D5.3: Preliminary report on AI-data driven networking and QoS management

Revision: v.1.0

Work package	WP 5
Task	Task 5.2 and Task 5.4
Due date	30/06/2024
Submission date	30/06/2024
Deliverable lead	Marius Corici (FHG)
Version	1.0.F
Authors	Daniele Tarchi, Swapnil Sadashiv Shinde, David Naseh (CNIT) Marius Corici, Hemant Zope, Hauke Buhr (FHG), Roshith Sebastian (DLR), Eddi González (HSP)
Reviewers	Justin Tallon (SRS)
Abstract	This deliverable presents the preliminary advancements of Task 5.2 and Task 5.4 in developing an AI framework and an end-to-end QoS mechanism for the 5G NTN.
Keywords	5G NTN, AI, end-2-end QoS

Co-funded by the European Union 6GSNS
Grant Agreement No.: 101096573 Topic: HORIZON-JU-SNS-2022-STREAM-A-01-02
Call: HORIZON-JU-SNS-2022 Type of action: HORIZON-JU-RIA

Document Revision History

Version	Date	Description of change	List of contributor(s)
V0.1	29/01/2024	1st edit, draft ToC	Marius Corici
V0.2	19/02/2024	ToC Update	Marius Corici, Daniele Tarchi
V0.3	05/04/2024	AI Data Driven Networking content definition	Daniele Tarchi
V0.4	29/05/2024	AI Data Driven Networking 1st Draft	Daniele Tarchi, Swapnil Sadashiv Shinde, David Naseh
V0.5	30/05/2024	End-to-end QoS First draft	Marius Corici, Hauke Buhr
V0.6	03/06/2024	First version of the core network split	Marius Corici, Hemant Zope
V0.7	14/06/2024	Completed the text of the deliverable	Daniele Tarchi, Marius Corici
V0.8	17/06/2024	Editorial changes	Marius Corici
V0.9	21/06/2024	Internal review	Justin Tallon
V0.10	24/06/2024	Responding to internal comments	Marius Corici, Daniele Tarchi
V0.11	26/06/2024	Final Editorial Adaptation	Marius Corici
V1.0.F	01/07/2024	Final review for approval and EC portal submission	Tomaso de Cola

DISCLAIMER

Co-funded by the European Union 6GSNS

5G-STARDUST (*Satellite and Terrestrial Access for Distributed, Ubiquitous, and Smart Telecommunications*) project has received funding from the Smart Networks and Services Joint Undertaking (SNS JU) under the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101096573.

This work has received funding from the Swiss State Secretariat for Education, Research and Innovation (SERI).

COPYRIGHT NOTICE

Project co-funded by the European Commission in the Horizon Europe Programme		
Nature of the deliverable:	**R**	
Dissemination Level		
PU	*Public, fully open, e.g. web (Deliverables flagged as public will be automatically published in CORDIS project's page)*	✓
SEN	*Sensitive, limited under the conditions of the Grant Agreement*	
Classified R-UE/ EU-R	*EU RESTRICTED under the Commission Decision No2015/ 444*	
Classified C-UE/ EU-C	*EU CONFIDENTIAL under the Commission Decision No2015/ 444*	
Classified S-UE/ EU-S	*EU SECRET under the Commission Decision No2015/ 444*	

* R: Document, report (excluding the periodic and final reports)

DEM: Demonstrator, pilot, prototype, plan designs

DEC: Websites, patents filing, press & media actions, videos, etc.

DATA: Data sets, microdata, etc.

DMP: Data management plan

ETHICS: Deliverables related to ethics issues.

SECURITY: Deliverables related to security issues

OTHER: Software, technical diagram, algorithms, models, etc.

EXECUTIVE SUMMARY

This deliverable, "Preliminary report on AI-data driven networking and QoS management," focuses on leveraging AI-driven networking concepts to enhance the overall network architecture for effective and flexible Quality of Service (QoS) management for 5G Non-Terrestrial Networks (NTN). The 5G-Stardust project aims to integrate satellite and terrestrial systems to provide global connectivity, especially in underserved areas. Within this context, this deliverable provides the following major technology advancements:

1) AI Data Driven Networking: The report discusses the integration of AI in network management, emphasizing the architectural framework of Open Radio Access Network (O-RAN) to ensure flexible and scalable network operations. AI techniques, including machine learning (ML) and reinforcement learning (RL), are proposed to optimize resource allocation, network slicing, and multi-connectivity solutions.

2) End-to-End AI-oriented Network Architecture: Two architectural schemes are proposed: direct connectivity and indirect connectivity, both designed for leveraging AI for dynamic resource management and seamless interoperability.

3) Controller Framework: A critical aspect of the network architecture is the controller, which manages network elements through traditional optimization methods and AI-based solutions. The controller is designed to handle various tasks such as RAN function deployment, Virtual Network Function (VNF) placement, and network slicing, ensuring optimal performance and resource utilization.

4) QoS Management: The report introduces a dynamic end-to-end QoS assurance mechanism tailored for mega-constellations, adapting routing at the mega-constellation level and implementing a cross-layer interface between the 5G service layer and the routing layer. This mechanism aims to maintain consistent QoS despite the variable conditions inherent in NTNs.

Being a preliminary deliverable, these activities concentrated on the development of framework elements necessary to sustain the optimization algorithms and not on punctual algorithms, providing a model how the different AI and QoS assurance algorithms can be later included.

TABLE OF CONTENTS

LIST OF FIGURES

ABBREVIATIONS

3GPP	3rd Generation Partnership Project
5G	5th Generation of Mobile Communications
AI	Artificial Intelligence
AMF	Access and Mobility Function
APN	Access Point Name
ATSSS	Access Traffic Steering, Switching and Splitting
B5G-XISP	Beyond 5G cross-industry service platform
BBU	Baseband Unit
CCP	Core Control Function
CHO	Conditional Handover
CN	Core Network
CP	Control Plane
CU	Central Unit
CUP	Core User Plane
DU	Distributed Unit
DVB	Digital Video Broadcasting
E2E	End-to-End
ECN	Explicit Congestion Notification
ENI	Experiential Networked Intelligence
ETSI	European Telecommunications Standards Institute
FFT	Fast Fourier Transform
FL	Federated Learning
HAPS	High Altitude Platform Systems
IAB	Integrated Access Backhaul
IETF	Internet Engineering Task Force
ISL	Inter-Satellite Link

ITU-T	International Telecommunication Union - Telecommunication Standardization
MAC	Medium Access Control
MEC	Mobile Edge Computing
mIoT	Massive Internet of Things
ML	Machine Learning
N3IWF	Non-3GPP Interworking Function
NF	Network Function
Near-RT RIC	near-Real Time RAN Intelligent Controller
NN	Neural Network
Non-RT RIC	non-Real Time RAN Intelligent Controller
NRF	Network Repository Function
NS	Network Slice
NTN	Non-Terrestrial Networks
NWDAF	Network Data Analytics Function
O-RAN	Open RAN
ONAP	Open Network Automation Platform
PCF	Policy Control Function
PDCP	Packet Data Convergence Protocol
QoS	Quality of Service
RAN	Radio Access Network
RIC	RAN Intelligent Controller
RL	Reinforcement Learning
RLC	Radio Link Control
RRC	Radio Resource Control
RU	Radio Unit
SDAP	Service Data Adaptation Protocol
SDN	Software Defined Networks
SDO	Standards Developing Organizations

SL	Supervised Learning
SLA	Service Layer Agreement
SMF	Session Management Function
SMO	Service Management and Orchestration
SON	Self-Organizing Networks
TL	Transfer Learning
TN	Terrestrial Network
TS	Tempo-Spatial
UAV	Unmanned Aerial Vehicle
UE	User Equipment
UL	Unsupervised Learning
UP	User Plane
UPF	User Plane Function
VNF	Virtual Network Function
ZSM	Zero-touch Service Management

1 INTRODUCTION

The deliverable titled "Preliminary report on AI-data driven networking and QoS management" is centered on the progress and innovations achieved in Work Package 5 (WP5) of the 5G-STARDUST project. WP5 focuses on integrating AI-driven networking solutions and comprehensive Quality of Service (QoS) management to enhance the performance and reliability of 5G networks in the context of Non-Terrestrial Networks (NTNs).

This deliverable is based on the architecture developments from D3.2 and specifically addresses the advancements in Task 5.2 and Task 5.4, which are pivotal in achieving the project's goals:

Task 5.2: AI Data Driven Network Management and Slicing

This task aims to develop optimizations for network management using AI insights. The high predictability and flexibility of NTNs, when combined with static Terrestrial Networks (TNs), promise to deliver superior service to subscribers through optimized network configurations.

To achieve this, a comprehensive end-to-end AI processing system is proposed in this deliverable, balancing the data acquisition, AI decisions and their dissemination needs with resource constraints.

Key objective of this deliverable was the development of this framework across the RAN, core network, virtual network functions deployments which may be composed in independent end-to-end slices setting up the scene for the D5.5 where exemplary optimizations are described. The main reason for this framework choice is that AI has already proved that it can provide specific punctual optimizations in the system, while at the current moment we are stringently requiring a framework which can include and run multiple AI decisions in parallel.

Task 5.4: End-to-End QoS Management

This task focuses on proposing a seamless resource reservation solution across the integrated NTN/TN system. The goal is to extend existing 5G QoS profiles to accommodate the unique capabilities of NTNs and enforce these profiles within both NTN and TN RAN as well as the 5G Core Network.

Special emphasis is placed on creating end-to-end QoS capabilities within the space segment that guarantee high reliability, utilizing multi-connectivity solutions and advanced network management strategies to enhance end-to-end communication reliability.

Please note that this deliverable represents only a first presentation of the activities in this work package. These activities will be continued and reported in D5.4 at a later date.

2 AI DATA DRIVEN NETWORKING

The purpose of this section is to define a proper framework for the networking management of the 5G-STARDUST architecture. To this aim, by resorting to the architectural framework defined in the D3.2 we progress the status of the networking element definition by integrating them in a networking architecture. We resorted to the O-RAN approach where proper elements have been defined by extending them to the case under study.

This section is organized in such a way. In Section 2.1, the End-to-End (E2E) Network Architecture is discussed; by leveraging from the Architectural framework proposed in D3.2, we focus here on the Networking aspects by showing the impact of the different elements in both direct and indirect connectivity scenarios. The O-RAN framework is considered as the reference one. In Section 2.2, the input to the controller is discussed, by considering both Non-Real time and Near Real-time processes. In Section 2.3, we focus on the network elements the Controller should have taken into account. In Section 2.4 instead the controller framework is presented where different possible options are discussed, also considering the impact of the ML blocks.

2.1 THE E2E NETWORK ARCHITECTURE

Non-Terrestrial Networks (NTN) are communication networks that utilize non-terrestrial platforms such as satellites, high-altitude platforms, or unmanned aerial vehicles. These networks are designed to provide global coverage, especially in areas where terrestrial networks are not feasible or cost-effective. An E2E architecture for NTN refers to a comprehensive approach that considers all aspects of the network, from the user equipment to the network core, and everything in between. This includes the radio access network, the transport network, the core network, and the application layer. The goal of E2E architecture is to ensure seamless interoperability and efficient resource management across the entire network.

In 5G-Stardust D3.2 the E2E architectures has been defined to meet project requirements. For this, the main elements that have been defined are:

- The focus of the system is to deploy an integrated TN/NTN system where the NTN part is considered as a self-organizing mega-constellation.
- Artificial Intelligence (AI) is part of the system aiming at coping with three main categories of events: varying conditions, exceptional situations, resource usage optimization. When any of the previous events occur, AI should be able to reorganize the system in order to cope with the changing situation.
    - AI can be executed on both the user and network side.
        - The user side is suggested once there is little data, so a global optimization is not viable.
        - Network side once there are big data available aiming at having a global optimization.
- The Controller whether based on AI, or through traditional mechanisms should be able to manage the following aspects.

- Multi-connectivity gains from the presence of multiple networks (TNs, NTNs), multiple connection types (IAB, direct), multiple transport layer configurations. AI- based control mechanisms are envisaged.

- Network slices may be configured. Network slices maps user requirements to RAN/CN configurations (policies-based allocation).

- RAN functions can be organized into several possible options. The most appropriate functional split should be selected. A companion idle satellite should be individualized to reorganize the split options.

- The core network functions may be relocated on different networking elements.

The Open Radio Access Network (O-RAN) is a new paradigm in network architecture that opens up traditionally closed RAN components, allowing for more flexibility and innovation. O-RAN promotes interoperability and competition by standardizing interfaces between different components of the RAN. This allows network operators to mix and match equipment from different vendors, leading to cost savings and improved network performance.

AI is being increasingly used in NTN to optimize network performance and improve user experience. AI can be used for predictive maintenance, anomaly detection, network optimization, and many other tasks. By analyzing network data, AI algorithms can make predictions and decisions in real time, leading to more efficient network management and better service quality.

In the context of NTN, leveraging AI and O-RAN in an E2E architecture can bring numerous benefits. AI can help optimize the allocation of network resources in real time, improving the efficiency and performance of the network. Meanwhile, O-RAN can provide the flexibility needed to adapt to the unique challenges of NTN, such as variable link conditions and limited resources.

Taking into account the previous elements, two architectural schemes have been designed with the aim of coping with the previous requirements. In addition to this, the O-RAN architectural framework has been considered to be integrated within the networking architecture. One architecture refers to the direct connectivity option where the UE is directly connected to the NTN. The second architecture is related to an indirect connectivity option, where the UE is connected to the NTN through a terrestrial IAB Node. The architecture is made up of several layers that identify the elements that make up the system.

2.1.1 O-RAN architectural framework

The Open Radio Access Network (O-RAN) specification aims to define open, interoperable RAN interfaces with virtualized and intelligent RAN functions. The open and interoperable RAN interfaces can enable RAN functions developed by multiple entities to work together, avoiding the possibility of vendor-locked systems. Through virtualization techniques, RAN functions can be distributed across several layers of networking infrastructures to enable a flexible access network.

RAN data collected from different RAN functions can be used to further optimize RAN operations by inducing ML solutions. In particular, the O-RAN architecture introduces the distributed control mechanism enabled through non-Real Time RAN Intelligent Controller (non-RT RIC) and near-Real Time RAN Intelligent Controller (near-RT RIC) that support the intelligent and optimized RAN control mechanisms over different time scales. Here, we summarize the logical O-RAN architecture, its basic elements, and various interfaces

presented in [1]. Figure 2-1 provides a logical description of various O-RAN elements and corresponding interfaces. In the following we describe these elements and interfaces in details.

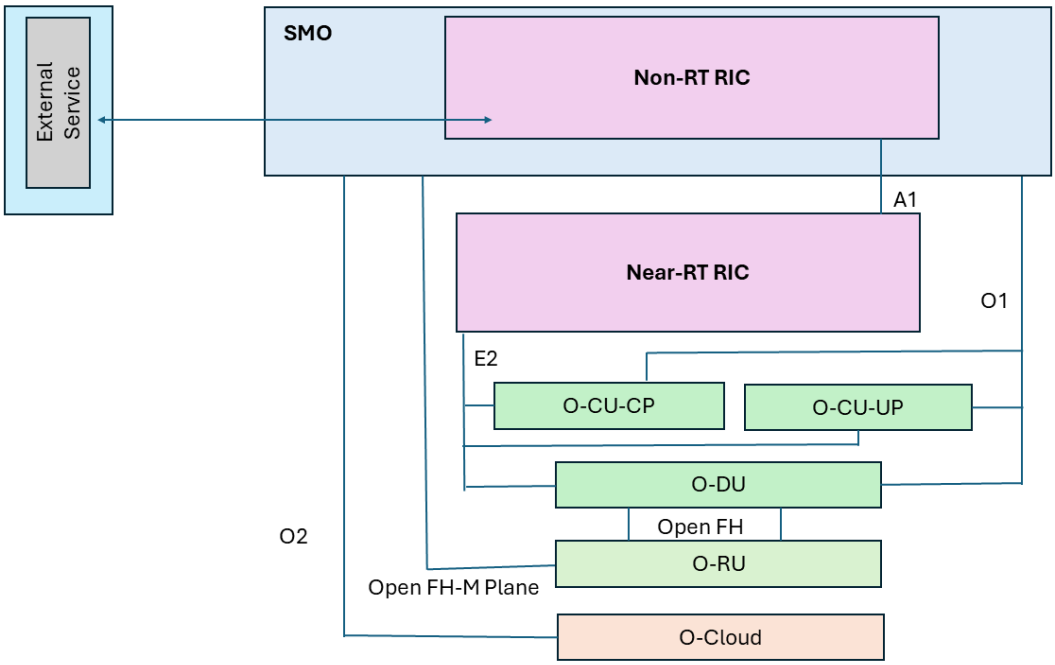

Figure 2-1: O-RAN Architecture

Service Management and Orchestration (SMO)

The Service Management and Orchestration (SMO) block includes various functionalities for managing the O-RAN functions and infrastructure resources. The interface O2 connects the SMO with O-RAN Cloud (O-Cloud), which is a cloud computing platform comprising a collection of physical infrastructure nodes that meet O-RAN requirements to host the relevant O-RAN Network Functions (NF). It provides platform resources and workload management services. SMO also includes the Non-Real-Time RAN Intelligent Controller (Non-RT RIC), which supports intelligent RAN operation and optimization.

Through the O1 interface, SMO can connect with E2 nodes (O-CU-UP, O-CU-CP and O-DU) to support these NFs with various services, including performance management, configuration management, fault management, file management, communications surveillance, etc. Additionally, SMO blocks can also access external services and datasets through external interfaces and provide various functionalities to the non-RT RIC through anchored/non-anchored functions defined in it.

Thus, SMO covers a wide range of functionalities including the functions specific to the RAN domain, some generic functions to integrate RAN, and other domains such as cloud infrastructure and other networks.

Non-RT RIC

In O-RAN architecture non-RT RIC is an internal functionality element of SMO block. Thus, non-RT RIC represents the subset of functionalities provided by the SMO block in particular by providing support to the intelligent RAN operation and optimization. It also provides policy-based guidance and enrichment information to Near-RT RICs over the A1 interface (the AI/ML

model on the A1 interface is not yet discussed in the O-RAN specifications). It supports the RAN optimization services through control loops with intervals greater than 1 s.

It is composed of a Non-RT RIC Framework and Non-RT RIC Applications (rApps). Non-RT RIC Framework terminates the A1 interface to Near-RT RIC Framework and provides R1 services to rApps. Various rApps can provide value-added services related to RAN operation and optimization including but not limited to radio resource management, data analytics, and providing enrichment information. The non-RT RIC framework should support the AI/ML workflow-related services including ML model training, storage and retrieval of trained ML models, real-time monitoring of the deployed ML model performance, retrieval of trained external models and corresponding metadata from external AI/ML service providers, etc.

