchestration framework tailored for NTN environments, leveraging AI-driven time-series forecasting to efficiently predict and allocate critical resources. This approach reduces delays, optimizes energy consumption, and ensures seamless scalability in dynamic NTN scenarios. Numerical results highlight the effectiveness of multivariate LSTM models, achieving a MAPE of 0.28 and significantly surpassing baseline methods like ARIMA, showcasing the advantages of incorporating multiple variables for enhanced prediction accuracy and resource optimization.

Through these contributions, the thesis established a foundation for O-RAN-based NTN integration. It is important to underline that the integration of O-RAN into NTNs not only addresses the dynamic and complex challenges of satellite-based networks but also significantly contributes to the sustainability goals of 6G. By enabling modular, AI-driven architectures and optimized resource utilization, O-RAN ensures energy-efficient and scalable operations. Its synergy with NTNs fosters a sustainable ecosystem, bridging the digital divide through reliable and affordable connectivity in underserved regions. Consequently, O-RAN's application in NTNs is a key enabler for achieving the ambitious energy efficiency and inclusivity objectives of a sustainable 6G future.

The work pursued in this thesis can be extended in several directions. The main topics left for future works are:

- Absence of NTN specific functional split requirements for interface throughput and latency: in this thesis, the data for interface throughput and latency are taken from 3GPP TR 38.801 [20]. While these figures provide valuable reference points, they are designed for terrestrial networks. For a more specific analysis of RAN disaggregation in NTNs, it is necessary to compute the required interface throughput for each functional split option, considering the particular configurations of NTN RANs. Additionally, the analysis of maximum interface latency should take into account the specific NTN user link characteristics and, where relevant, the processing time of satellite hardware.

- Lack of NTN-specific security frameworks for O-RAN in TN-NTN integration: This thesis does not address security considerations for O-RAN, focusing instead on the architectural and operational aspects of TN-NTN integration. However, the open and distributed nature of O-RAN introduces unique security challenges, particularly in NTN environments with dynamic topologies, satellite mobility, and latency constraints. Future work should explore the development of NTN-specific security frameworks, including lightweight encryption protocols for resource-constrained nodes, robust defense mechanisms against AI/ML adversarial attacks in near-RT RICs, and scalable identity and access management solutions. These measures would ensure the secure and reliable operation of TN-NTN integrated networks within the O-RAN ecosystem.

Bibliography

[1] W. Jiang *et al.*, "The Road Towards 6G: A Comprehensive Survey," *IEEE Open Journal of the Communications Society*, vol. 2, pp. 334–366, 2021, Conference Name: IEEE Open Journal of the Communications Society, ISSN: 2644-125X. DOI: 10.1109/OJCOMS.2021.3057679.

[2] Cisco, *Cisco Annual Internet Report (2018-2023)*, [Online], Mar. 2020. [Online]. Available: https://www.cisco.com/c/en/us/index.html.

[3] The 5G Infrastructure Association, *European Vision for the 6G Network Ecosystem*, [Online], 2021. [Online]. Available: https://5g-ppp.eu/.

[4] D. Soldani, "6G Fundamentals: Vision & Enabling Technologies Towards Trustworthy Solutions & Resilient Systems," *6G Fundamentals*, vol. 18, 2021.

[5] R. Campana, C. Amatetti, and A. Vanelli-Coralli, "O-ran based non-terrestrial networks: Trends and challenges," in *2023 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit)*, 2023, pp. 264–269. DOI: 10.1109/EuCNC/6GSummit58263.2023.10188308.

[6] 6G-NTN, *Deliverable 3.4, "Report on vleo space segment (1st version)"*, 2024.

[7] 6G-NTN, *Deliverable 3.5, "Report on 3d multi layered ntn architecture (2nd version)"*, 2024.

[8] 5G-STARDUST, *Deliverable 3.1, "System requirements analysis and specifications"*, 2023.

[9] 5G-STARDUST, *Deliverable 4.4, "Preliminary report on AI-based radio resource management and onboard processing"*, 2023.

[10] R. Campana, C. Amatetti, and A. Vanelli-Coralli, "RAN Functional Splits in NTN: Architectures and Challenges," *arXiv*, vol. 2309.14810, 2023. DOI: 10.48550/arXiv.2309.14810.

[11] 6G-NTN, *Deliverable 4.2, "Report on 6g-ntn radio controller (1st version)"*, 2024.

[12] H. Shahid, C. Amatetti, R. Campana, *et al.*, "Emerging advancements in 6g ntn radio access technologies: An overview," in *2024 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit)*, 2024, pp. 593–598. DOI: 10.1109/EuCNC/6GSummit60053.2024.10597037.

[13] B. De Filippo, R. Campana, A. Guidotti, C. Amatetti, and A. Vanelli-Coralli, "Cell-free mimo in 6g ntn with ai-predicted csi," in *2024 IEEE 25th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, 2024, pp. 631–635. DOI: 10.1109/SPAWC60668.2024.10694298.

[14] 3GPP, "TR 21.917: Release 17 description; summary of rel-17 work items," Technical report, Jan. 2023.

[15] C. Amatetti, "Nb-iot via non terrestrial networks," Ph.D. dissertation, alma, 2023. [Online]. Available: http://amsdottorato.unibo.it/11058/.

[16] R. Campana, "Nb-iot synchronization procedure analysis," Ph.D. dissertation. [Online]. Available: https://amslaurea.unibo.it/id/eprint/22583/.

[17] NASA, "Catalog of earth satellite orbits," *[online]*, 2009. [Online]. Available: https://earthobservatory.nasa.gov/features/OrbitsCatalog .