The SMO and the Non-RT RIC framework can provide several logical functions, including the functions anchored inside the non-RT RIC framework, functions anchored outside the non-RT RIC framework, and the non-anchored functions. The AI/ML workflow function can be a part of a non-anchored function group.

Near-RT RIC

Near-RT RIC includes a set of xApps (both third-party and RIC vendor enabled) and the generic platform functions demanded by the specific functions implemented through xApps. It also provides an RAN database by collecting information on network states, E2 nodes, cells, UEs, etc. It can provide support for AI / ML application development (through xApps) by hosting data pipeline, model management, ML training, and inference processes through AI/ML support functionality. xApps can use none, part of, or all AI/ML support functionalities based on their designs and requirements. In general, Near-RT RIC can connect to a single non-RT RIC and one or more E2 nodes.

O-RAN Functional Split and RAN Functions

Among several possible split option defined by 3GPP, O-RAN embraces and extends the 3GPP NR 7.2 split for base stations, which disaggregates the base station functionalities into a Central Unit (CU), a Distributed Unit (DU), and a Radio Unit (RU). Split 7.2 allows a relatively simple RU design with proper data rates and latency required on the interface between the RU and the DU. From now on we will consider split option 7.2 as the only one to be deployed.
The O-RAN based functional blocks are called **O-RAN Distributed Unit (O-DU)**, **O-RAN Central Unit (O-CU)** and **O-RAN Radio Unit (O-RU)** in the O-RAN specification documentation. The O-CU is further split into two logical components associated with the Control Plane (CP) and the User Plane (UP). These are called **O-RAN Central Unit –Control Plane (O-CU-CP) and O-RAN Central Unit -User Plane (O-CU-UP)**. Split 7.2x, supports the simplified and inexpensive RU designs where only FFT and cyclic prefix addition/removal operations are performed. The DU includes the remaining physical layer functionalities along with Medium Access Control (MAC) and Radio Link Control (RLC) layers. The CU functions implement the Radio Resource Control (RRC) layer, the Service Data Adaptation Protocol (SDAP) layer, and the Packet Data Convergence Protocol (PDCP) layer from the protocol stack. The functional split allows for the virtualization and flexible deployment of RAN functions over different networking facilities.

2.1.2 Proposed E2E Reference NTN Architectures

Two reference architecture frameworks are considered. A direct connectivity architecture is envisaged to have direct connectivity with the NTN elements, where eventually a terrestrial gateway is present. Instead, an indirect connectivity architecture is considered to exploit an integrated access and backhaul (IAB) node, acting as an intermediary link.

2.1.2.1 Direct Connectivity Architecture

In the Direct Connectivity architecture, Figure 2-2, the User Equipment (UE) can be directly linked to the Non-Terrestrial Network (NTN) without the need for a Terrestrial Network (TN) node as an intermediary. This architecture takes advantage of advancements in 5G new radio (NR) technology to establish efficient communication channels over satellite links, potentially replacing traditional technologies like DVB for various communication links.

Components and Connections:

- **User Equipment (UE):** Represents the devices directly connecting to the NTN.
- **NTN Access Network:** Provides the direct link between the UEs and the NTN.
- **Near-RT RIC:** This device is responsible for controlling the radio access network (RAN) in near-real time.
- **Near-RT Ground RIC:** This is a device that is responsible for controlling the RAN in near-real time at the ground level.
- **Near-RT NTN RIC:** This device is responsible for controlling the RAN in near real time at the NTN level.
- **Terrestrial Gateway:** This is a device that connects the terrestrial network to the NTN network.
- **Serving Satellite:** This is the satellite to which the UE is connected.
- **Companion Satellite(s):** These are satellites that are working together with the serving satellite to provide coverage and capacity to the UE.
- **Ground Station:** This is a station on the ground that communicates with the satellites.
- **5G Core Network:** Facilitates the interconnection between the NTN Access Network and the core network through the Non-3GPP Interworking Function (N3IWF).

Advantages:

- **Seamless Roaming:** Enables seamless authentication and roaming for users between terrestrial cellular networks and NTN access without requiring additional equipment.
- **Cost Efficiency:** Reduce barriers to entry and satellite constellation systems by promoting common platforms for service management and coherent orchestration of resource, mobility, and forwarding path functions.

Challenges:

- **Interoperability:** Ensuring seamless integration and interoperability between different vendor systems and satellite providers.
- **Dynamic Network Management:** Managing dynamic and wide-ranging types of networks with satellite and/or High-Altitude Platform Systems (HAPS) elements efficiently.

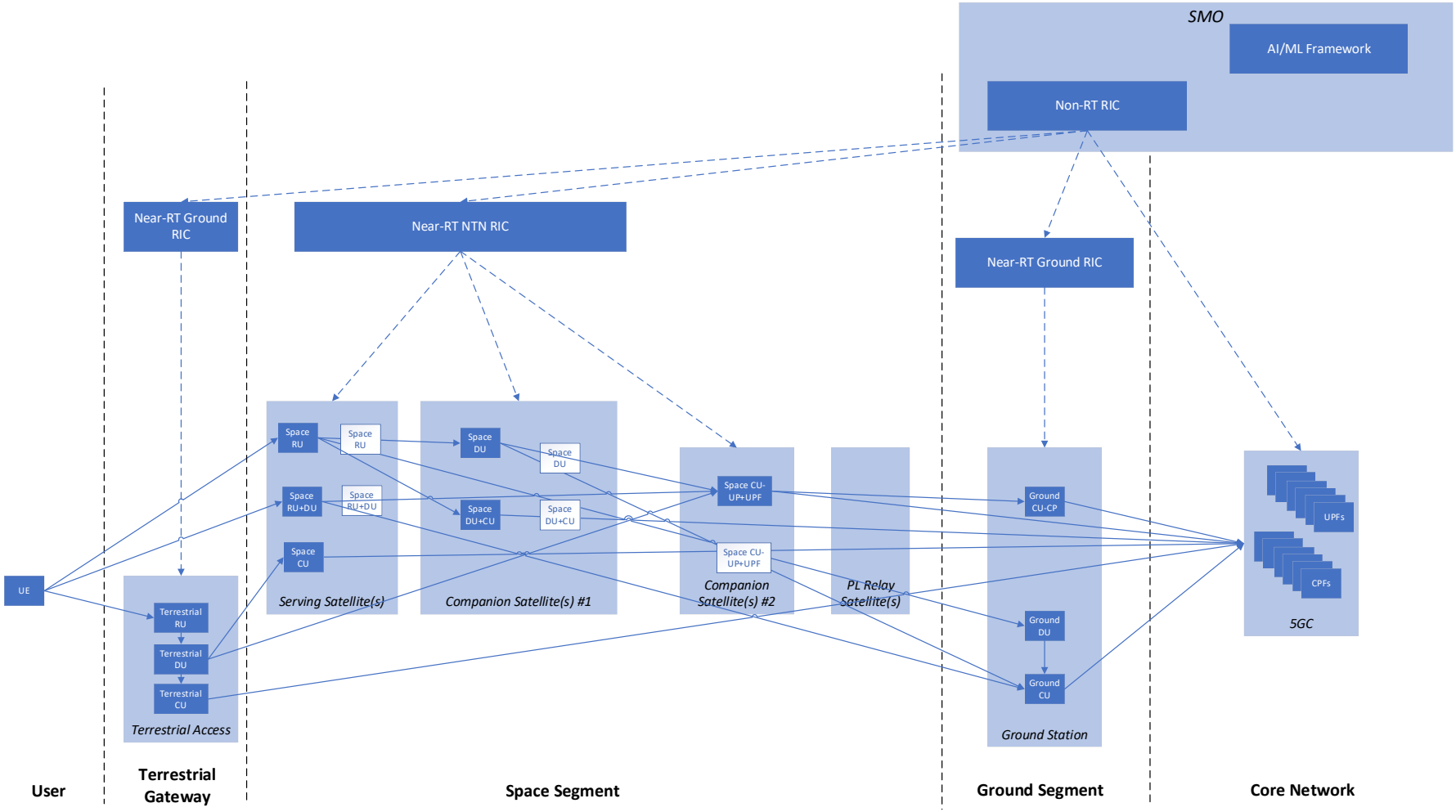

Figure 2-2 Direct Connectivity

2.1.2.2 Indirect Connectivity Architecture

In the Indirect Connectivity architecture, Figure 2-3, the UE connects to the NTN through a Terrestrial Integrated Access and Backhaul (IAB) Node, acting as an intermediary link, supposed to be composed by a RU+CU+DU terminal and a Mobile Terminal (MT). This architecture capitalizes on the capabilities of 5G IAB to establish a mesh backhaul network, managed and orchestrated using similar techniques as the direct connectivity architecture.

Components and Connections:

- **UE:** Represents devices connected to the NTN through the Terrestrial IAB Node.
- **Terrestrial IAB Node:** Acts as a backhaul node for the NTN Access Network, providing connectivity to remotely deployed gNodeB units.
- **NTN Access Network:** Offers connectivity to the Terrestrial IAB Node, serving as backhaul for gNB units, and interconnects with the 5G Core Network through the N3IWF.
- **Serving Satellite:** This is the satellite to which the UE is connected.
- **Companion Satellite(s):** These are satellites that are working together with the serving satellite to provide coverage and capacity to the UE.
- **Ground Station:** This is a station on the ground that communicates with the satellites.

Advantages:

- **Unified Management:** Enables the unification of RAN and backhaul mesh networks, allowing Service Management and Orchestration (SMO) capabilities for 5G NTN to control and manage both networks efficiently.
- **Optimized satellite backhaul:** Maps 5G bearers to different backhaul channels, utilizing the SMO's knowledge of available carriers and their capabilities to enhance efficiency.

Challenges

- **Complex Network Integration:** Managing the integration of various network components and ensuring seamless operation.
- **Resource Optimization:** Balancing resources and optimizing network performance in a dynamic environment.

These architectures present distinct advantages in terms of connectivity and efficiency, yet they also pose challenges related to interoperability, dynamic network management, and resource optimization in the context of Non-Terrestrial Networks.

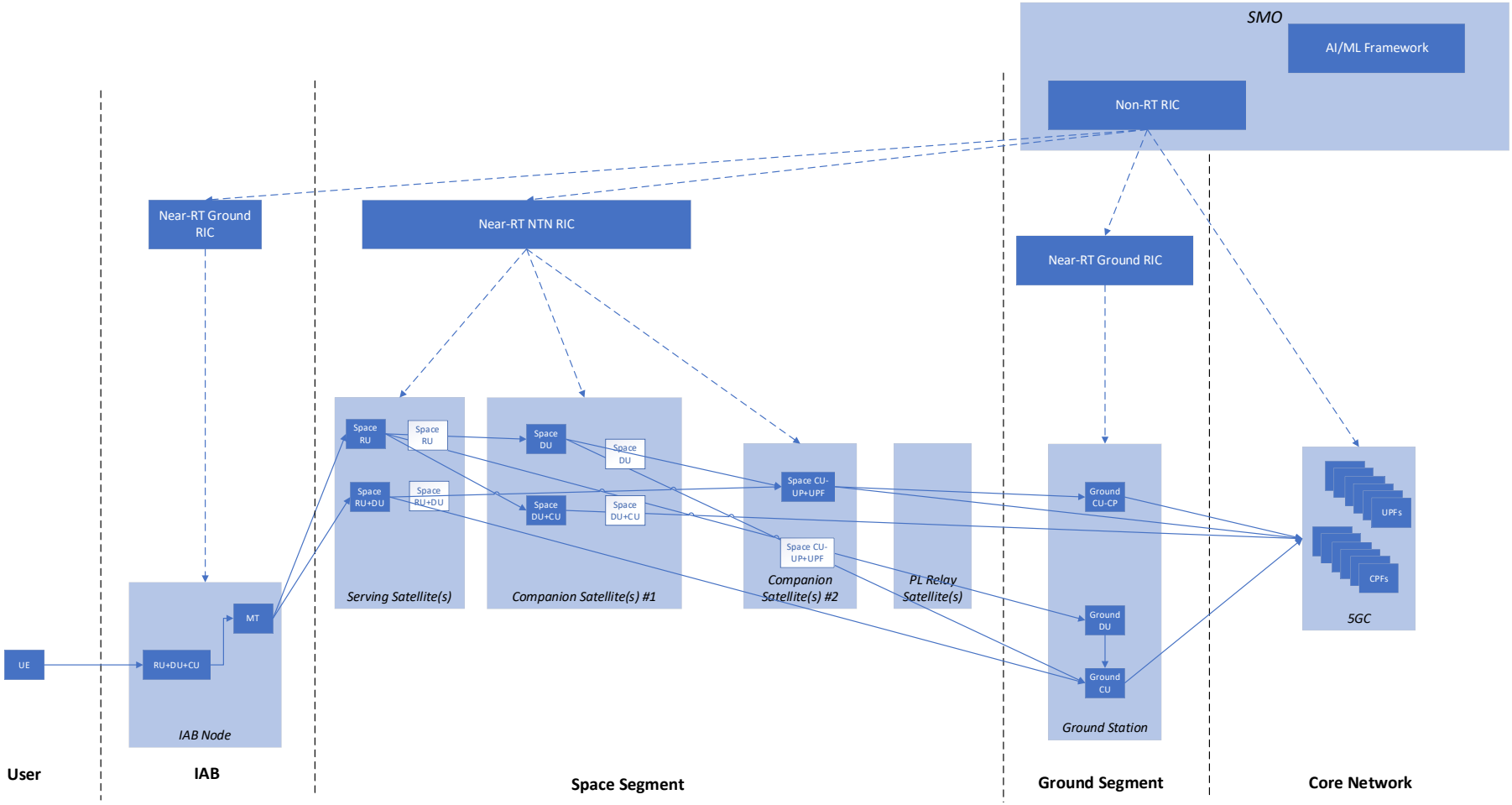

Figure 2-3 Indirect Connectivity

2.1.3 Goals of the Controller

In this subsection we will put emphasis on the Controller and its role in the Network Architecture, focusing on the fact that it is the network element on which we will mainly focus our attention as controller of the networking elements.

In the following paragraphs we introduce the literature in the area. The 3GPP standardization group has mandated that 5G systems support services through satellites, allowing the integration of satellite assets into the 5G framework. This integration improves terrestrial infrastructure by providing global connectivity, extending coverage to underserved areas, and improving reliability for mission critical and disaster scenarios. However, managing a 3D-integrated TN-NTN network architecture is complex due to the variety of architecture options, the need for coordination, and evolving standards.

NTNs introduce additional layers of complexity, such as varying propagation characteristics and latency issues. Effective management requires sophisticated orchestration mechanisms for dynamic resource allocation, optimized routing, and seamless handovers between terrestrial and non-terrestrial components. This involves developing intelligent algorithms and protocols to balance traffic, mitigate interference, and adapt to real-time network conditions. Ensuring interoperability and standardization of interfaces and protocols is also crucial to smooth integration and efficient data exchange between network segments. Security and privacy are significant concerns in the integration of NTNs with 5G, necessitating robust encryption and secure authentication protocols to protect sensitive data. Additionally, regulatory and spectrum management issues must be addressed to ensure efficient utilization of the frequency band and minimize interference. Consequently, integrating NTNs with 5G offers significant benefits, but also poses challenges that must be addressed through advanced orchestration and management. The evolution of orchestration in telecommunications has seen significant advancements. Since 2008, 3GPP has introduced AI concepts in mobile networks, such as Self Organizing Networks (SON), to automate specific control and management tasks [1]. By 2017, 3GPP introduced domain-specific data analytics functions, improving network control and management. The first release of 5G-Advanced, 3GPP Release 18, includes various AI-focused study items [3].

The O-RAN Alliance is notable for its efforts to disaggregate the traditional RAN architecture, promoting multivendor interoperability, intelligence, and programmability. The RAN Intelligent Controller (RIC) is a key component of this architecture, optimizing RAN functions. Two significant ETSI standardization initiatives are ETSI Experiential Networked Intelligence (ENI) [4] and ETSI Zero-touch Service Management (ZSM) [5]. ETSI ENI provides a centralized AI-based network management framework, while ETSI ZSM offers a reference architecture for distributed management and orchestration. The Open Network Automation Platform (ONAP), an open-source project by the Linux Foundation, aims to automate network service management, supporting distributed and federated learning and closed-loop automation. From 3GPP Release 17, the NT environment has been considered, initially treated as separate from the conventional 5G network. Progress in the O-RAN Alliance includes extending interfaces to support 5G/6G NTN capabilities, though maturity levels vary. Emerging approaches for the integration of TN-NTN are being explored in the scientific community. The ITU-T Study Group 13 has proposed a reference architecture for the TN-NTN interaction, but a standardized approach for convergence and interoperability is still needed.

A foundational work discusses the approach to unified service management and orchestration (SMO) for NTNs, using open RAN technology and software-defined networking (SDN) [6]. This paper describes the implementation of Temporospatial SDN (TS-SDN) controller software, enhanced to harmonize with O-RAN Alliance specifications. This unified SMO framework supports 5G NTNs, DVB, and other satellite networks, with the goal of reducing costs for users

and operators, facilitating seamless authentication, and enabling dynamic reconfiguration of network resources. TS-SDN efficiently manages dynamic and wide-ranging NTNs by integrating satellite and aerial nodes into the radio access network (RAN), crucial for seamless integration into the 5G and beyond ecosystem.

An emulation framework for managing NTNs and satellites within 5G networks, using the Open-Air Interface platform to reflect OpenRAN's disaggregated network architecture, is another significant development [7]. This framework emphasizes enhancing the lifecycle management of virtual network functions (VNF), particularly through Kubernetes integration for resource sharing among VNFs. The primary goals include optimizing VNF deployment and resource allocation, ensuring scalability and efficiency in hybrid networks. The effectiveness of the framework in optimizing resource utilization and enabling new use cases like MEC and massive IoT (mIoT) is validated through performance metrics.

AI techniques for optimizing resource allocation and utilization in satellite-based communication systems are explored in another study [8]. The orchestrator in this framework comprises data collection, analysis, and data-driven decision-making using AI and ML techniques such as DL, RL, and FL. The orchestrator aims to improve communication efficiency, reliability, and network management by learning optimal resource allocation strategies through Q-learning. Practical applications include improving IoT systems, mobile services, and overall network management efficiency.

The O-RAN architecture, which uses open interfaces and standardized protocols to enable interoperability and flexibility of multiple vendors, also plays a crucial role [9]. Central to this architecture is the RAN Intelligent Controller (RIC), divided into near-real-time and non-real-time components. The near-RT RIC uses AI/ML algorithms for real-time network optimization, while the non-RT RIC handles longer-term management and guidance. The O-Cloud supports these functions by providing a scalable cloud computing environment, aiming to optimize RAN functionalities and ensure efficient resource management and network performance.

A 5G smart connectivity platform integrates dynamic control and orchestration of NTNs, including satellites and UAVs, with IoT and cellular 5G networks [10]. This platform uses real-time monitoring and dynamic resource allocation to meet different QoS requirements and improve energy efficiency. The slicing of the network and integration with 5G technologies improve connectivity, particularly in remote areas. The platform's automation capabilities aim to manage the increasing number of connected devices and diverse business verticals autonomously. The Beyond 5G cross-industry service platform (B5G-XISP) optimizes service environments by orchestrating systems from different industries [11]. A cross-industry orchestrator manages, and coordinates subsystems based on service requirements, ensuring seamless integration and scalability. The orchestrator selects and configures subsystems to create optimized service environments, facilitating efficient resource allocation and service delivery across various industries. Efficient management and orchestration are critical to integrating satellite components into 5G networks, as highlighted by standardization efforts [12]. These efforts emphasize the increased complexity of integrating satellite access networks, necessitating efficient orchestration to optimize resources. The study presents reference architectures and requirements for managing integrated satellite components, focusing on network slice management, monitoring, and optimization. The goal is to streamline processes and enhance the overall performance of 5G networks with satellite integration.

In conclusion, the literature underscores the critical role of advanced controllers and orchestrators in managing and optimizing non-terrestrial networks. The primary goals include seamless integration of satellite and terrestrial networks, efficient resource allocation, and enhanced network performance through AI and ML techniques. Achieving these objectives is based on unified service management frameworks, emulation platforms, and rigorous

standardization efforts. These elements are critical for the successful deployment and operation of 5G and 6G NTNs, paving the way for improved connectivity and performance in next-generation networks.

2.2 INPUT OF THE CONTROLLER

In O-RAN-based access networks, both SMO/non-RT RIC and near-RT RIC can collect and process data to generate ML datasets. The data can be collected over various interfaces and used according to the ML application's demands.

2.2.1 Input Data to Non-RT RIC

The non-RT RIC can collect data from various RAN elements, O-cloud infrastructure, and external sources through the following mechanisms:

- **Data Collected on the O2 Interface through SMO**: The O2 interface connects the SMO facility to the O-Cloud, allowing it to manage the infrastructure resources and the NF deployment-related services. The O2 interface can be explored to collect infrastructure and NF placement-related data to allow the proper allocation of resources and NF placements.

- **Data Collected on O1**: The O1 interface connects the SMO platform with Near-RT RIC, O-DU, and O-CU NFs. Therefore, it can be explored to collect data related to the NFs' performance, configuration, fault management, file management, communications surveillance, etc. Through the O1 interface, SMO can also collect feedback from near-RT RIC and E2 nodes regarding the performance of control policies implemented over them.

- **Data Collected over FH-M Interface**: O-RU performance-related data can be collected through the FH-M interface.

- **Data Collected over the External Interface through SMO**: External enrichment Information from application servers and other policy-related data can be collected by SMO through an external service interface.

- **Internal Data Collection**: Additionally, various internal elements monitoring and feedback-related data from SMO/non-RT RIC i.e., rAPP, SMO, and Non-RT APP functionality Data components can be collected and utilized for ML model training.

2.2.2 Input Data to Near-RT RIC

The near-RT RIC can collect data from various RAN elements, as well as from non-RT RIC through the following mechanisms:

- **Data Collected on E2**: Near-RT RIC can explore the E2 interface for collecting data from O-DU and O-CU nodes including their performance in real time, resource consumption, policy feedback, etc.

- **Data Collected on A1**: Near-RT RIC can also use the A1 interface to receive useful information and policy-related data from non-RT RIC. It can also include the external enrichment data and the policies generated over SMO/non-RT RIC for improving the near-RT RICs performance.

2.2.3 Data Requirements

For optimizing the RAN/core network performance through proper function placement, network slice deployments, RAN functional split, and multi-connectivity related solutions, various RAN, other parts of networks, and external data sources can be integrated to form a RAN control policy for a specific task. In the following, we introduce the data requirements for each of these use cases and the type of interface that can be explored to collect such data.

a) **Network Slicing options data**: In the case of slice deployment options, the O1 interface can be used to collect the data from the O-Cloud network infrastructure including node capabilities and characteristics, location, speed, and resource availability. These data can be provided alongside the application requirements to non-RT RIC for policy developments, and these policies and application/slice requirements can be communicated over the A1 interface to near-RT RIC for further policy enhancements and policy deployments.

b) **RAN Function deployment Data**: In the case of RAN/core network function placements in multi-user scenarios, SMO can collect network traffic, geographical layout, existing infrastructure, cost considerations, regulatory data, technology metrics, and customer experience feedback through O1 and external interfaces. These data can be provided to non-RT RIC for developing the function placement policies over different infrastructure layers. Near-RT RIC can collect real-time user distribution, and resource consumption, i.e., bandwidth data from the RAN elements over the E2 interface. Additionally, it can also explore the A1 interface to collect the service requirements, i.e., latency demands, QoS demands, and other policies developed by non-RT RIC.

c) **VNF placement Data**: After selecting the infrastructure layer for placing the specific functions, it is important to fine-tune the performance by placing the specific VNFs that correspond to specific UEs according to the service demands and availability of multi-server processing units. In such cases, infrastructure data, UE demands, resource availability, and specific VNF demands can be considered. Such data can be collected through the O2 and O1 interface over SMO/non-RT RIC and used to form a VNF placement policy for multi-user multi-server cases.

d) **Multi-connectivity Data**: In the case of multi-connectivity-based solutions, the external data, in the form of past connectivity trends, network coverage maps, etc., can be collected by SMO. These data along with the connectivity requirements can be provided to non-RT RIC for generating the user connection policies. The E2 interface can be used to collect real-time information on real-time network performance data. The performance requirements and the policies generated by non-RT RIC can be provided to near-RT RIC for further enhancement and possible deployments.

2.3 NETWORK CONTROLLER ACTIONS

In this section, the network elements to be managed/controlled/optimized are discussed. The goal is to have an optimal placement of the VNFs, depending on their requirements, on the network configurations, and different slicing options. In Figure 2-4, a graphical representation of the controller functional scheme is given. The controller, eventually AI-based, works by managing four different elements.

The RAN controllers are expected to control and optimize the various parts of RAN including O-Cloud infrastructure resource management, RAN function deployments over heterogeneous infrastructures, VNF placements and reconfigurations over multi-server sites, slice-based service provisioning, handling the multi-connectivity and corresponding network selection process, etc. Based on the availability of distributed networking and processing infrastructures and service requirements, RAN controllers are expected to deploy RAN functions at multiple infrastructure sites to optimize RAN performance. Next, each infrastructure site can have multiple servers with restricted resources. Proper policies for the VNF placements based on the VNF's demands, service requirements, and server characteristics should be developed by RAN controllers. Additionally, for the case of multi-connectivity, where each UE can connect to multiple access nodes, a proper UE-RU assignment, i.e., network selection is required to satisfy the service demands. Such multi-connectivity-related policies should be enabled and optimized by RAN controllers based on UE demands and characteristics. Finally, by integrating these elements, proper slice oriented VNF chains should be formed for multiple UEs with diverse service requirements on a common networking infrastructure. Such network slicing policies should be developed by RAN controllers by monitoring the O-Cloud infrastructure and VNF deployment policies.

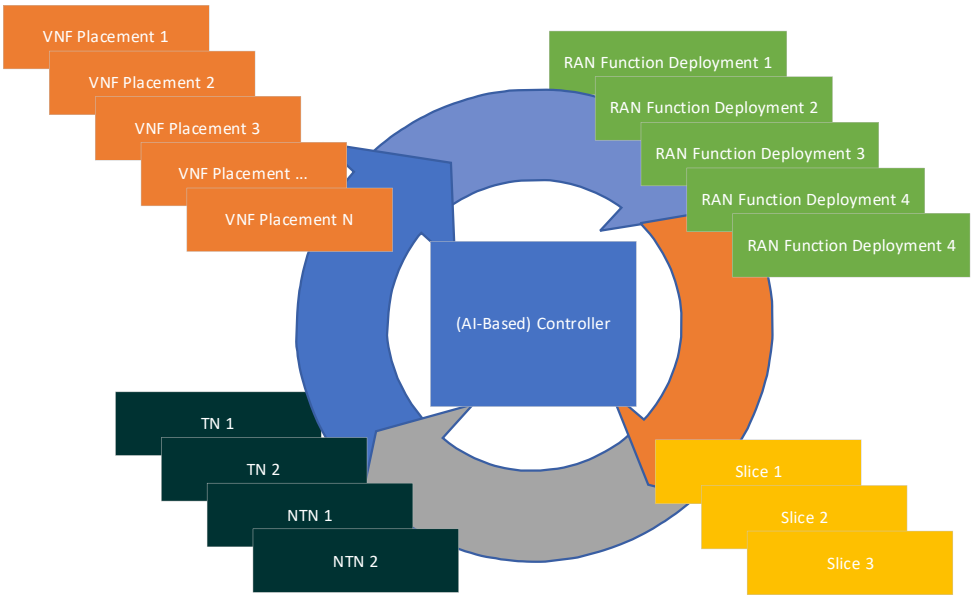

Figure 2-4 Controller Functional Scheme

2.3.1 RAN Function Deployment

Here, we introduce several possible options for the deployment of the RAN functions. In particular, several of these functions can be virtualized and deployed at various sites including terrestrial gateways, satellite access, satellite backhaul, terrestrial control stations, or in cloud facilities depending upon the service demands imposed by the users. The possible set of RAN functions including RU, DU, and CU correspond to the set of stack function functions of the O-RAN protocol. Additionally, core networking functions can be also deployed in some options where the Core User Plane (CUP), and the Core Control Function (CCP) are also considered managing the activities of the user and control plane.

While in this deliverable we focus on several different possible deployment options, in the following of the work we will shortlist the possible options to one, or two at the maximum by selecting the most promising.

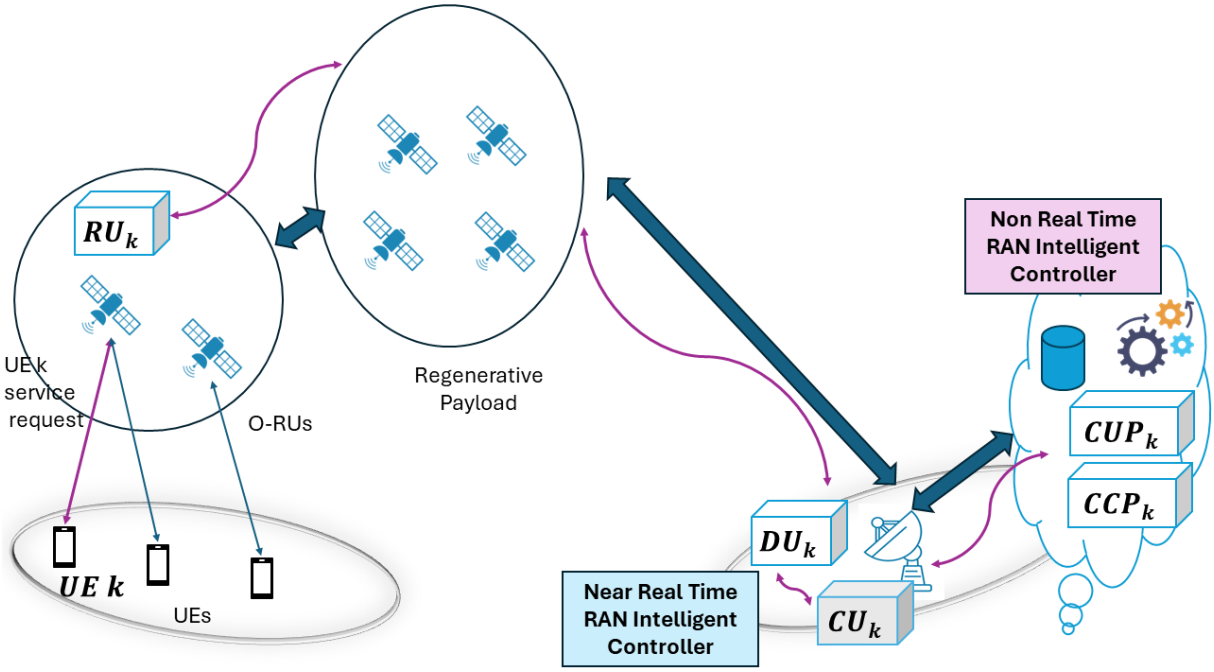

Figure 2-5: RAN Function Deployment- Option 1

Figure 2-5 presents the possible deployments of RAN/core functions over space-ground networks defined as option 1. Several UEs, from the service area, are connected to the satellite access and demand specific services. RU functions are deployed over the satellite access site, providing direct connectivity to the UEs under service. The user service data is then routed through the satellite transport network and delivered to the ground station site, where the DU and CU functionalities are implemented. Both sets of core functions CUP and CCP are implemented over the terrestrial cloud facilities with huge amount of resources.

Such deployments can have reduced costs and centralized control, allowing a single DU to handle a set of RUs and their data traffic. However, such solutions can suffer from non-real-time interface management and corresponding latency costs. Services demanding non-real-time latency and abundant data communication resources can take advantage of such O-RAN solutions where data polling and centralized management through non-real-time interfaces are possible.

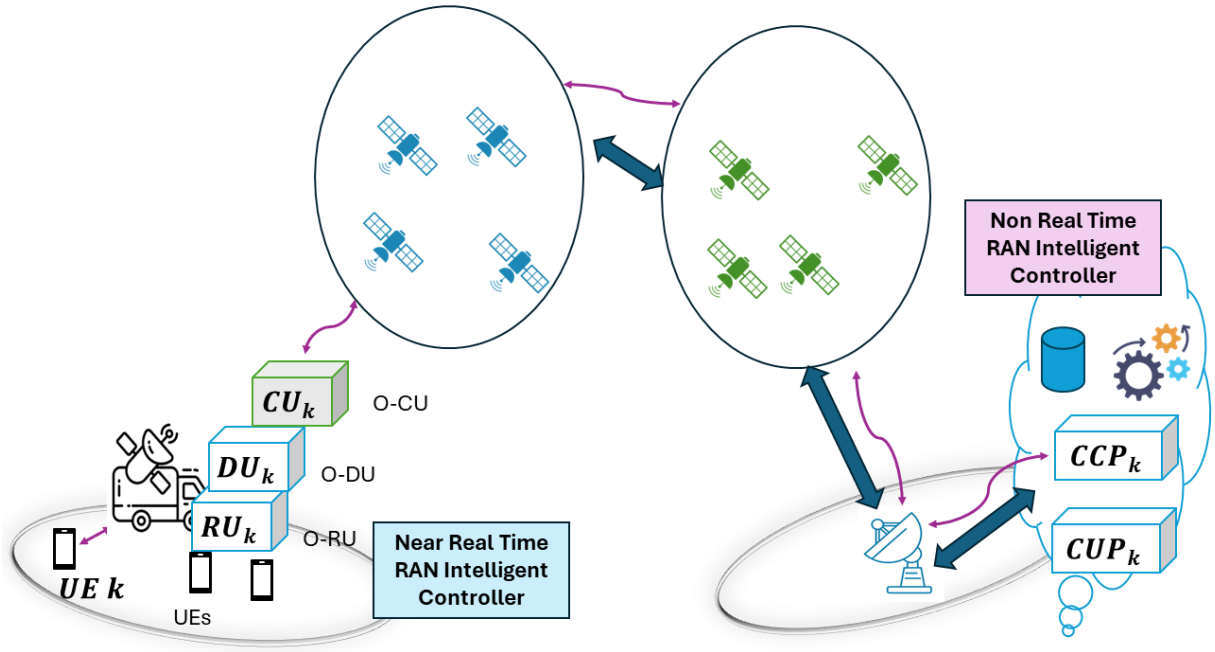

Figure 2-6: RAN Function Deployment- Option 2

Figure 2-6 shows the possible deployment option for the RAN function through the exploration of terrestrial gateway resources. This possibility is defined as option 2 where UEs are connected to terrestrial gateways that can route the service data towards distributed satellite nodes. In this case, the terrestrial O-RAN facility is enabled through the deployment of RU, DU, and CU functions on the gateways. Next, the UE service data are routed through the satellite network to the terrestrial cloud facilities having core networking functions onboard.

Such solutions can enable real-time access for terrestrial UEs with near-zero latency. However, centralized core networking facilities can induce higher latency while driving the backhaul data traffic. Such deployments fail to explore the space edge computing resources and explore the satellite networks only for relaying the user data to/from core networks. Such solutions can be adequate for services that require real-time data communication.

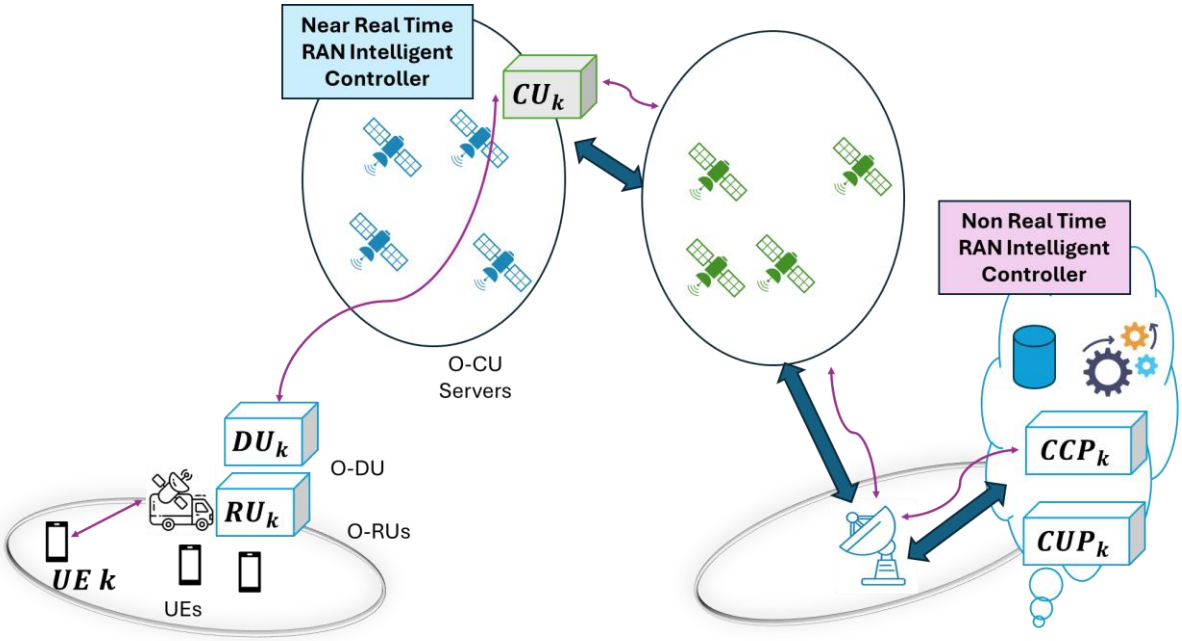

Figure 2-7: RAN Function Deployment - Option 3

Figure 2-7 shows the possibility of implementing the O-RAN functions over the space-ground network through the CU-DU function split. This possibility is defined as option 3, where the RU and DU functions are implemented on the terrestrial gateway site, while the CU functions are implemented in the space in the proximity of UEs. The data is then routed through satellite transport to terrestrial control stations, which then provide connectivity to the core networking facilities enabled at nearby cloud processing units.