[18] O. Kodheli, N. Maturo, S. Chatzinotas, S. Andrenacci, and F. Zimmer, "On the random access procedure of nb-iot non-terrestrial networks," in *2020 10th Advanced Satellite Multimedia Systems Conference and the 16th Signal Processing for Space Communications Workshop (ASMS/SPSC)*, 2020, pp. 1–8. DOI: 10.1109/ASMS/SPSC48805.2020.9268788.

[19] G. Maral, M. Bousquet, and Z. Sun, *Satellite communications systems: Systems, techniques and technology*, Sixth edition. Hoboken, N.J: John Wiley & Sons, 2020, ISBN: 978-1-119-38212-6 978-1-119-38207-2.

[20] 3GPP, "TR38.801, Technical Specification Group Radio Access Network; Study on new radio access technology: Radio access architecture and interfaces (Release 14), V14.0.0," Technical Report, Mar. 2017.

[21] 3GPP, "TS 38.470, Technical Specification Group Radio Access Network; NG-RAN; F1 general aspects and principles (Release 16), V16.3.0," Technical report, Sep. 2020.

[22] 3GPP, *TR38.821 - Solutions for NR to support Non-Terrestrial Networks (NTN)*, Jun. 2021.

[23] A. Guidotti, S. Cioni, G. Colavolpe, *et al.*, "Architectures, standardisation, and procedures for 5G Satellite Communications: A survey," *Computer Networks*, vol. 183, p. 107 588, 2020.

[24] M. Conti, "Satellite systems in the era of 5g internet of things," Ph.D. dissertation, alma, 2021. [Online]. Available: http://amsdottorato.unibo.it/9879/.

[25] 3GPP, "TS 38.410 Technical Specification Group Radio Access Network; NG-RAN; NG general aspects and principles," Technical report, Sep. 2020.

[26] "ETSI EN 302 307-1, V1.1.1, Digital Video Broadcasting (DVB); Second generation framing structure, channel coding and modulation systems for Broadcasting, Interactive Services, News Gathering and other broadband satellite applications; Part 2: DVB-S2X," Tech. Rep., Oct. 2014.

[27] C. Amatetti, A. Guidotti, and A. Vanelli-Coralli, "NTN support for mMTC: Architectural and channel model considerations."

[28] I. Ali, N. Al-Dhahir, and J. E. Hershey, "Doppler characterization for leo satellites," *IEEE Transactions on Communications*, vol. 46, no. 3, pp. 309–313, 1998. DOI: `10.1109/26.662636`.

[29] A. Guidotti, A. Vanelli-Coralli, M. Caus, *et al.*, "Satellite-enabled LTE systems in LEO constellations," in *2017 IEEE International Conference on Communications Workshops (ICC Workshops)*, 2017, pp. 876–881. DOI: `10.1109/ICCW.2017.7962769`.

[30] 3GPP, "TR38.811, Technical Specification Group Radio Access Network; Study on New Radio (NR) to support non-terrestrial networks (Release 15), V15.3.0," Technical Report, Jul. 2020.

[31] ITU, "ITU-R P.676 : Attenuation by atmospheric gases and related effects," Tech. Rep., Aug. 2019.

[32] ITU, "ITU-R P.618 : Propagation data and prediction methods required for the design of Earth-Space telecommunication systems," Tech. Rep., Dec. 2017.

[33] ITU, "ITU-R P.531 : Ionospheric propagation data and prediction methods required for the design of satellite networks and systems," Tech. Rep., Aug. 2019.

[34] 3GPP, "TR38.401, NR NG-RAN; Architecture description, (Release 16), V16.3.0," Technical Specification, Nov. 2020.

[35] "O-ran architecture description 5.00, oran. wg1.o-ran-architecture-description-v05.00 technical specification, july 2021.,"

[36] O-RAN Working Group 3, *O-RAN Near-Real-time RAN Intelligent Controller, E2 Application Protocol 2.00*, O-RAN.WG3.E2AP-v02.00, 2021.

[37] O-RAN Working Group 3, *O-RAN Near-Real-time RAN Intelligent Controller E2 Service Model (E2SM) KPM 2.0*, ORAN-WG3.E2SMKPM-v02.00 Technical Specification, 2021.

[38] *O-RAN E2 Service Model (E2SM), Cell Configuration and Control 1.0*, ORAN.WG3.E2SM-CCC-v01.00 Technical Specification, 2022.

[39] O-RAN Working Group 1, *O-RAN Operations and Maintenance Interface 4.0*, ORAN.WG1.O1-Interface.0-v04.00 Technical Specification, 2020.

[40] R. Enns, M. Bjorklund, J. Schoenwaelder, and A. Bierman, *Network Configuration Protocol (NETCONF)*, Internet Requests for Comments, RFC Editor, RFC 6241, [Online], 2011. [Online]. Available: `http://www.rfc-editor.org/rfc/rfc6241.txt`.

[41] O-RAN Working Group 2, *A1 Interface: General Aspects and Principles 2.03*, ORANWG2.A1.GAP-v02.03 Technical Specification, 2021.

[42] O-RAN Working Group 4, *O-RAN Fronthaul Control, User and Synchronization Plane Specification 7.0*, ORAN-WG4.CUS.0-v07.00 Technical Specification, 2021.

[43] CPRI Consortium, *Common Public Radio Interface: eCPRI Specification V2.0*, ORAN.SFG.O-RAN-Security-Protocols-Specificationsv02.00 Technical Specification, 2019.

[44] IEEE SA, *IEEE Standard for Radio over Ethernet Encapsulations and Mappings*, 2018.

[45] *O-RAN Management Plane Specification 7.0*, Alfter, Germany, Technical Specification, 2021.