This deployment option implements the O-RAN solutions through integrated space-ground resources. In particular, with RU-DU on terrestrial gateways, the baseband processing is performed in proximity of UEs while CU functions in space and a centralized coordination of UEs is possible. The backhaul latency to reach the core networks can be higher depending on the satellite transport network and the UE traffic demands.

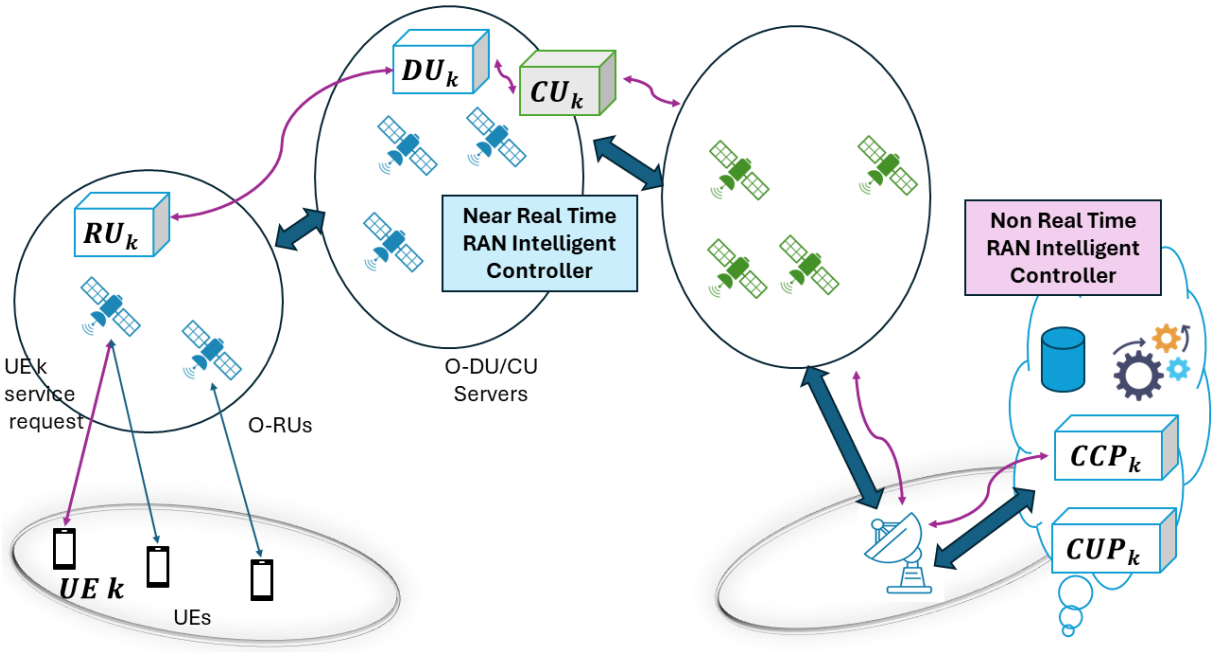

Figure 2-8: RAN Function Deployment- Option 4

Figure 2-8 shows the possibility of implementing the O-RAN architecture in space by distributing the RU, DU, and CU functions in space. In particular, UEs can access space network services by directly connecting to the satellite nodes in proximity that host RU functions. Next, the DU and CU functions are implemented on nearby satellite nodes to reduce the latency requirements for the O-RAN data processing functions. The service data are then routed through the satellite nodes towards the ground control station which can access the core networking functions through backhaul links.

In this case, the O-RAN functions are distributed over the space networks between multiple sites. With RU-DU split multiple RUs can be handled by DUs and through a non-real-time interface. Additionally, implementing CU-DU functions together can reduce the data processing capacity. However, the non-real-time interface and the longer satellite transport and backhaul delay can limit the applicability to real-time applications.

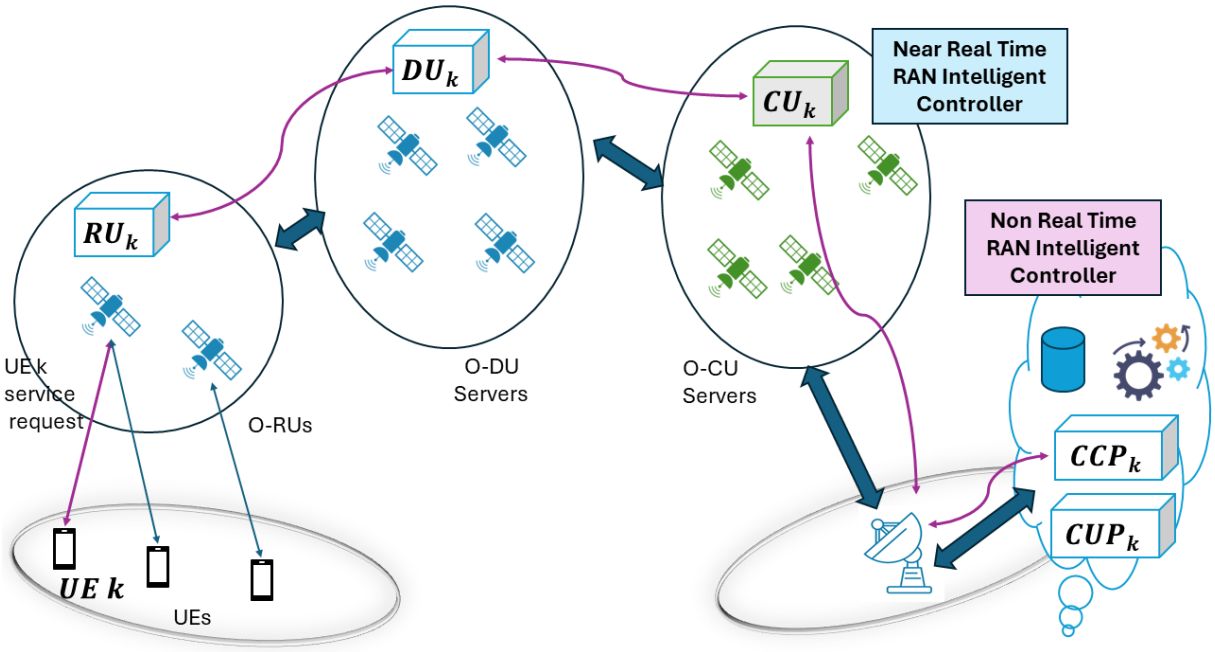

Figure 2-9: RAN Function Deployment- Option 5

The implementation of space-O-RAN in Figure 2-9 achieves the full distribution by allocating the RU, DU, and CU functions at different satellite sites. With this approach, distributed satellite resources can be properly used to enable efficient RAN architectures supporting multiple user services. Similar to the previous cases, the service data is then forwarded to the terrestrial control station with access to the core networking facilities.

In this case, the O-RAN solution is implemented through complete isolation of RU, DU, and CU functions by implementing them on different satellite sites. Such solutions can be critical for improving satellite resource utilization; however, the interface between different functions can be non-RT. Additionally, centralized control and resource pooling is feasible for such solutions with the centralized BBU operations.

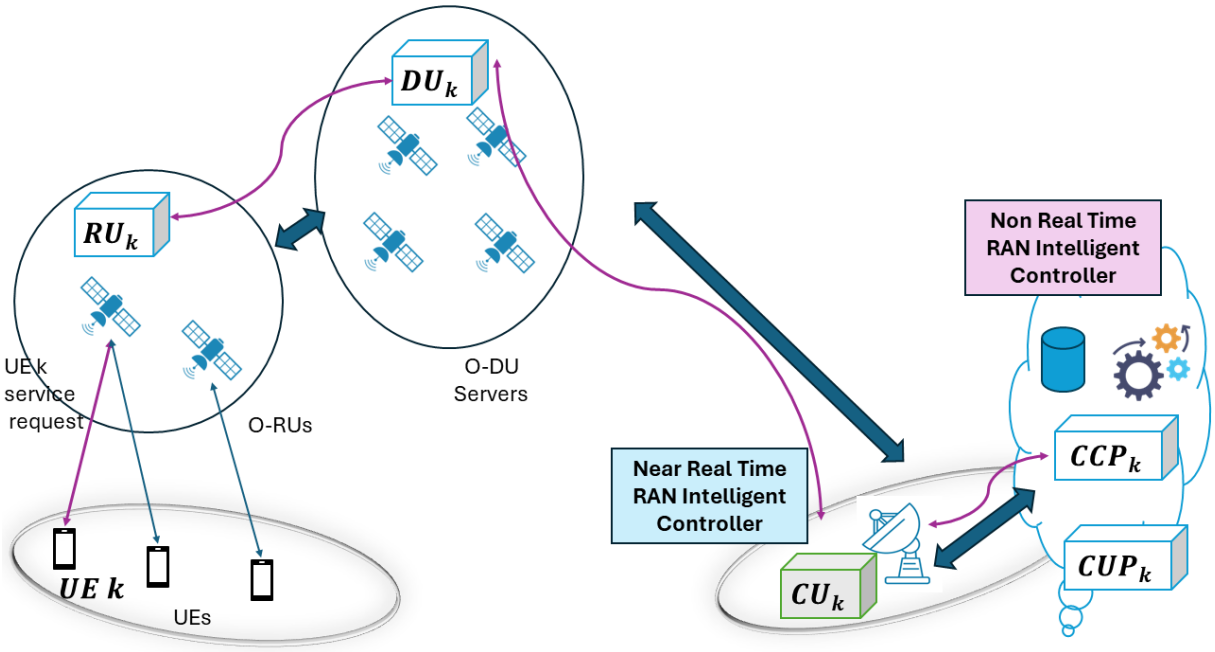

Figure 2-10: RAN Function Deployment- Option 6

Figure 2-10 shows another possible O-RAN function deployment option over space-ground networks. This option is defined as option 6 where the RU and DU functions are implemented over satellite nodes connected through ISL. The CU facilities are located at terrestrial control stations which have backhaul connectivity to cloud-based core networking facilities.

In this case, the RAN functions are split between the multiple satellite sites and the control station, creating a distributed RAN solution. However, the induced latency can be higher and is not suitable for real-time applications.

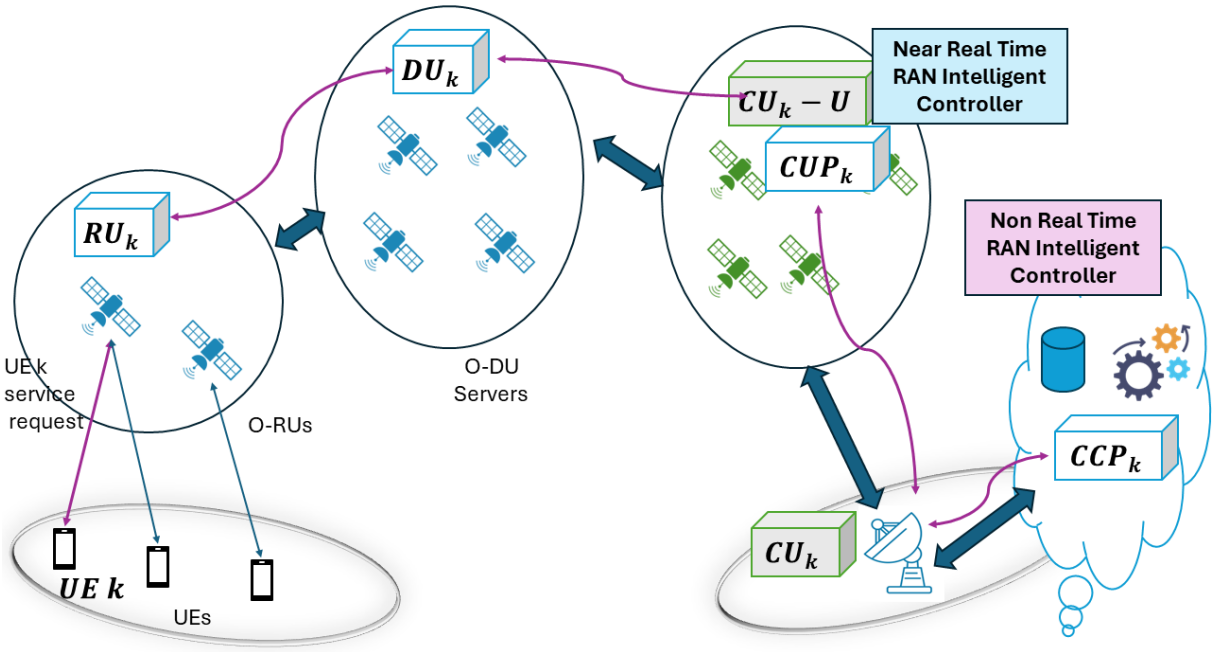

Figure 2-11: RAN Function Deployment - Option 7

In another case (Figure 2-11), to avoid excessive latency for user plane data traffic, the user place functions of O-CU and the core network, i.e., UPF are implemented over the space segment. Thus, O-CU functions are deployed in over two different sites. The CU-U corresponds to the O-CU user plane functions that are implemented over the satellite networks connecting them to the DU functions. The remaining CU functionalities correspond to the control options are implemented over the ground control stations. Similarly, user plane functions of core networks accompany the CU-U functions, while the remaining control functions are placed on cloud facilities located on the ground. Such architectures can effectively serve the users with reduced user plane latencies.

2.3.2 VNF Placement

In the previous parts, we defined the various possible options for O-RAN/core network function deployments over ground and/or space networks. A virtualized form of O-RAN and core networking functions can be chained to form a slice corresponding to the UE service request. Such network-slicing solutions induce the flexibility and possibility of implementing multiple services over the common infrastructure. Each slice can provide an independent service with specific performance. Given the dynamic nature of space networks and the varying demands of UEs, providing static slice solutions can be challenging and has limited performance. This requires the possibility of online updates of slice functions over time to achieve the service demands. In such cases, the VNFs should be moved from one satellite to another depending upon the mobility, available resources, and distance measures from UEs/ other satellites.

Figure 2-12, provides a possible VNF deployment solution for a dynamic satellite environment. In this case, initially, the service is enabled through the deployments of RU, DU, CU, and core networking functions over distributed space-ground nodes. The figure also shows the possibility of a change in the location of DU functionalities from one satellite node to another during the service lifetime, to avoid the drop in service quality due to mobility and other competing users. Note that as an example case, we have considered a RAN Function Deployment- Option 4 in Figure 2-10. However, such a VNF deployment solution can also be explored for other scenarios.

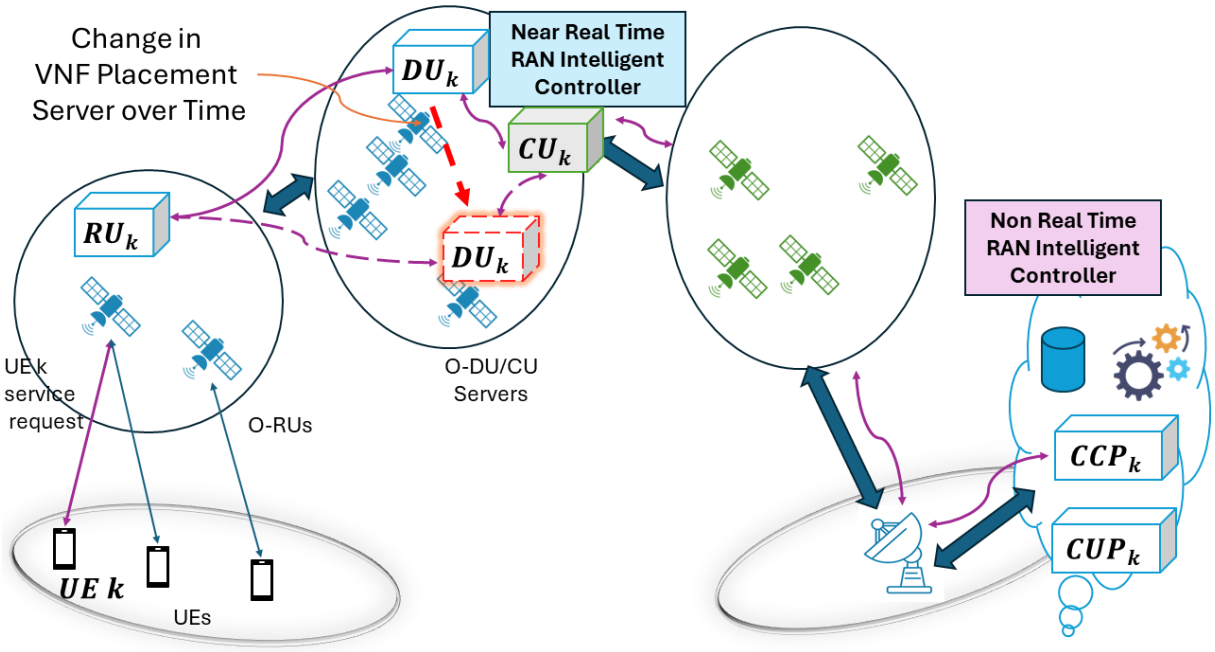

Figure 2-12: VNF Placement in Space

2.3.2.1 Core Network Functionality Placement

The deployment of core network functionalities is critically influenced by the location of the RAN network, with three main strategies tailored to enhance efficiency through the interdependent grouping of network functions as illustrated in Figure 2-12.

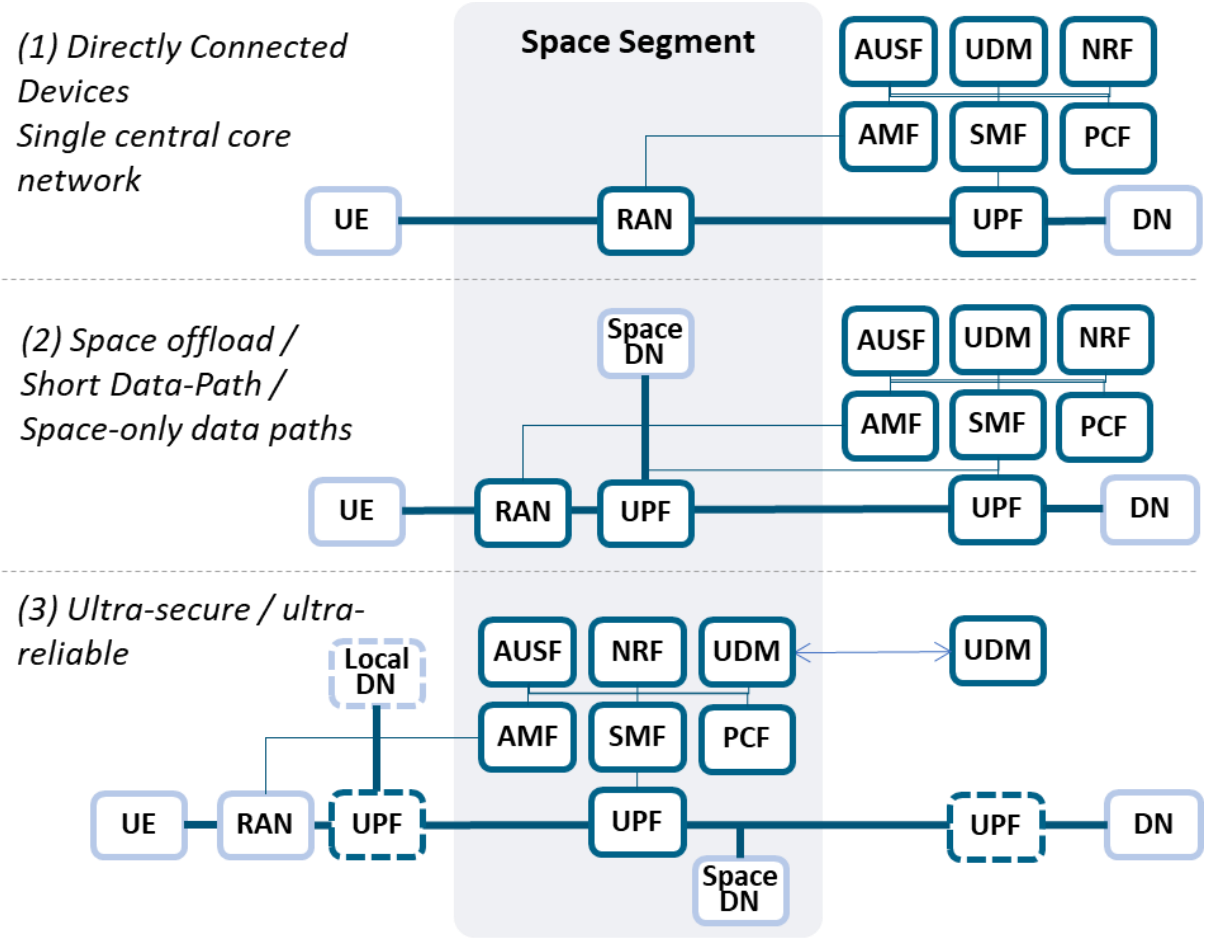

Figure 2-13: Core Network Functionality Placement Options

1) Single Ground Core Network - This approach maintains a central core network to which all ground stations connect to. It requires few resources, has a simple network management and integrates seamlessly with terrestrial networks. It is beneficial due to its simplicity and ability to manage all satellite-capable subscribers. Multiple UPFs may be deployed within different ground stations to reduce the end-to-end data path. Still, its major drawback remains the distance of the UPFs from the UEs requiring long delay data paths even for low Earth orbit and a significantly long signaling plane to reach the UEs and the UPFs.

2) Space Offloading - By relocating the UPF to space nodes, this strategy enables the direct connection to space-deployed services and shortens the path between UEs, effectively bypassing extended routing through ground stations. This setup allows for quicker data transfers and reduced latency. Also, the space-UPF functionality can be significantly reduced as to conserve the space node resources, with deferring non-essential tasks to be processed asynchronously by a terrestrial UPFs.

3) Fully Integrated Space Core Network - This approach places the UPFs and the entire control plane in space, creating an ultra-secure, reliable connectivity framework without interactions with any terrestrial network. Depending on the RAN's location, UPFs can also be placed at the terrestrial endpoints for example to facilitate direct connectivity between globally distributed enterprise locations. However, the communication service requires inputting user

profile data from a terrestrial administrative portal into the UDM as potentially synchronizing the communication accounting reports.

Options 2 and 3 require specific self-organizing features, adapting dynamically to operational demands and satellite movements. Space UPFs may be collocated with the RAN CU to minimize data paths for local connectivity, although this arrangement demands frequent handovers due to satellite mobility. An alternative is to maintain a UPF on a single satellite longer, reducing handover frequency but potentially distancing it from its service area, even on the other side of the Earth. A better alternative is to make decisions on dynamic spawning of ne UPFs and whether to have logically localized or orbiting around in the same satellite depending on the 5G system load. This flexibility allows the network to respond to varying load conditions and communication requirements dynamically, adjusting UPF deployment based on real-time data.

Moreover, in Option 3, the complete core network potentially deployed in designated space nodes facilitates enterprise connectivity by allowing direct links across a its distributed locations without reliance on external networks where control plane is less important compared to increased security and reliability. Additional dynamic deployment of UPFs at ground enterprise locations enable and specific APNs for them, ensure the availability of tailored and secure enterprise data paths.

2.3.3 Network Slicing

The network-slicing solutions over shared resources of space-ground networks can be implemented through different options. Here, we present the E2E network slicing framework, how resources can be allocated for implementing slices and two possible ways to implement the network slice solutions defined as independent and shared resource-based slice options.

2.3.3.1 E2E network slicing

An end-to-end slicing model considers slicing in RAN, Transport Network (TN), and CN. The network slice modelling approach from Standards Developing Organizations (SDOs) do not consider end-to-end slicing model but each domain independently. One of the main reasons to adopt this approach is due to the fact that each domain is standardized by different SDOs. While the standardization in RAN and CN domains is handled by 3GPP, TN falls under IETF. Traditionally only backhaul links were considered as part of TN, but the 5G RAN architecture discussed in Section 2.1.1 opens up new dimensions to the TN. The RU, DU, CU separation necessitates the presence of Fronthaul link between RU and DU, Midhaul link between DU and CU in addition to the Backhaul link which connects RAN to the CN. The overall end-to-end architecture is depicted in Figure 2-14.

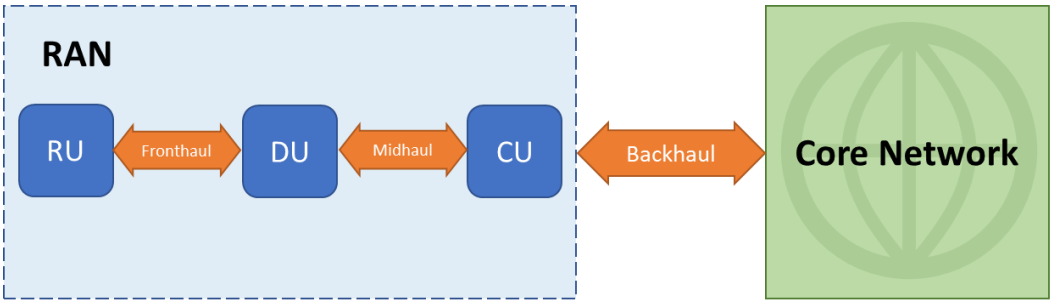

Figure 2-14: End-to-end Network

A multi-tier orchestrator shown in Figure 2-15 can help in the efficient management of the slices. The end-to-end orchestrator receives the overall specifications for a slice, which is then passed to the respective domain orchestrators. Each domain orchestrator is responsible for creating the network sub-slices required to satisfy the SLAs. The orchestrators at either level can take advantage of Network Data Analytics Function (NWDAF) [26] to draw conclusions for decision making. The NWDAF can provide slice load level information and slice specific network status analytics information to the orchestrator. It can also train Machine Learning (ML) models and expose new training services. Hence, it plays a pivotal role in the implementation of AI services for orchestration.

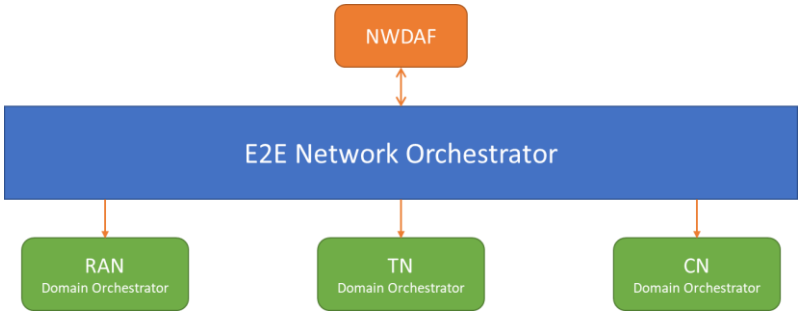

Figure 2-15: End-to-end Network Orchestrator

2.3.3.2 Network slice resource allocation approaches

A network slice consists of multiple network slice subnets. The network slice subnet represents a group of network functions (including their corresponding resources) that form part or complete constituents of a network slice [27]. Multiple slice subnets might be needed to realize a slice in a domain (let's say, sub-slice). This division of a network slice which spans across different domains help in independent management of the network slices. The allocation of resources in each of these subnets/sub-slices can be carried out in two ways:

- Dedicated resource allocation: The resources allocated are not shared among different network slices.
- Shared resource allocation: The resources allocated are shared among multiple network slices. This sharing may be direct or indirect. The direct sharing implies that the network slice subnet is offered as network slice multiple times. The indirect sharing implies that the network slice subnet is either a constituent of a network slice subnet shared by two or more network slices or is shared by two or more network slice subnet(s), which are in turn offered as different network slices.

The above-mentioned resource allocation options can be considered at intra-slice and inter-slice level. The end-to-end orchestrator handles inter-slice level slicing, and the domain orchestrator is in charge of slicing at intra-slice level.

Inter-slice resource allocation approaches are shown with examples in Figure 2-16. The end-to-end network slices are labelled with NS, while RAN, TN, and CN slices are represented with labels R, T and C respectively. For the first slice NS1, all the domains have assigned dedicated resources. This approach is useful for mission critical applications to guarantee uninterrupted connectivity without congestion. The configuration can also be deployed in scenarios where security is an important factor. In the cases of NS2 and NS3, only the sub-slice T2 is shared. This configuration ensures good performance as only the backhaul link is shared. The other configurations NS4, NS5, NS6 share resources in multiple domains. These slices can be used in scenarios where an isolated, dedicated resources are not of major concern. Resource sharing at multiple domains is an efficient approach in scenarios where the slices are deployed in the network infrastructure (TN/NTN). In integrated TN/NTN scenarios, the domain orchestrator should take into account the fact that the resources required to satisfy the same SLAs across type of networks differ, hence the resources have to be assigned accordingly.

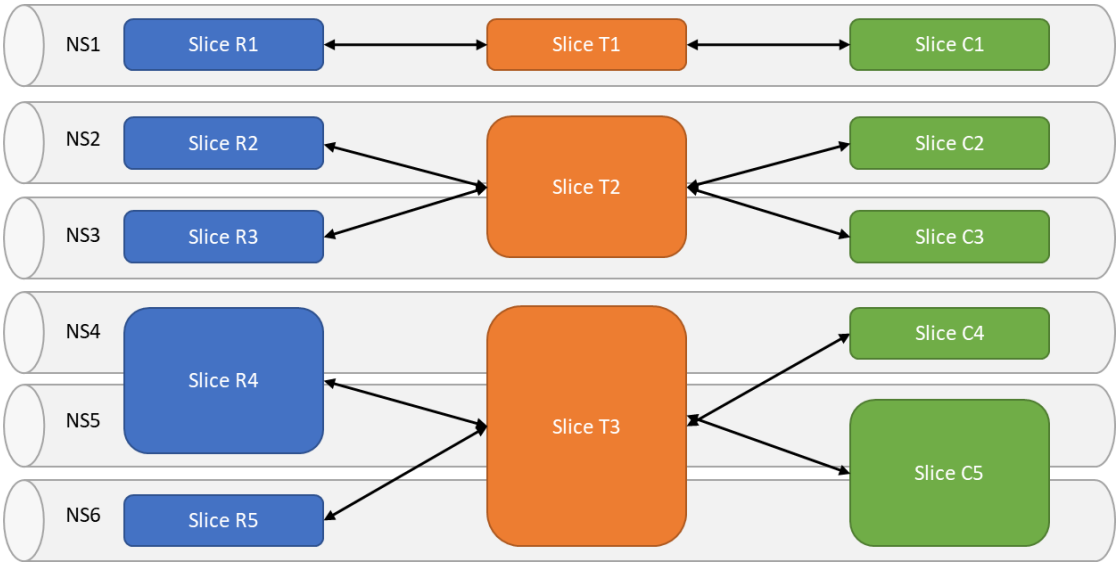

Figure 2-16: End-to-end Network Slices - Dedicated and Shared

Intra-slice level resource allocation can be performed in each domain. As an example, we have considered the Intra-RAN slicing using O-RAN architecture for the space segment below.

2.3.3.3 Multiple Network Slicing Deployment Options

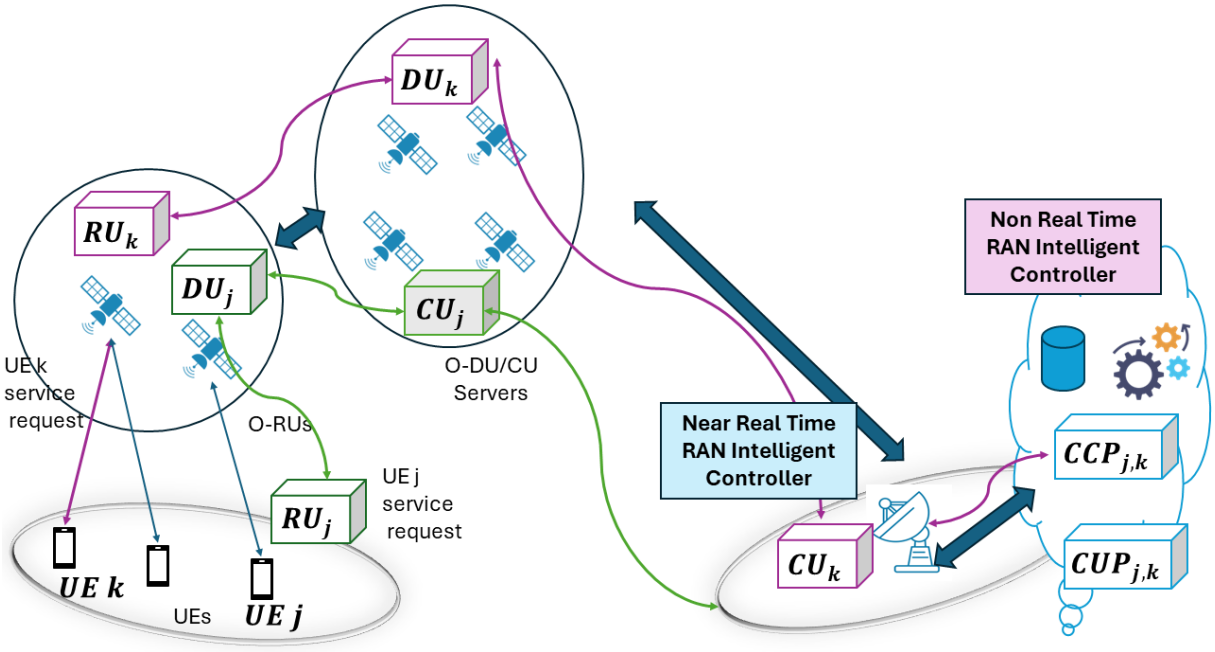

Figure 2-17: Multi-user independent slicing solutions

In the case of an independent slice solution, each user service demand can be satisfied through the independent slice function chain. In such cases, each slice function can have access to the dedicated resources for the entire lifetime of the service. In Figure 2-17, we have shown the example case of two slices implemented through the space-ground network facilities where both slices are completely independent and do not share any functions. In addition to this, dedicated communication links with proper data rates are also allocated to each service. Such an approach can provide the highest reliability in terms of service guaranty for the cost of poor resource utilization.

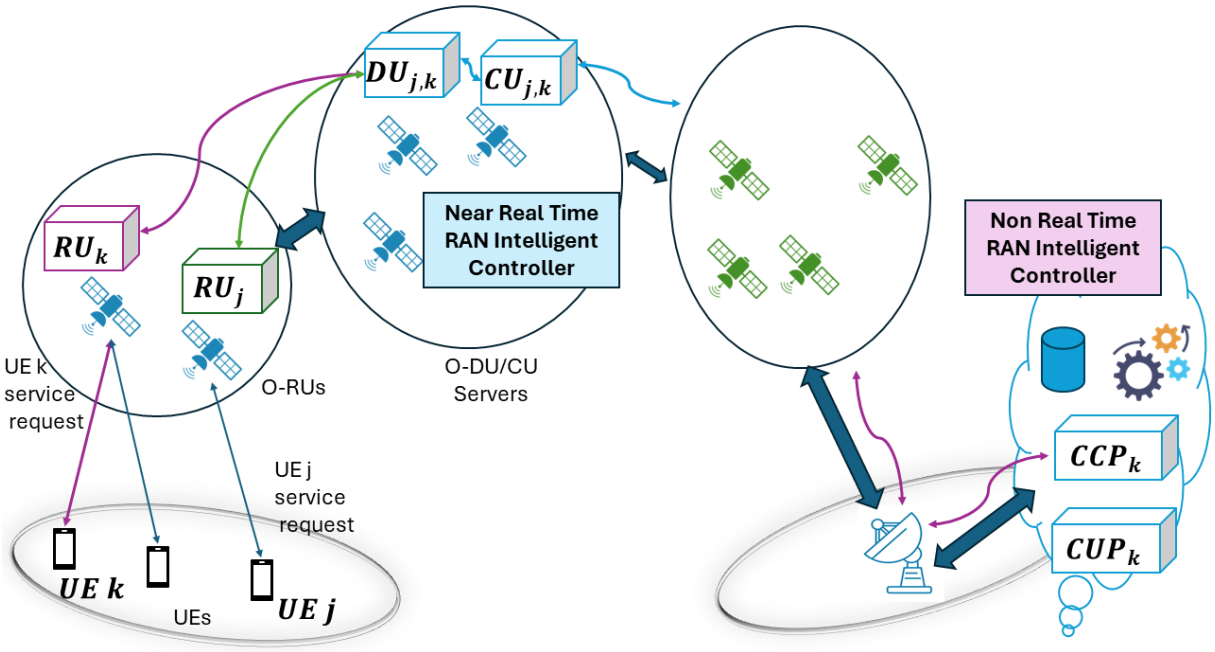

Figure 2-18: Multi-user Shared/dependent Slicing Solution

In another case, network slices can be implemented by combining network functions shared by different service requests. In such cases, slice functions can be implemented through the resource-sharing approach, allowing them to serve different users simultaneously. Proper resource allocation and session management strategies are required for adequate performance. In Figure 2-18, we have presented the example solutions where two users are served through the dependable slices sharing the multiple RAN functions. Such an approach can improve the overall utilization of distributed edge computing/communication resources. However, without proper resource allocation and session management strategies in place, such solutions can supper with reduced reliability.