[46] "Ai/ml workflow description and requirements, o-ran.wg2.aiml- v01.03, 2021,"

[47] S. D'Oro *et al.*, "OrchestRAN: Network Automation through Orchestrated Intelligence in the Open RAN," in *IEEE INFOCOM 2022 - IEEE Conference on Computer Communications*, ISSN: 2641-9874, May 2022, pp. 270–279. DOI: `10.1109/INFOCOM48880.2022.9796744`.

[48] T. Pamuklu *et al.*, "Reinforcement Learning Based Dynamic Function Splitting in Disaggregated Green Open RANs," in *ICC 2021 - IEEE International Conference on Communications*, ISSN: 1938-1883, Jun. 2021, pp. 1–6. DOI: `10.1109/ICC42927.2021.9500721`.

[49] E. Amiri *et al.*, "Optimizing Virtual Network Function Splitting in Open-RAN Environments," in *2022 IEEE 47th Conference on Local Computer Networks (LCN)*, ISSN: 0742-1303, Sep. 2022, pp. 422–429. DOI: `10.1109/LCN53696.2022.9843310`.

[50] H. Kumar *et al.*, "O-RAN based proactive ANR optimization," in *2020 IEEE Globecom Workshops (GC Wkshps*, Dec. 2020, pp. 1–4. DOI: `10.1109/GCWkshps50303.2020.9367582`.

[51] y. Cao *et al.*, "User Access Control in Open Radio Access Networks: A Federated Deep Reinforcement Learning Approach," *IEEE Transactions on Wireless Communications*, vol. 21, no. 6, pp. 3721–3736, Jun. 2022, Conference Name: IEEE Transactions on Wireless Communications, ISSN: 1558-2248. DOI: `10.1109/TWC.2021.3123500`.

[52] M. Motalleb *et al.*, "Resource Allocation in an Open RAN System using Network Slicing," *IEEE Transactions on Network and Service Management*, pp. 1–1, 2022, Conference Name: IEEE Transactions on Network and Service Management, ISSN: 1932-4537. DOI: `10.1109/TNSM.2022.3205415`.

[53] *Intelligent 5G L2 MAC Scheduler: Powered by Capgemini NetAnticipate 5G on Intel Architecture*, en. [Online]. Available: `https://builders.intel.com/docs/networkbuilders/intelligent-5g-l2-mac-scheduler-powered-by-capgemini-netanticipate-/5g-on-intel-architecture-v13.pdf` (visited on 10/18/2022).

[54] S. Lagén *et al.*, "Modulation Compression in Next Generation RAN: Air Interface and Fronthaul Trade-offs," *IEEE Communications Magazine*, vol. 59, no. 1, pp. 89–95, Jan. 2021, Conference Name: IEEE Communications Magazine, ISSN: 1558-1896. DOI: 10.1109/MCOM.001.2000453.

[55] T. Hewavithana *et al.*, "Overcoming Channel Aging in Massive MIMO Basestations With Open RAN Fronthaul," in *2022 IEEE Wireless Communications and Networking Conference (WCNC)*, ISSN: 1558-2612, Apr. 2022, pp. 2577–2582. DOI: 10.1109/WCNC51071.2022.9771697.

[56] C. Shen *et al.*, "Security Threat Analysis and Treatment Strategy for ORAN," in *2022 24th International Conference on Advanced Communication Technology (ICACT)*, ISSN: 1738-9445, Feb. 2022, pp. 417–422. DOI: 10.23919/ICACT53585.2022.9728862.

[57] D. Dik *et al.*, "Transport Security Considerations for the Open-RAN Fronthaul," in *2021 IEEE 4th 5G World Forum (5GWF)*, Oct. 2021, pp. 253–258. DOI: 10.1109/5GWF52925.2021.00051.

[58] L. Bertizzolo *et al.*, "Streaming from the Air: Enabling Drone-sourced Video Streaming Applications on 5G Open-RAN Architectures," *IEEE Transactions on Mobile Computing*, pp. 1–1, 2021, Conference Name: IEEE Transactions on Mobile Computing, ISSN: 1558-0660. DOI: 10.1109/TMC.2021.3129094.

[59] C. Pham *et al.*, "When RAN Intelligent Controller in O-RAN Meets Multi-UAV Enable Wireless Network," *IEEE Transactions on Cloud Computing*, pp. 1–15, 2022, Conference Name: IEEE Transactions on Cloud Computing, ISSN: 2168-7161. DOI: 10.1109/TCC.2022.3193939.

[60] L. Larsen *et al.*, "Xhaul Latency Dimensioning of 5G Drone Control," in *2022 International Conference on Unmanned Aircraft Systems (ICUAS)*, ISSN: 2575-7296, Jun. 2022, pp. 762–771. DOI: 10.1109/ICUAS54217.2022.9836059.

[61] R. Smith *et al.*, "An O-RAN Approach to Spectrum Sharing Between Commercial 5G and Government Satellite Systems," in *MILCOM 2021 - 2021 IEEE Military Communications Conference (MILCOM)*, ISSN: 2155-7586, Nov. 2021, pp. 739–744. DOI: 10.1109/MILCOM52596.2021.9653140.

[62] N. Siasi, N. I. Sulieman, and R. D. Gitlin, "Ultra-reliable nfv-based 5g networks using diversity and network coding," in *2018 IEEE 19th Wireless and Microwave Technology Conference (WAMICON)*, 2018, pp. 1–4. DOI: 10.1109/WAMICON.2018.8363900.

[63] W. Azariah, F. A. Bimo, C. Lin, R. Cheng, N. Nikaein, and R. Jana, "A survey on open radio access networks: Challenges, research directions, and open source approaches," *Sensors*, vol. 24, no. 3, p. 1038, 2024. DOI: 10.3390/s24031038.

[64] Seeram, L. Feltrin, M. Ozger, S. Zhang, and C. Cavdar, "Feasibility study of function splits in ran architectures with leo satellites," 2024.

[65] M Saravanan and R. Pratap Sircar, "Quantum evolutionary algorithm for scheduling resources in virtualized 5g ran environment," in *2021 IEEE 4th 5G World Forum (5GWF)*, 2021, pp. 111–116. DOI: 10.1109/5GWF52925.2021.00027.

[66] L. Lo Schiavo, G. García-Avilés, A. García-Saavedra, *et al.*, "CloudRIC: Open radio access network (o-ran) virtualization with shared heterogeneous computing," 2024. DOI: 10.5281/zenodo.13889362.

[67] Ericsson. "Mobile network traffic q3 2024 - ericsson mobility report." Accessed 2 May 2025. (2024), [Online]. Available: https://www.ericsson.com/en/reports-and-papers/mobility-report/dataforecasts/mobile-traffic-update.

[68] A. Jalalian, S. Yousefi, and T. H. Kunz, "Network slicing in virtualized 5g core with vnf sharing," *Journal of Network and Computer Applications*, vol. 215, p. 103 631, 2023. DOI: 10.1016/j.jnca.2023.103631.

[69] A. A. Barakabitze, A. Ahmad, R. Mijumbi, and A. Hines, "5g network slicing using sdn and nfv: A survey of taxonomy, architectures and future challenges," *Computer Networks*, vol. 167, p. 106 984, 2020, ISSN: 1389-1286. DOI: https://doi.org/10.1016/j.comnet.2019.106984. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1389128619304773.

[70] o. A.Guidotti, "The path to 5g-advanced and 6g non-terrestrial network systems," *arXiv preprint arXiv:2209.11535*, 2022.

[71] 3GPP, "TR 22.822: Study on using satellite access in 5G," Technical report, Jul. 2018.

[72] 3GPP, "TS 22.261: Service requirements for the 5G system," Technical specification, Jul. 2018.

[73] 3GPP, "TR 28.808: Study on management and orchestration aspects of integrated satellite components in a 5G network," Technical report, Jul. 2018.

[74] 3GPP, "TS 22.261: Service requirements for the 5g system," Technical specification, Sep. 2020.

[75] 3GPP, *TS 38.314: Layer 2 Measurements*, Technical Specification, Available at: https://www.3gpp.org/.

[76] 3GPP, *TS 38.321: Medium Access Control (MAC) Protocol Specification*, Technical Specification, Available at: https://www.3gpp.org/.

[77] 3GPP, "TR38.821, Solutions for NR to support non-terrestrial networks (NTN), (Release 17), V16.0.0," Technical Specification, Jan. 2020.

[78] 3GPP, "TS 38.101-5: NR; user equipment (ue) radio transmission and reception; part 5: Satellite access radio frequency (rf) and performance requirements," Technical specification, Sep. 2024.

[79] 3GPP, "TS 38.108: NR; satellite access node radio transmission and reception," Technical specification, Sep. 2024.

[80] 3GPP, "TS 38.104: NR; base station (bs) radio transmission and reception," Technical specification, Sep. 2024.

[81] 3GPP, "TS 38.133: NR; requirements for support of radio resource management," Technical specification, Oct. 2024.

[82] 3GPP, "TS 23.737: Study on architecture aspects for using satellite access in 5g," Technical specification, Mar. 2021.

[83] 3GPP, "TS 23.501: System architecture for the 5g system (5gs)," Technical specification, Sep. 2024.

[84] 3GPP, "TS 23.502: Procedures for the 5g system (5gs)," Technical specification, Sep. 2024.

[85] 3GPP, "TS 23.503: Policy and charging control framework for the 5g system (5gs); stage 2," Technical specification, Sep. 2024.

[86] 3GPP, "TS 23.122: Non-Access-Stratum (nas) functions related to mobile station (ms) in idle mode," Technical specification, Sep. 2024.

[87] 3GPP, "TS 24.501: Non-Access-Stratum (nas) protocol for 5g system (5gs); stage 3," Technical specification, Sep. 2024.

[88] 3GPP, *TS 38.331: Radio Resource Control (RRC); Protocol Specification*, Technical Specification, Available at: https://www.3gpp.org/.

[89] 3GPP, "TS 38.213: NR; physical layer procedures for control," Technical specification, Sep. 2024.

[90] A. Vanelli-Coralli, N. Chuberre, G. Masini, A. Guidotti, and M. El Jaafari, "Nr ntn architecture and network protocols," in *5G Non-Terrestrial Networks: Technologies, Standards, and System Design*. 2024, pp. 91–109. DOI: 10.1002/9781119891185.ch4.

[91] 3GPP, *TS 38.300: NR and NG-RAN Overall Description*, Technical Specification, Available at: https://www.3gpp.org/.

[92] C. B. E. J. M. C. G. R. L. Cioni S Lin X, "Physical layer enhancements in 5g-nr for direct access via satellite systems," *International Journal of Satellite Communications and Networking*, vol. 41, no. 3, pp. 262–275, Jun. 2023. DOI: 10.1002/sat.1461.

[93] 3GPP, *TS 38.304: User Equipment (UE) Procedures in Idle Mode and in RRC Inactive State*, Technical Specification, Available at: https://www.3gpp.org/.

[94] 3GPP, "TS 38.306: NR; user equipment (ue) radio access capabilities," Technical specification, Sep. 2024.