2.3.4 Multi-connectivity/Network Selection

Multi-connectivity solutions can allow users to connect to the different access networks. In Figure 2-19, we present the example case of a multi-connectivity solution utilized by ground users while accessing space networking resources. In particular, two O-RAN base stations are deployed: one for the NTN network and one for the TN. The two base stations are pertaining to two separate 3GPP systems converging only at the core network level. As there is no communication between the two base stations, a multi-3GPP connectivity is considered. This solution is preliminary developed in D5.2 where a large number of details are described.

This deliverable is concerned to the decision of which of the multiple connections to be used in order to increase the network resilience. These algorithms will enable the splitting of the data traffic across multiple links based on the architecture depicted at high level in Figure 2-19 and described in detail in D5.2.

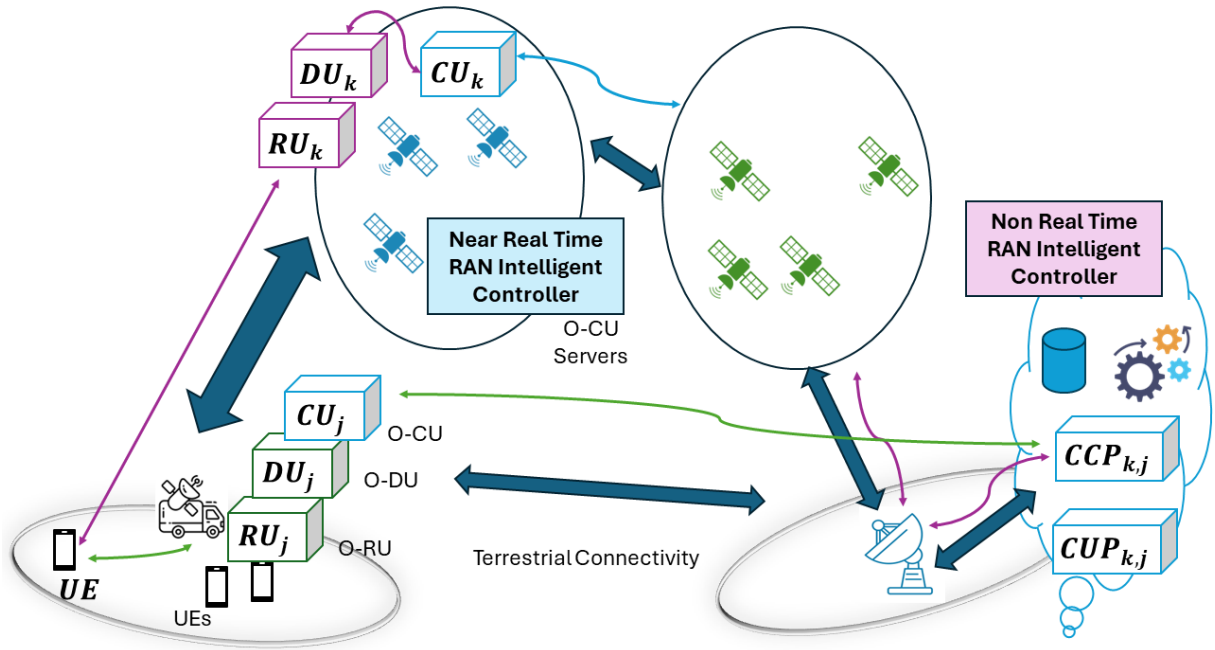

Figure 2-19: Multi-connectivity solution RAN function deployment

The link selection decisions have to be taken very fast after an event has happened such as the loss of a link or the drastic reduction of its capacity. Furthermore, as the main events related to the data path selection are related to the measurements taken at the UE and as these events may not be easy notified to the network (especially link loss), such decisions should be taken by the UE. This is in line with other decisions such as the access network discovery and selection or the user data traffic routing considered in the 4G and 5G architectures. However, the network may transmit indications on how to behave on the specific events to be able to reach the appropriate decision. Depending on the accuracy and the time duration of such indications, they can be generally classified into three levels:

- Policies – the network transmits general policies to the devices on how to handle the different events, similar to the Access Traffic Steering, Switching and Splitting (ATSSS) policies (3GPP TS24.193). In this case the algorithm making the selection is completely running in the UE adapting in real-time to the specific changes.

- Conditional modifications – derived from the Conditional Handover (CHO) mechanism of the 5G RAN (3GPP TS38.401,"NG-RAN; Architecture description."), the network transmits indications to the UEs how to behave on certain conditions when the communication with the network would not be possible. In this case, the algorithm is split between the network which defines the conditional modifications and the UE which has to apply them when the conditions happen. Depending on how granular the conditions are defined, the conditional modifications may have a long-term effect similar to policies or very short-term effect similar to commands.

- Commands – the network transmits behavioral commands based on decisions taken in the network to the UEs. This type of operations is considered too fine granular for the network to be considered for further implementation. As mentioned before, it requires a large amount of information from the UEs and the possibility to send commands to the

UE after the decisions are taken, which would result into a very large delay considering at least one of the networks is non-terrestrial.

As such most of the decisions would be taken by the UE depending on its own measurements. In order to be able to implement AI optimizations for such decisions a basic federated learning UE solution may be deployed.

2.4 CONTROLLER FRAMEWORK

For the case of the proposed O-RAN-based space networking infrastructure, non-/near-RT RIC controllers can be explored for enabling the intelligent and optimized access network for terrestrial User Equipments (UEs). RAN controllers can be placed at various locations to optimize service-based performance. In the O-RAN-based NTN framework considered, RAN controller functionalities can be placed at different locations depending on the service demands. Placing the controllers at centralized locations, i.e., the cloud, can enable the centralized control of multiple RAN elements with relatively higher latency demands for E2E control loops. In such scenarios, controllers can explore the abundant resources of centralized processing sites for various control operations. On the other hand, controller blocks can also be placed in proximity of UE locations, i.e., on the local cloud to enable real-time control operations. However, in such cases, the performance of controllers can be limited due to the resource-limited nature of platforms, and centralized policies might not be feasible. Some studies have also shown the possibility of disaggregating the RAN controller functionalities and placing them at different locations based on the end user demands. Such an approach can enable flexible and service-based RAN control processes [13].

Various policies can be developed and implemented through SMO and controllers for efficient placement of RAN functions, network slice implementations, etc. Policies can be developed over RAN controllers by exploring the data collected through various interfaces. In such cases, various AI/ML tolls can be explored. The ML-based approach can enable controller devices to explore the collected datasets to extract useful information and develop intelligent policies.

In the O-RAN environment, ML solutions can be deployed through various options. For a given problem and the corresponding requirements, various ML deployment policies can be adapted over a distributed RAN environment. In particular, ML process involves the collection and preprocessing of ML datasets, ML model training, model management, inference, collection of deployed model performance, and corresponding feedback information.

In particular, O-RAN controllers can develop optimal policies by exploring various traditional optimization approaches, including the heuristic and metaheuristic approaches developed and deployed as rApps or xApps. On the other hand, controllers also support the development and deployment strategies of ML solutions through the exploration of RAN data collected over various interfaces. In the following, we introduce the possible options for implementing AI-based and optimization solutions in O-RAN-supported network infrastructures. While in this deliverable we focus on several different possible options for optimizing the controller and placing the different AI blocks, in the following of the work we will shortlist the possible options to one, or two at the maximum, by selecting the most promising.

2.4.1 Traditional Optimization Solutions

Various control and optimization policies can be developed in O-RAN-based networking infrastructure to optimize the performance. The traditional optimization solutions, i.e., convex optimization along with heuristic and meta-heuristic approaches, can be used to define RAN policies based upon service requirements. The data required for such solutions can be

collected through different RAN interfaces and external sources. Providing the global optimal solution over a dynamic and complex multi-layered space network infrastructure can be hard and may not be feasible. In such cases, optimization problems can often be relaxed and simplified to certain assumptions. In another case, simplified heuristic and meta-heuristic solutions can be developed by inducing the static RAN policies with suboptimal performances.

The O-RAN controllers can support such traditional optimization solutions through rAPP/xAPP environments based on real/non-real-time performance requirements. Optimization engines and data collection strategies can be developed and deployed in non-/near-RT RIC as applications. Figure 2-20 visualizes the possibility of deploying the traditional optimization solutions over O-RAN control environments including non-RT RIC and near-RT RIC.

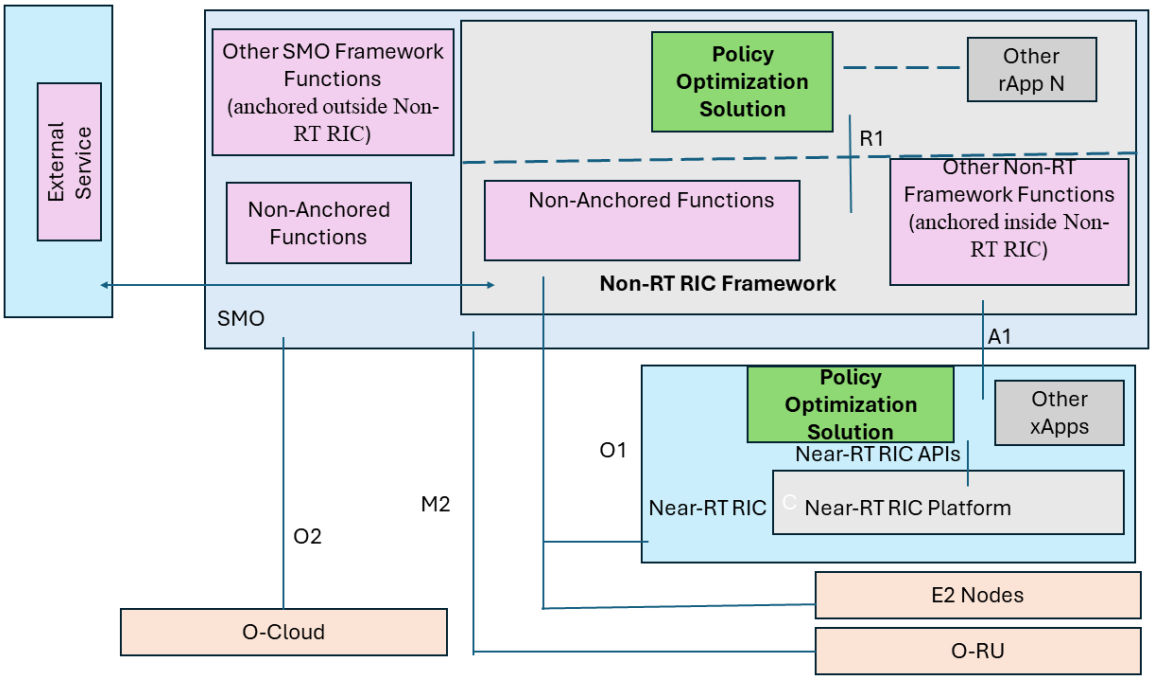

Figure 2-20: O-RAN Policy Optimization Solutions Hosts

2.4.2 AI-based solutions

AI can be a powerful tool to enable optimal policies based on past experiences. In particular, various ML algorithms can be applied to the proposed O-RAN-based space networking scenarios to enable intelligent RAN policies for generating various control actions. O-RAN controllers can collect data from different NFs, external sources, and O-cloud infrastructure to generate ML data sets that can be used to train the ML models. Based on the intelligent applications demands, various ML strategies can be adapted over O-RAN architectures. In particular, with its centralized location with a large amount of computation and storage resources, SMO/non-RT RIC can train sufficiently complex ML models. However, it can induce higher data collection costs and the corresponding training latencies. Such an approach cannot guarantee the real-time performance. However, the ML models can be pre-trained at non-RT RIC and then deployed over other RAN elements for further adapting the policies according to the local datasets. Figure 2-21 highlights the various ML functionalities provided by SMO/non-RT RIC for enabling the ML solutions. In particular, the ML functionalities are part of a non-anchored group of functions inside SMO/non-RT RIC which can be supported by the other anchored functions. ML solutions can be deployed through one or more rApps that consume or generate various ML-related services.

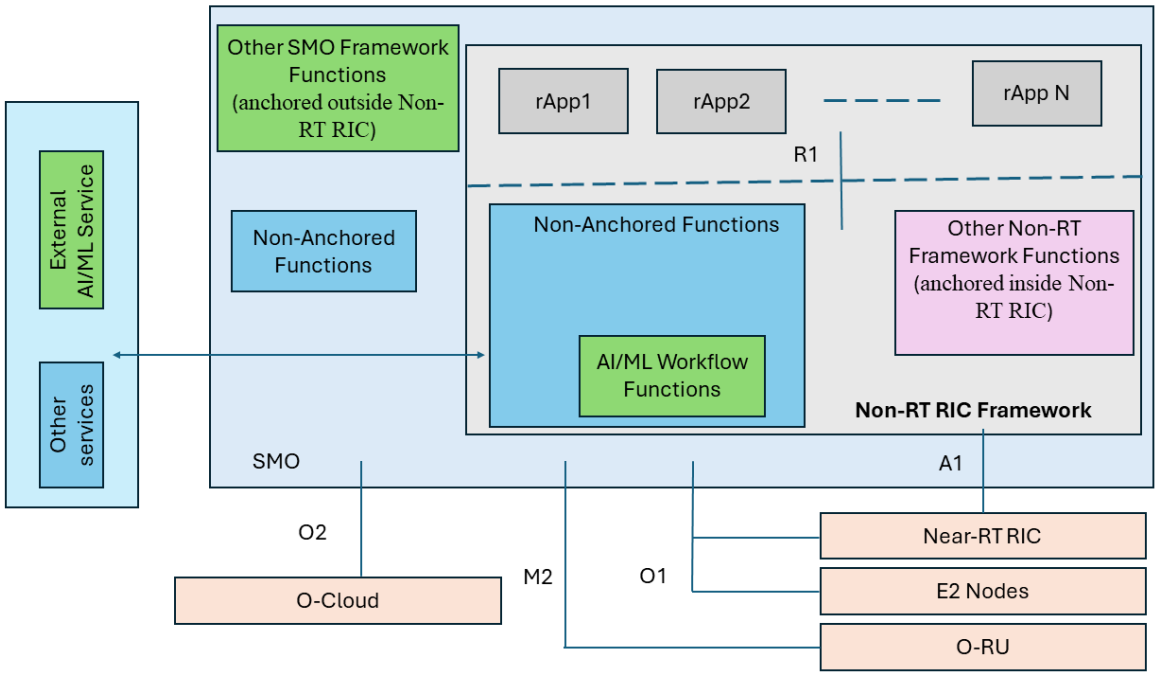

Figure 2-21: O-RAN non-RT RIC ML functionalities.

Another possibility to implement ML over O-RAN architectures is to explore near-RT RIC functionalities to enable ML-based solutions. In such cases, data and policies can be collected over E2/A1 interfaces and used to train ML models in a near-RT RIC environment. Figure 2-22, shows the various ML functionalities provided by near-RT RIC through its platform. The AI/ML support function along with the database and its dedicated space for xApps can implement the entire ML process over near-RT RIC through various functionalities. Data can be collected from E2 nodes or from non-RT RIC through the E2/A1 interface. The trained models and their performance can be stored in a database for future uses.

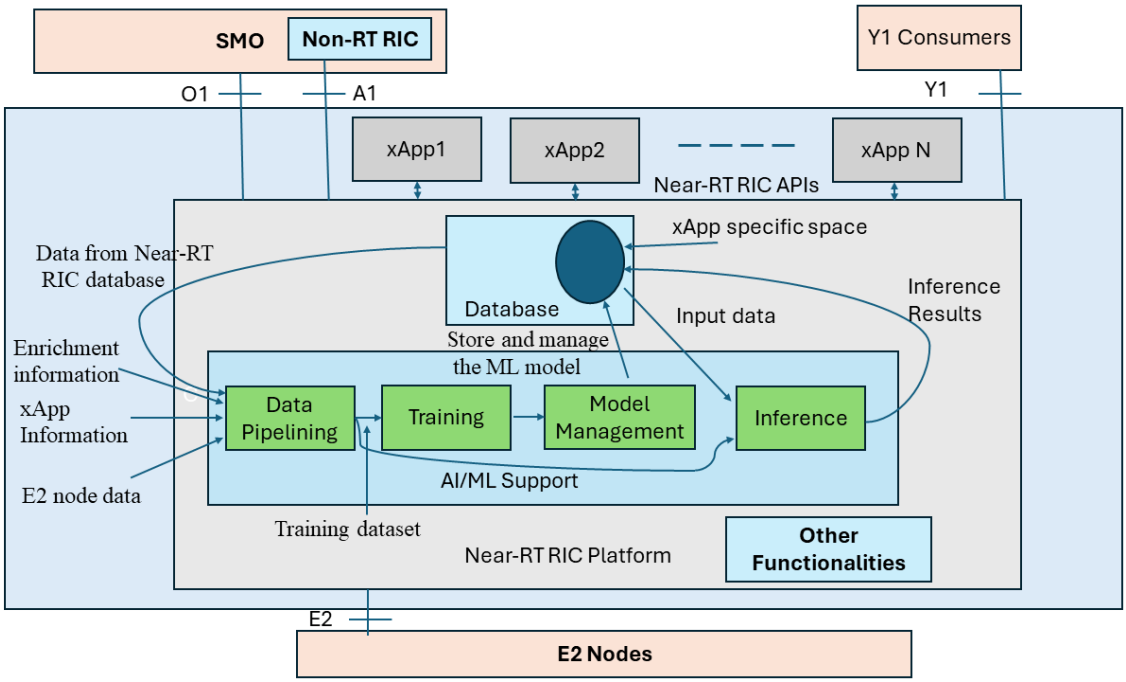

Figure 2-22: O-RAN near-RT RIC ML functionalities.

To enable the intelligent ML-based solution over the O-RAN framework one has to select the proper facilities for data collection, model training i.e., training host, model inference, the entity over which control action is applied, i.e., actor node, etc. Based upon the proposed O-RAN-based space networking framework and the O-RAN specifications, here we introduce several options for the development of intelligent solutions.