[95] 3GPP, "TS 38.413: NG-RAN; ng application protocol (ngap)," Technical specification, Sep. 2024.

[96] 3GPP, “TS 38.424: NG-RAN; xn data transport,” Technical specification, Apr. 2024.

[97] 3GPP, “TR 21.918: Release 18 description; summary of rel-18 work items,” Technical report, Jul. 2024.

[98] 3GPP, “TR 23.700-27: Study on 5g system with satellite backhaul (release 18),” Technical report, Dec. 2022.

[99] 3GPP, “TS 29.502: 5G system; session management services; stage 3,” Technical specification, Sep. 2024.

[100] 3GPP, “TS 29.508: 5G system; session management event exposure service; stage 3,” Technical specification, Sep. 2024.

[101] 3GPP, “TS 29.512: 5G system; session management policy control service; stage 3,” Technical specification, Sep. 2024.

[102] 3GPP, “TS 29.514: 5G system; policy authorization service; stage 3,” Technical specification, Sep. 2024.

[103] 3GPP, “TS 29.518: 5G system; access and mobility management services; stage 3,” Technical specification, Sep. 2024.

[104] 3GPP, “TS 29.523: 5G system; policy control event exposure service; stage 3,” Technical specification, Jun. 2024.

[105] 3GPP, “TS 29.525: 5G system; ue policy control service; stage 3,” Technical specification, Sep. 2024.

[106] 3GPP, “TS 29.571: 5G system; common data types for service based interfaces; stage 3,” Technical specification, Sep. 2024.

[107] 3GPP, “TS 28.541: Management and orchestration; 5g network resource model (nrm); stage 2 and stage 3,” Technical specification, Sep. 2024.

[108] 3GPP, “TS 28.552: Management and orchestration; 5g performance measurements,” Technical specification, Sep. 2024.

[109] 3GPP, “TR 23.700-28: Study on integration of satellite components in the 5g architecture; phase 2,” Technical report, Mar. 2023.

[110] 3GPP, “TS 38.215: NR; physical layer measurements (release 18),” Technical specification, Jun. 2024.

[111] 3GPP, “TR 38.882: Study on requirements and use cases for network verified ue location for non-terrestrial-networks (ntn) in nr,” Technical report, Jun. 2022.

[112] 3GPP, “TS 37.355: LTE positioning protocol (lpp),” Technical specification, Sep. 2024.

[113] 3GPP, “TR 22.865: Study on satellite access phase 3; (release 19),” Technical report, Dec. 2023.

[114] 3GPP, "TR 23.700-29: Study on integration of satellite components in the 5g architecture; phase 3," Technical report, Jun. 2024.

[115] 3GPP, "TR 28.874: Study on management aspects of ntn phase 2," Technical report, Sep. 2024.

[116] 3GPP, "TS 28.530: Management and orchestration; concepts, use cases and requirements," Technical specification, Sep. 2024.

[117] 3GPP, "R1-2407771: Work plan for rel-19 nr$_{n}tn_{p}h3$," Technical report, Oct. 2024.

[118] A. Guidotti, "Eager white paper:Architectures, services, and technologies towards 6G Non-Terrestrial Networks," White paper, 2023.

[119] 3GPP, "TR38.473, Technical Specification Group Radio Access Network; f1 application protocol (f1ap) (Release 17), V17.2.0," Technical Specification, Oct. 2022.

[120] L. M. P. Larsen, A. Checko, and H. L. Christiansen, "A survey of the functional splits proposed for 5g mobile crosshaul networks," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 1, pp. 146–172, 2019. DOI: 10.1109/COMST.2018.2868805.

[121] N. DOCOMO, *R3-162102: Cu-du split: Refinement for annex a*, 2016.

[122] CMCC, *R3-161813: Transport requirement for cu&du functional splits options*, 2016.

[123] N. Alliance, "Further study on critical c-ran technologies," *Next Generation Mobile Networks*, vol. 64, 2015.

[124] A. Tukmanov, M. A. Lema, I. A. Mings, *et al.*, "Fronthauling for 5g and beyond," in IET, 2017.

[125] A. Guidotti, C. Sacchi, and A. Vanelli-Coralli, "Feeder link precoding for future broadcasting services," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 58, no. 4, pp. 3126–3146, 2022.

[126] 3GPP, "TR 38.213, Physical layer procedures for control," Technical Specification, Jul. 2023.

[127] 3GPP, "TR38.821, Technical Specification Group Radio Access Network; Solutions for NR to support non-terrestrial networks (NTN) (Release 16), V16.0.0," Technical Report, Dec. 2019.

[128] F. Dong, T. Huang, Y. Zhang, C. Sun, and C. Li, "A computation offloading strategy in leo constellation edge cloud network," *Electronics*, vol. 11, no. 13, 2022, ISSN: 2079-9292. DOI: 10.3390/electronics11132024. [Online]. Available: https://www.mdpi.com/2079-9292/11/13/2024.

[129] A. Lacava, M. Polese, R. Sivaraj, *et al.*, "Programmable and customized intelligence for traffic steering in 5g networks using open ran architectures," *IEEE Transactions on Mobile Computing*, vol. 23, no. 4, pp. 2882–2897, 2024. DOI: 10.1109/TMC.2023.3266642.

[130] H.-M. Yoo, J.-S. Rhee, S.-Y. Bang, and E.-K. Hong, "Load balancing algorithm running on open ran ric," in *2022 13th International Conference on Information and Communication Technology Convergence (ICTC)*, 2022, pp. 1226–1228. DOI: 10.1109/ICTC55196.2022.9952635.

[131] B. Agarwal, M. A. Togou, M. Ruffini, and G.-M. Muntean, "Qoe-driven optimization in 5g o-ran-enabled hetnets for enhanced video service quality," *IEEE Communications Magazine*, vol. 61, no. 1, pp. 56–62, 2023. DOI: 10.1109/MCOM.003.2200229.

[132] C. Ge, S. Xia, Q. Chen, and F. Adachi, "2-layer interference coordination framework based on graph coloring algorithm for a cellular system with distributed mu-mimo," *IEEE Transactions on Vehicular Technology*, vol. 72, no. 3, pp. 3557–3568, 2023. DOI: 10.1109/TVT.2022.3219411.

[133] Z. Liu, J. Wu, W. Lu, *et al.*, "A general downlink frequency-domain icic framework for next-generation ran," in *2022 IEEE Globecom Workshops (GC Wkshps)*, 2022, pp. 980–985. DOI: 10.1109/GCWkshps56602.2022.10008558.

[134] F. Mungari, "An rl approach for radio resource management in the o-ran architecture," in *2021 18th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*, 2021, pp. 1–2. DOI: 10.1109/SECON52354.2021.9491579.

[135] F. Rezazadeh, L. Zanzi, F. Devoti, *et al.*, "A multi-agent deep reinforcement learning approach for ran resource allocation in o-ran," in *IEEE INFOCOM 2023 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, 2023, pp. 1–2. DOI: 10.1109/INFOCOMWKSHPS57453.2023.10226154.

[136] B. H. Prananto, Iskandar, and A. Kurniawan, "A new method to improve frequent-handover problem in high-mobility communications using ric and machine learning," *IEEE Access*, vol. 11, pp. 72 281–72 294, 2023. DOI: 10.1109/ACCESS.2023.3294990.

[137] Z. Mahrez, M. B. Driss, E. Sabir, W. Saad, and E. Driouch, "Benchmarking of anomaly detection techniques in o-ran for handover optimization," in *2023 International Wireless Communications and Mobile Computing (IWCMC)*, 2023, pp. 119–125. DOI: 10.1109/IWCMC58020.2023.10183347.

[138] V. H. L. Lopes, G. M. Almeida, A. Klautau, and K. Cardoso, "A coverage-aware vnf placement and resource allocation approach for disaggregated vrans," in *GLOBECOM 2022 - 2022 IEEE Global Communications Conference*, 2022, pp. 185–190. DOI: 10.1109/GLOBECOM48099.2022.10000776.

[139] A. El-Amine, M. Iturralde, H. A. Haj Hassan, and L. Nuaymi, "A distributed q-learning approach for adaptive sleep modes in 5g networks," in *2019 IEEE Wireless Communications and Networking Conference (WCNC)*, 2019, pp. 1–6. DOI: `10.1109/WCNC.2019.8885818`.

[140] I. Rego, L. Medeiros, P. Alves, *et al.*, "Prototyping near-real time ric o-ran xapps for flexible ml-based spectrum sensing," in *2022 IEEE Conference on Network Function Virtualization and Software Defined Networks (NFV-SDN)*, 2022, pp. 137–142. DOI: `10.1109/NFV-SDN56302.2022.9974940`.

[141] S. D'Oro, M. Polese, L. Bonati, H. Cheng, and T. Melodia, "Dapps: Distributed applications for real-time inference and control in o-ran," *IEEE Communications Magazine*, vol. 60, no. 11, pp. 52–58, 2022. DOI: `10.1109/MCOM.002.2200079`.

[142] o. M. Polese, *Understanding O-RAN: Architecture, Interfaces, Algorithms, Security, and Research Challenges*, arXiv:2202.01032 [cs, eess], Aug. 2022. DOI: `10.48550/arXiv.2202.01032`. [Online]. Available: `http://arxiv.org/abs/2202.01032` (visited on 10/28/2022).

[143] L. Bonati, M. Polese, S. D'Oro, S. Basagni, and T. Melodia, "Open, programmable, and virtualized 5g networks: State-of-the-art and the road ahead," *Computer Networks*, vol. 182, p. 107 516, 2020, ISSN: 1389-1286. DOI: `https://doi.org/10.1016/j.comnet.2020.107516`. [Online]. Available: `https://www.sciencedirect.com/science/article/pii/S1389128620311786`.

[144] M. Dryjański, Kułacz, and A. Kliks, "Toward Modular and Flexible Open RAN Implementations in 6G Networks: Traffic Steering Use Case and O-RAN xAPPs," *Sensors*, vol. 21, no. 24, p. 8173, 2021. DOI: `10.3390/s21248173`.

[145] O. Orhan, V. N. Swamy, T. Tetzlaff, M. Nassar, H. Nikopour, and S. Talwar, "Connection Management xAPP for O-RAN RIC: A Graph Neural Network and Reinforcement Learning Approach," in *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, IEEE, 2021, pp. 936–941. DOI: `10.1109/ICMLA52953.2021.00156`.

[146] H. Erdol, X. Wang, P. Li, *et al.*, "Federated meta-learning for traffic steering in o-ran," in *2022 IEEE 96th Vehicular Technology Conference (VTC2022-Fall)*, 2022, pp. 1–7. DOI: `10.1109/VTC2022-Fall57202.2022.10012789`.

[147] R. Rashad and S. Sudhir, "Load Balancing Technique Based on Network Segmentation and Adaptive Sleep Scheduling for 5G-IoT Networks," 2022, Preprint. DOI: `10.21203/rs.3.rs-825633/v1`.

[148] K. Addali and M. Kadoch, "Enhanced Mobility Load Balancing Algorithm for 5G Small Cell Networks," in *2019 IEEE Canadian Conference of Electrical and Computer Engineering (CCECE)*, IEEE, 2019, pp. 1–5. DOI: `10.1109/CCECE.2019.8861651`.