2.4.2.1 Scenario 1

Scenario 1 (Figure 2-23), includes the possibility of deploying the ML solution over SMO/non-RT RIC functional blocks. Here, non-RT RIC is selected as the training and inference host while the near-RT RIC or E2 can be actor. Training data can be collected by SMO through the O1 interface. Data processing is done inside non-RT RIC for generating the training/inference datasets. The training process and the AI/ML model management are part of non-RT RIC where model training, storage, and model management functions are performed. The inference process is also performed inside non-RT RIC. The policies/control actions are then provided to the near-RT RIC/E2 nodes according to the actor node selection.

In general, non-RT RIC can be associated with platforms having centralized locations with higher resource capabilities. Such ML deployment scenarios allow for centralized control policy generation and deployment through the use of SMO/non-RT RIC functionalities. With the availability of centralized datasets and an enormous amount of resources, such ML deployment policies can be effective in enabling complex neural networks (NN)-based solutions with non-RT performance.

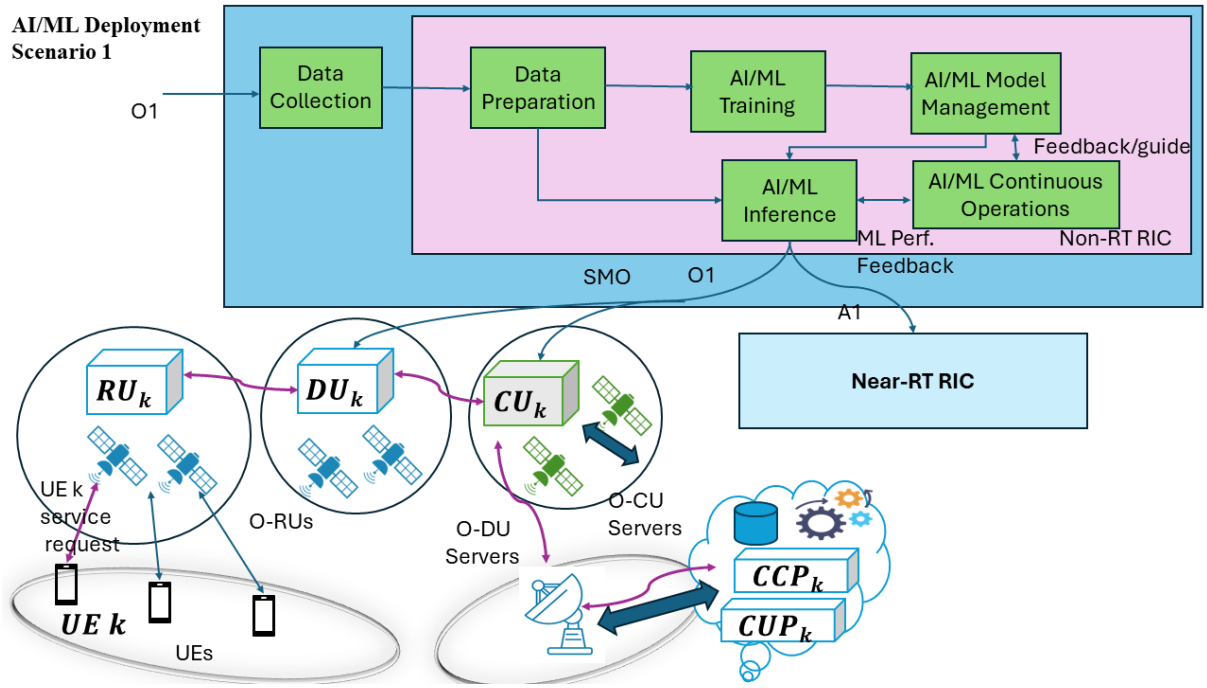

Figure 2-23: O-RAN ML deployment: Scenario 1

2.4.2.2 Scenario 2

In scenario 2 (Figure 2-24), the ML process is hosted between SMO, non-RT RIC, and near-RT RIC. In particular, the data collection and model management processes are performed by SMO, while the data preparation and model training are performed by non-RT RIC. Next, model inference is hosted in near-RT RIC where online information collected from E2 nodes is used to make decisions. Control actions can be implemented on E2 nodes for various purposes.

In this case, both SMO/non-RT RIC and near-RT RIC are participating in the ML process. With the use of SMO/non-RT RIC for data collection and model training, such an approach can enable complex ML solutions. Additionally, with the near-RT RIC-based inference process, the latency requirements for control loops can be reduced. Additionally, the real-time RAN performance data can be used during the inference process. However, it is important to analyze the cost required to deploy the models based on the locations of two different controllers. Such solutions cannot be ideal for cases where frequent retraining of the ML models is required due to changes in environments, dynamicity, etc.

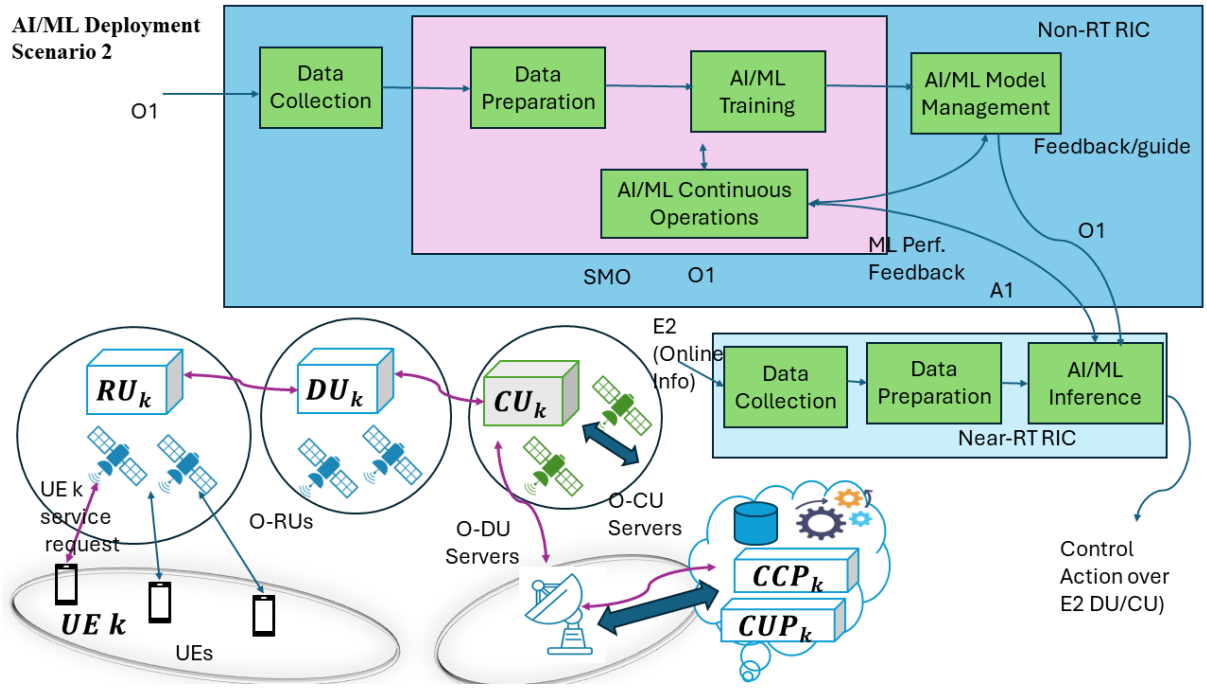

Figure 2-24: O-RAN ML deployment: Scenario 2

2.4.2.3 Scenario 3

Scenario 3 (Figure 2-25) provided the possibility of distributing the ML process between SMO and near-RT RIC through non-anchored ML dataflow functions. This includes SMO functionalities such as ML data collection, processing, ML training, and model management functions. Additionally, the inference process is hosted by the non-RT RIC.

In this case, the complete ML process is divided between SMO and non-RT RIC. In particular, the data collection, pre-processing, and ML model training are performed by SMO while model inference is hosted inside non-RT RIC. Such solutions can be ideal for ML scenarios where the ML data required to train the models is much higher. This includes data from O-cloud, O-RAN NFs, and external sources. However, such training approaches cannot guarantee real-time performance and are thus suitable for generating long-term policies with non-RT control requirements.

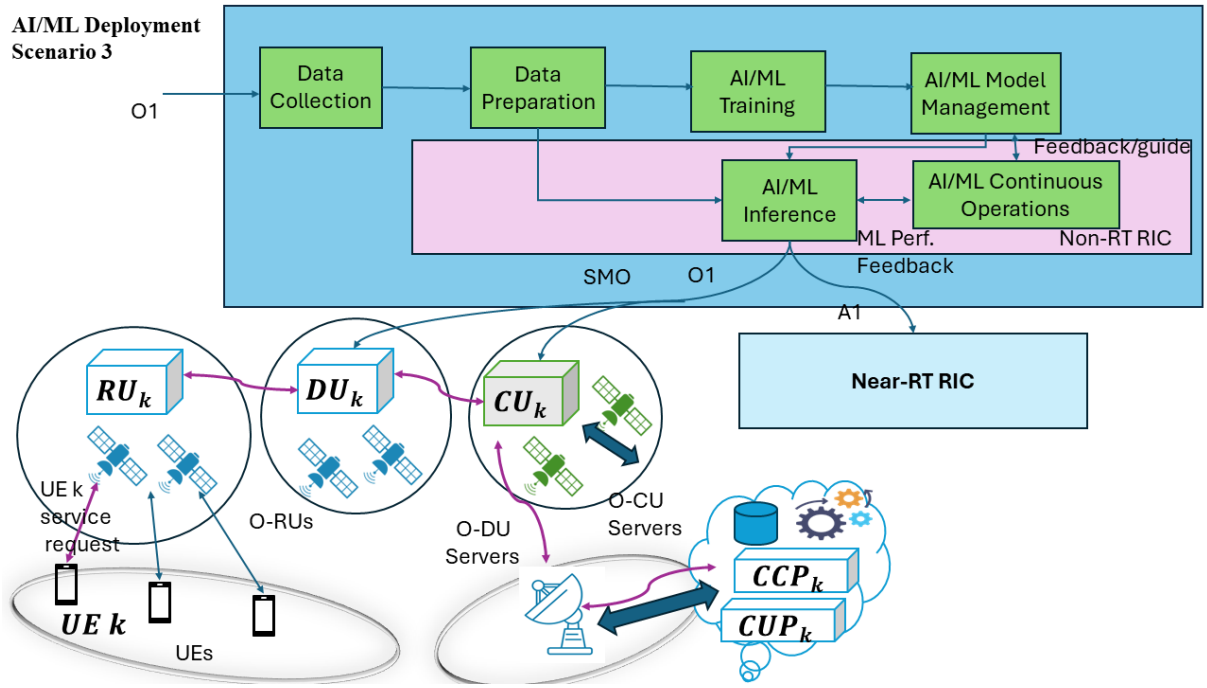

Figure 2-25: O-RAN ML deployment: Scenario 3

2.4.2.4 Scenario 4

Scenario 4 (Figure 2-26), includes the possibility of online training of the ML models pre-trained by SMO/non-RT RIC. The online training is performed by the near-RT RIC by downloading the model through the A1 interface and using the online data from the E2 nodes collected over the E2 interface. Data collection is performed by SMO over O2 for pre-training operations. The remaining functions of the pre-training process are performed within a non-RT RIC. Next, the online training and the model inference are hosted by near-RT RIC.

In this case, the initial model is generated by non-RT RIC and SMO together. Next, this model is used to drive the near-RT RIC policies by updating it through the on-line data collected from the E2 nodes. Such solutions can be ideal where complex ML models are considered to generate the control policies with near-RT performance through frequent updates.

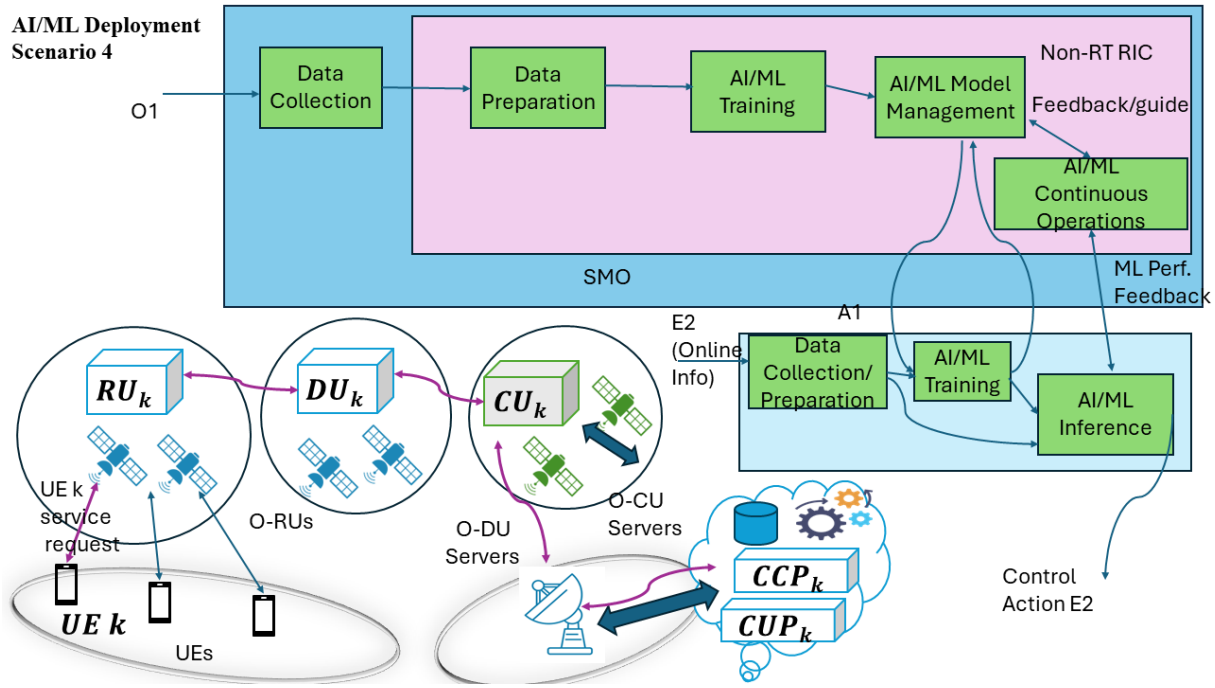

Figure 2-26: O-RAN ML deployment: Scenario 4

2.4.2.5 Scenario 5

Scenario 5 (Figure 2-27), allows the model inference to be hosted by CU/DU functions for reduced latency performances. The ML process is distributed between SMO/non-RT RIC and E2 nodes.

This scenario allows model data collection and training inside centralized SMO/non-RT RIC facilities, while model inference is hosted over E2 nodes. Such solutions can be ideal for cases demanding real-time control with long-term policy generations. Frequent updates of control policies according to the change in the RAN environments might not be suitable for such solutions given the cost required to download the models at E2 facilities from centralized non-RT RIC platforms.

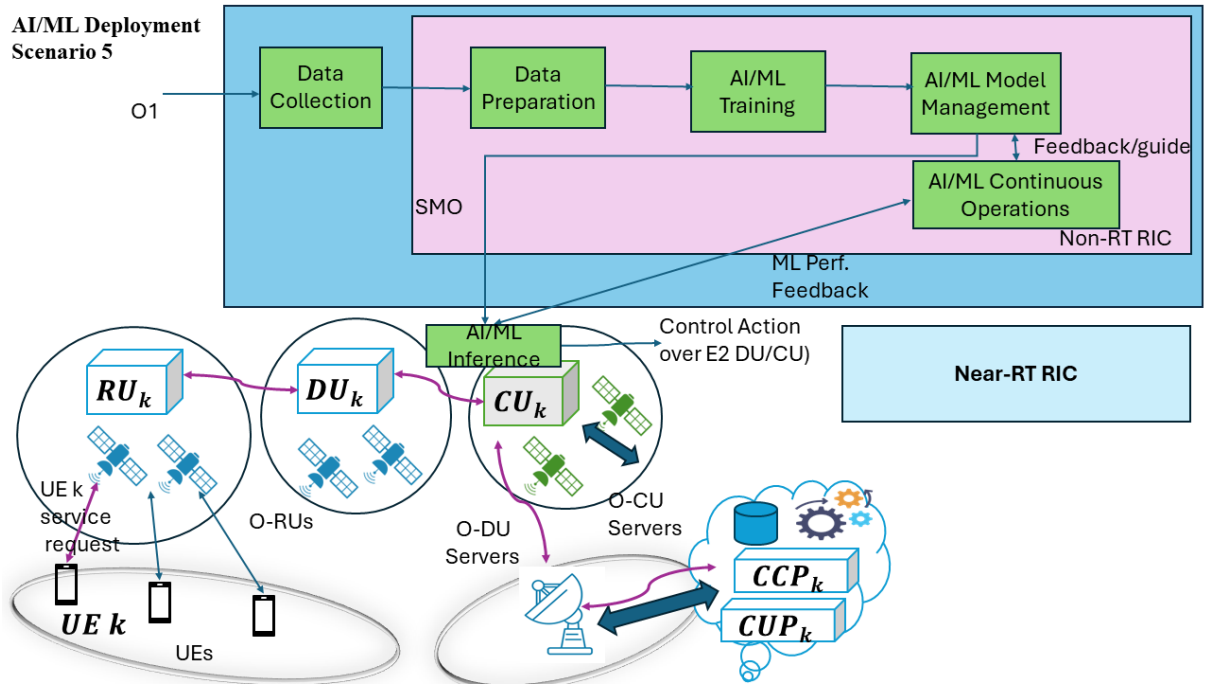

Figure 2-27: O-RAN ML deployment: Scenario 5

2.4.3 Centralized vs. Distributed AI

The O-RAN framework can support both centralized and distributed AI deployment options. In the case of centralized AI solutions, data from various RAN elements, O-Cloud, and external sources can be collected at the centralized location with a sufficient amount of processing and storage resources. In such cases, non-RT RIC can act as a centralized entity collecting and processing the entire data from different elements. Such a process can induce higher costs in terms of overall data transmissions and model training latencies and may not be feasible for real-time control actions. However, with centralized datasets, advanced ML solutions with complex models can be trained through such centralized solutions. Various solutions can be adapted to O-RAN-based networks to enable efficient centralized AI solutions. For example, it is possible to use the pretrained ML models and apply the real-time data collected over near-RT RIC through the E2 interface for updating the models according to the real-time RAN performance. Figure 2-28, shows the possible centralized AI solutions deployment policy over distributed O-RAN architecture. In particular, data can be supplied by O-RAN NFs, O-cloud, and external sources through various interfaces. These data can be collected over non-RT RIC acting as a centralized server. The ML model can be trained over non-RT RIC by using the collected data and then deployed over different actor nodes based on the performance requirements.