[149] K. M. Addali, S. Y. B. Melhem, Y. Khamayseh, Z. Zhang, and M. Kadoch, "Dynamic Mobility Load Balancing for 5G Small-cell Networks Based on Utility Functions," *IEEE Access*, vol. 7, pp. 126 998–127 011, 2019. DOI: 10.1109/ACCESS.2019.2939158.

[150] S. Saibharath, S. Mishra, and C. Hota, "Swap-based Load Balancing for Fairness in Radio Access Networks," *IEEE Wireless Communications Letters*, vol. 10, no. 11, pp. 2412–2416, 2021. DOI: 10.1109/LWC.2021.3109601.

[151] B. Ma, B. Yang, Y. Zhu, and J. Zhang, "Context-aware Proactive 5G Load Balancing and Optimization for Urban Areas," *IEEE Access*, vol. 8, pp. 8405–8417, 2020. DOI: 10.1109/ACCESS.2020.2963957.

[152] S. M. Shahid, Y. T. Seyoum, S. H. Won, and S. Kwon, "Load balancing for 5g integrated satellite-terrestrial networks," *IEEE Access*, vol. 8, pp. 132 144–132 156, 2020. DOI: 10.1109/ACCESS.2020.3010059.

[153] Z. Lin, Z. Ni, L. Kuang, C. Jiang, and Z. Huang, "Multi-satellite Beam Hopping based on Load Balancing and Interference Avoidance for NGSO Satellite Communication Systems," *IEEE Transactions on Communications*, vol. 71, no. 1, pp. 282–295, 2022. DOI: 10.1109/TCOMM.2022.3181769.

[154] N. Badini, M. Jaber, M. Marchese, and F. Patrone, "Reinforcement Learning-Based Load Balancing Satellite Handover Using NS-3," in *ICC 2023 - IEEE International Conference on Communications*, IEEE, 2023, pp. 2595–2600. DOI: 10.1109/ICC2023.XXXXXXX.

[155] 3GPP, "Study on Channel Model for Frequencies from 0.5 to 100 GHz," 3rd Generation Partnership Project (3GPP), Technical Report (TR).

[156] 3GPP, "Study on International Mobile Telecommunications (IMT) Parameters for 6.425 - 7.025 GHz, 7.025 - 7.125 GHz and 10.0 - 10.5 GHz," 3rd Generation Partnership Project (3GPP), Technical Report (TR).

[157] C. Amatetti, M. Alsenwi, H. Chougrani, A. Vanelli-Coralli, M. R. Palattella, *et al.*, "A novel twofold approach to enhance nb-iot mac procedure in ntn," *IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS*, vol. 1, pp. 1–12, 2024.

[158] A. Guidotti, A. Vanelli-Coralli, and C. Amatetti, "Federated cell-free mimo in nonterrestrial networks: Architectures and performance," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 60, no. 3, pp. 3319–3347, 2024. DOI: 10.1109/TAES.2024.3362769.

[159] D. G. Riviello, B. Ahmad, A. Guidotti, and A. Vanelli-Coralli, "Joint graph-based user scheduling and beamforming in leo-mimo satellite communication systems," in *2022 11th Advanced Satellite Multimedia Systems Conference and the 17th Signal Processing for Space Communications Workshop (ASMS/SPSC)*, 2022, pp. 1–8. DOI: 10.1109/ASMS/SPSC55670.2022.9914723.

[160] B. Ahmad, D. G. Riviello, A. Guidotti, and A. Vanelli-Coralli, "Graph-based user scheduling algorithms for leo-mimo non-terrestrial networks," in *2023 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit)*, 2023, pp. 270–275. DOI: `10.1109/EuCNC/6GSummit58263.2023.10188287`.

[161] Y. Mao, O. Dizdar, B. Clerckx, R. Schober, P. Popovski, and H. V. Poor, "Rate-splitting multiple access: Fundamentals, survey, and future research trends," *IEEE Communications Surveys & Tutorials*, vol. 24, no. 4, pp. 2073–2126, 2022.

[162] A. Schroeder, M. Roeper, D. Wuebben, B. Matthiesen, P. Popovski, and A. Dekorsy, "A comparison between rsma, sdma, and oma in multibeam leo satellite systems," in *WSA & SCC 2023; 26th International ITG Workshop on Smart Antennas and 13th Conference on Systems, Communications, and Coding*, 2023, pp. 1–6.

[163] B. Clerckx, Y. Mao, Z. Yang, *et al.*, "Multiple access techniques for intelligent and multi-functional 6g: Tutorial, survey, and outlook," *arXiv preprint arXiv:2401.01433*, 2024.

[164] R. Hoshyar, F. P. Wathan, and R. Tafazolli, "Novel low-density signature for synchronous cdma systems over awgn channel," *IEEE Transactions on Signal Processing*, vol. 56, no. 4, pp. 1616–1626, 2008.

[165] R. Hoshyar, R. Razavi, and M. Al-Imari, "Lds-ofdm an efficient multiple access technique," in *2010 IEEE 71st Vehicular Technology Conference*, IEEE, 2010, pp. 1–5.

[166] H. Nikopour and H. Baligh, "Sparse code multiple access," in *2013 IEEE 24th annual international symposium on personal, indoor, and mobile radio communications (PIMRC)*, IEEE, 2013, pp. 332–336.

[167] J. Zeng, B. Li, X. Su, L. Rong, and R. Xing, "Pattern division multiple access (pdma) for cellular future radio access," in *2015 international conference on wireless communications & signal processing (WCSP)*, IEEE, 2015, pp. 1–5.

[168] 3GPP, "'Initial LLS results for UL non-orthogonal multiple access', documentR1-164329,3GPPTSG-RANWG185," Tech. Rep., 2016.

[169] Z. Yuan, G. Yu, W. Li, Y. Yuan, X. Wang, and J. Xu, "Multi-user shared access for internet of things," in *2016 IEEE 83rd Vehicular Technology Conference (VTC Spring)*, IEEE, 2016, pp. 1–5.

[170] LG Electronics, ""Considerations on DL/UL multiple access for NR," documentR1-162517,3GPPTSG-RANWG184b," Technical Report, 2014.

[171] Nokia, Alcatel-Lucent, Shanghai Bell, "Non-orthogonal multiple access for newer adio," document R1-165019,3GPP TS G-RANWG185," Technical Report, 2016.

[172] L. Ping, L. Liu, K. Wu, and W. Leung, "Interleave division multiple-access," *IEEE Transactions on Wireless Communications*, vol. 5, no. 4, pp. 938–947, 2006.

[173] Q. Xiong, C. Qian, B. Yu, and C. Sun, "Advanced noma scheme for 5g cellular network: Interleave-grid multiple access," in *2017 IEEE Globecom Workshops (GC Wkshps)*, 2017, pp. 1–5.

[174] Y. Zhang, Y. Wu, A. Liu, X. Xia, T. Pan, and X. Liu, "Deep learning-based channel prediction for leo satellite massive mimo communication system," *IEEE Wireless Communications Letters*, vol. 10, no. 8, pp. 1835–1839, 2021.

[175] Y. Zhang, A. Liu, P. Li, and S. Jiang, "Deep learning (dl)-based channel prediction and hybrid beamforming for leo satellite massive mimo system," *IEEE Internet of Things Journal*, vol. 9, no. 23, pp. 23 705–23 715, 2022.

[176] G.-Y. Chang, C.-K. Hung, and C.-H. Chen, "A csi prediction scheme for satellite-terrestrial networks," *IEEE Internet of Things Journal*, 2022.

[177] M. Gao, T. Liao, and Y. Lu, "Fully connected feed-forward neural networks based csi feedback algorithm," *China Communications*, vol. 18, no. 1, pp. 43–48, 2021. DOI: `10.23919/JCC.2021.01.004`.

[178] W. Jiang and H. D. Schotten, "Deep learning for fading channel prediction," *IEEE Open Journal of the Communications Society*, vol. 1, pp. 320–330, 2020. DOI: `10.1109/OJCOMS.2020.2982513`.

[179] S. Agarwal, M. A. Rodriguez, and R. Buyya, "A reinforcement learning approach to reduce serverless function cold start frequency," in *2021 IEEE/ACM 21st International Symposium on Cluster, Cloud and Internet Computing (CCGrid)*, IEEE, 2021, pp. 797–803.

[180] Z. Zhong, M. Xu, M. A. Rodriguez, C. Xu, and R. Buyya, "Machine learning-based orchestration of containers: A taxonomy and future directions," *ACM Computing Surveys (CSUR)*, vol. 54, no. 10s, pp. 1–35, 2022.

[181] K. B. Letaief, W. Chen, Y. Shi, J. Zhang, and Y.-J. A. Zhang, "The roadmap to 6g: Ai empowered wireless networks," *IEEE communications magazine*, vol. 57, no. 8, pp. 84–90, 2019.

[182] H.-G. Kim, J.-H. Yoo, and J. W.-K. Hong, "Ai-based network function virtualization orchestration," in *NOMS 2024-2024 IEEE Network Operations and Management Symposium*, IEEE, 2024, pp. 1–5.

[183] Y. Xiao, Q. Zhang, F. Liu, *et al.*, "Nfvdeep: Adaptive online service function chain deployment with deep reinforcement learning," in *Proceedings of the International Symposium on Quality of Service*, 2019, pp. 1–10.

[184] V. Sciancalepore, F. Z. Yousaf, and X. Costa-Perez, "Z-torch: An automated nfv orchestration and monitoring solution," *IEEE Transactions on Network and Service Management*, vol. 15, no. 4, pp. 1292–1306, 2018.

[185] R. Bianchini, M. Fontoura, E. Cortez, *et al.*, "Toward ml-centric cloud platforms," *Communications of the ACM*, vol. 63, no. 2, pp. 50–59, 2020.

[186] K. Ali and M. Jammal, "Proactive vnf scaling and placement in 5g o-ran using ml," *IEEE Transactions on Network and Service Management*, 2023.

[187] N. Kazemifard and V. Shah-Mansouri, "Minimum delay function placement and resource allocation for open ran (o-ran) 5g networks," *Computer Networks*, vol. 188, p. 107 809, 2021.

[188] L. Toka, G. Dobreff, B. Fodor, and B. Sonkoly, "Machine learning-based scaling management for kubernetes edge clusters," *IEEE Transactions on Network and Service Management*, vol. 18, no. 1, pp. 958–972, 2021.

[189] Google, *Pipeline ml*. [Online]. Available: https://developers.google.com/machine-learning/managing-ml-projects/pipelines.

[190] A. Piemonti, *6g-ntn resource forecasting*, https://github.com/martel-innovate/6G-NTN-resource-forecasting/tree/main, 2024.

[191] M. Mekki, N. Toumi, and A. Ksentini, *Benchmarking on microservices configurations and the impact on the performance in cloud native environments*, version 1.0, Aug. 2022. DOI: 10.5281/zenodo.6907619. [Online]. Available: https://doi.org/10.5281/zenodo.6907619.