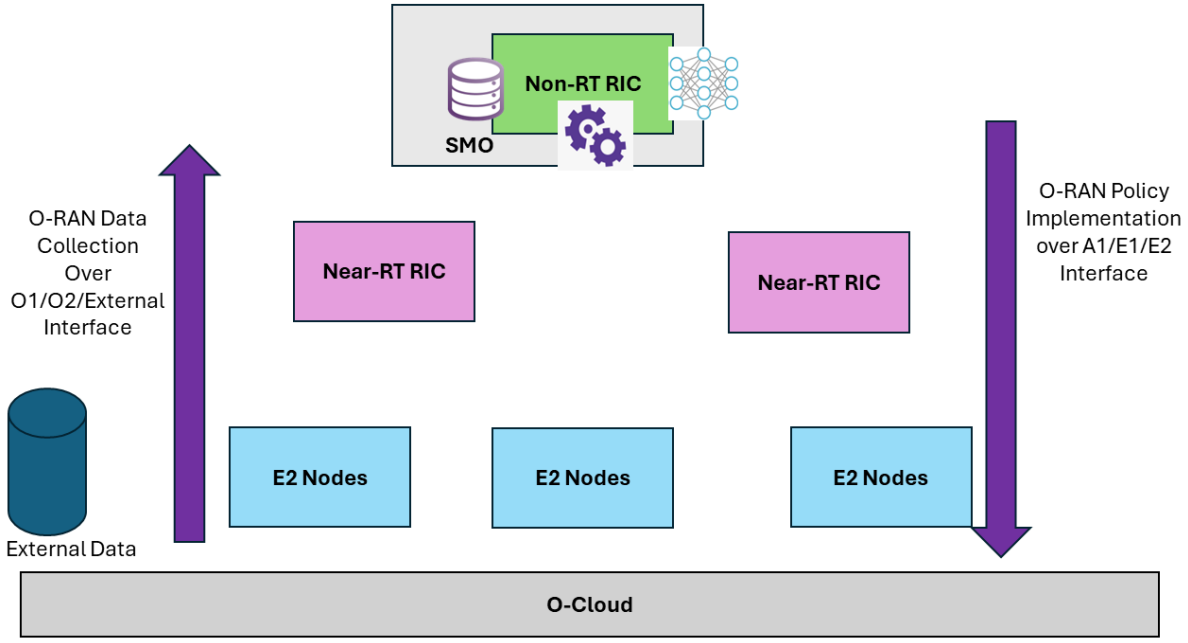

Figure 2-28: O-RAN Centralized AI Solutions

On the other hand, with O-RAN distributed control mechanisms, the distributed AI solutions can also be applied where multiple entities can train the ML models through local datasets, and with the support of a centralized processor they can share their knowledge to form a global model. Federated Learning (FL) is one such distributed AI solution that can enable the efficient distributed learning process over O-RAN solutions. In the case of O-RAN architecture, non-RT RIC can handle one or more near-RT RICs. The near-RT RIC can connect with a single non-RT RIC and can handle one or more E2 (i.e., DU, CU functions) nodes. Such a hierarchy can be explored by the FL approach, where real-time data is available at near-RT RIC. With its centralized position, the non-RT RIC can act as an FL server. In general, the FL process involves multiple iterations, where in each iteration the ML model is trained on local FL devices through local datasets, and the global model is collected from the FL server in the previous round. Next, the trained model parameters are shared with a centralized FL server which then aggregates the received models from all the FL devices to form a global model characterizing the training performance of all the devices. Next, this global model is shared with the local FL devices for the next round of the training process. The training process lasts until certain stoppage criteria are matched in terms of model performance of training time. Figure 2-23 shows the FL deployment case implemented over distributed O-RAN architecture. In this case, the E2 nodes are generating real-time O-RAN performance data and forwarding them to near-RT RIC. The near-RT RICs act as FL devices and perform the ML model training through local datasets. The non-R RIC is considered an FL server. Various other options for distributed AI deployment can also be verified through O-RAN architecture. architecture.

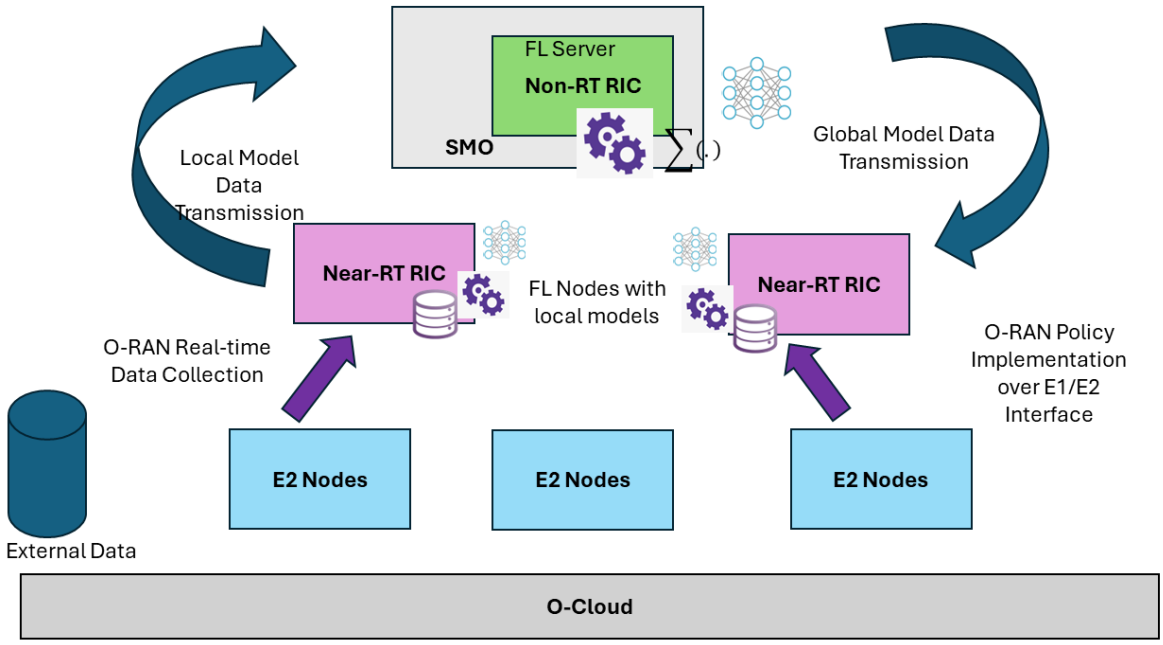

Figure 2-29: O-RAN distributed AI (FL) solution.

Beyond FL, several other distributed learning methods can be incorporated into O-RAN architectures to train ML models more effectively. These methods include:

1. **Split Learning**: This method divides the ML model into two segments: one part is trained on the device, and the other on the server. This not only enhances privacy, but also makes it more resource efficient for devices with limited capacity.

2. **Transfer Learning (TL)**: This technique uses existing pre-trained models as a baseline for training new ones. It can greatly improve model accuracy and reduce training time, especially when the number of training data is limited.

3. **Hierarchical Federated Learning**: This method structures the FL process hierarchically, involving multiple levels of FL servers and devices. It enhances the efficiency and scalability of the FL process in large-scale O-RAN implementations.

4. **Asynchronous Federated Learning**: This approach allows devices to participate in the training process at different times, without needing to wait for others to finish. This approach can reduce the overall training time and better utilize available resources.

By integrating these distributed learning methods into O-RAN architectures, the benefits of distributed processing can be harnessed, leading to more efficient training of the ML model.

2.4.4 ML Procedures

In this section we will compare different ML options (Supervised, Unsupervised, and Reinforcement Learning) and their pros and cons in the considered framework, along with a short description for each:

1. **Supervised Learning (SL)**

   Supervised learning involves training a model using labeled data to make predictions about new, unseen data. The model learns to map inputs to outputs based on the labeled training data.

   Pros:

   - **Accurate predictions:** SL models can make accurate predictions about new data.
   - **Efficient:** SL models can be trained quickly and efficiently, especially for classification tasks.
   - **Easy to understand:** SL models are relatively simple to understand and interpret.

   Cons:

   - **Requires labeled data:** SL requires large amounts of labeled data, which can be time-consuming and costly to obtain.
   - **Limited applicability:** SL is less effective for tasks that involve complex patterns or relationships in data.
   - **Overfitting:** SL models can overfit the training data, leading to poor performance on unseen data.

2. **Unsupervised Learning (UL)**

   Unsupervised learning involves training a model using unlabeled data to identify patterns and relationships within the data. The model learns to group similar data points together without any external guidance.

   Pros:

   - **Discover hidden patterns:** UL models can identify complex patterns and relationships in data without requiring labeled data.
   - **Cost-effective:** UL models can be trained using unlabeled data, reducing costs.
   - **Flexibility:** UL models can be used for various tasks, such as clustering, dimensionality reduction, and anomaly detection.

   Cons:

   - **Less accurate:** UL models are generally less accurate than SL models due to the lack of labeled data.
   - **Interpretability:** UL models can be difficult to interpret and understand.
   - **Requires human intervention:** UL models often require human intervention to understand the results and make decisions.

3. **Reinforcement Learning (RL)**

   Reinforcement learning involves training a model to make decisions in an environment by interacting with it and receiving rewards or penalties based on the actions taken. The model learns to optimize its behavior to maximize the rewards.

   Pros:

   - Real-time data processing: RL models can process real-time data and adapt to changing conditions.
   - Flexibility: RL models can be used for various tasks, such as control systems, robotics, and game playing.
   - Autonomous decision-making: RL models can make autonomous decisions without human intervention.

   Cons:

   - **Data-hungry:** RL models require large amounts of data to learn and improve.
   - **Complexity:** RL models can be complex and difficult to understand.
   - **Slow learning:** RL models can learn slowly, especially in complex environments.

Comparison and Recommendation

In the context of satellite non-terrestrial networks, RL-based solutions may be the best candidate for real-time data processing. RL models can adapt to changing network conditions and make autonomous decisions, which is crucial for ensuring reliable communication in these networks. For example, RL can be used to optimize network resource allocation, manage network congestion, and improve network performance.

However, the choice of ML approach ultimately depends on the specific requirements and constraints of the project. If accurate predictions are necessary, SL may be the best choice. If discovering hidden patterns in data is the primary goal, UL may be more suitable.

2.4.5 Main Use Cases and Optimization Framework

As mentioned above, various optimization and ML-based solution methods can be induced to optimize the RAN performance. Various control policies can be generated to optimize the RAN performance in terms of resource allocation, network selection, function placements, etc. to satisfy the UE's demands. Network slicing technology can be adapted to provide a set of heterogeneous services with specific performance over the same physical infrastructure. However, in such diverse service provisioning scenarios with stringent user demands, a proper selection of RAN split options, a proper placement of network functions, network selection, and optimal slice provisioning are important. In Figure 2-30, we have defined an optimization problem framework where this operation and its impacts are visualized.

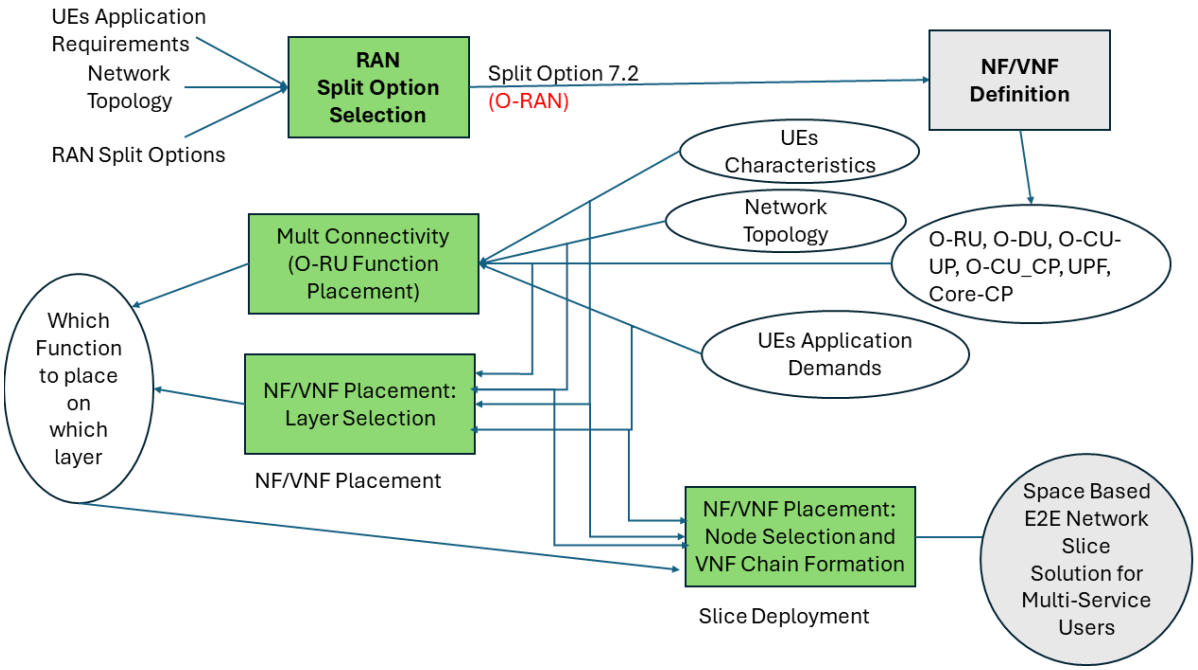

Figure 2-30: O-RAN Optimization Framework for NF/VNF Placements and Slice Deployments with Multi-connectivity.

RAN Split Option Selection: 3GPP has introduced several RAN functional split options to be implemented in distributed RAN environments. For a considered space system, we have introduced O-RAN as an access network technology, which by default considered 7.2 as a typical functional split option of the RAN.

NF/VNF Definition: Based on a selected RAN split option and the core networking functions, a set of NF/VNFs can be defined to facilitate the networking services over the proposed space networking infrastructure. Here, we define O-RU, O-DU, O-CU-UP, and O-CU-CP as RAN network functions. Additionally, O-RAN controller blocks can also be deployed as VNFs. Next, we also define UPF and core-CU as two network functions defining the user plane and control plane functionalities of core networks.

Multi-connectivity (O-RU Function Placement): In a considered scenario, UEs can connect to multiple access networks and should select appropriate RU facilities based on their own characteristics, network topology, application demands, and possible deployment options for O-RU. Therefore, proper deployments of O-RU functions are required according to the demands and characteristics of UE. Hence, we define an O-RU Function Placement for a multi-connectivity environment where the aim is to define an optimization/ML-based engine to find a proper O-RU function placement over a considered space networking infrastructure.

NF/VNF placement (layer selection): In a multilayered infrastructure, it is important to select a proper layer for NF/VNF placement according to the function demands, network topology, application requirements, etc.

NF / VNF placement (node selection and VNF chain formation): After selecting a proper edge layer for placing the NF/VNFs, it is also important to select a proper edge node/server for deploying the specific VNF over a given layer. Various potions can be available depending on the needs of the function, the availability of resources, and the application requirements. After deploying the VNFs over different edge nodes, they can be chained to form a network slice to serve UEs according to their requirements.

2.4.6 Time Scale of the Controller

The O-RAN architecture supports both real-time and non-real-time control loops through desegregated control mechanisms. The non-RT RIC located at centralized locations with large amounts of resources is expected to provide centralized control with fairly large time scales. On the other hand, the near-RT RIC can generate control actions on a moderate time scale to support the near-RT control processes. The RAN E2 nodes, that is, the nodes implementing the CU/DU functionalities, can monitor the RAN operations in real-time and support the real-time RAN control.

The control policies generated by non-RT RIC can also be shared with near-RT RIC through the A1 interface, where it can be used to refine the policies of near-RT RIC. In return, non-RT RIC can receive feedback from lower layers near RT RIC and E2 nodes in terms of RAN performance. Similarly, the policies generated by near-RT RIC can be forwarded to E2 nodes through the E2 interface to guide the E2 node's policies.

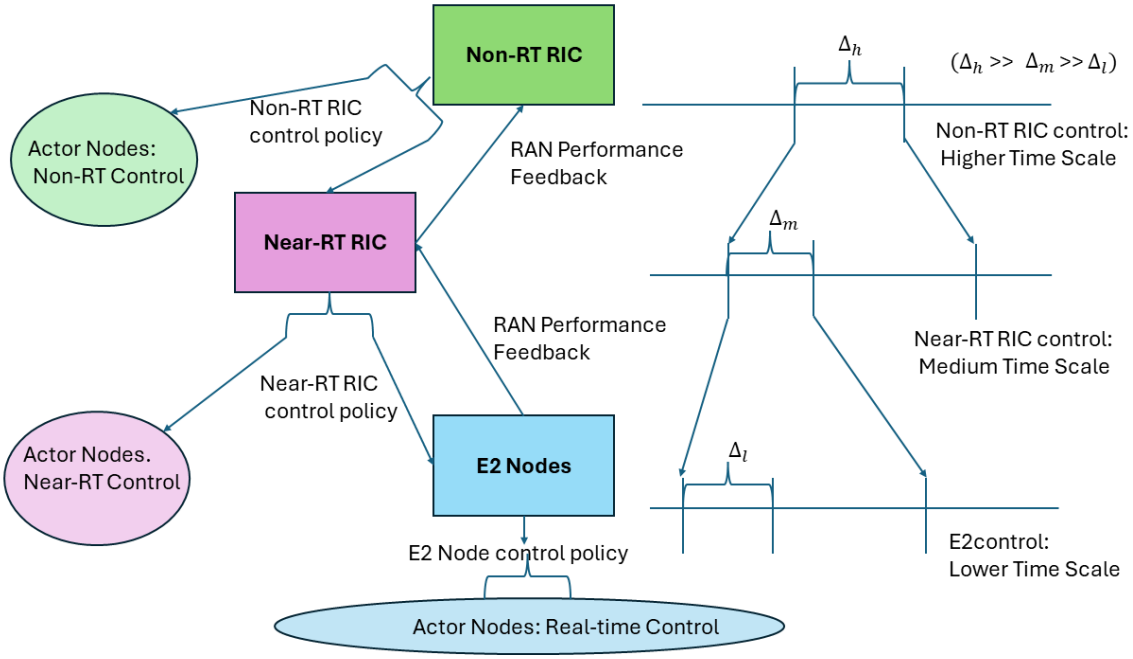

Figure 2-31: O-RAN Multi-Time Scale Control Process

Figure 2-31 presents the multi-time scale control process implemented over the O-RAN distributed controller blocks and E2 nodes. In particular, three-time scales are defined with different time steps. Non-RT RIC can provide non-RT control with a slow time scale with time step Δ_h. Next, the non-RT RIC control action can be forwarded to the near-RT RIC to enable the near-RT RIC control policies with near-RT performance. Non-RT RIC control action/policy can fine-tune the near-RT RIC control mechanisms through centralized control. In return, near-RT RIC can provide feedback in terms of RAN performance to guide and update non-RT policies in the next iteration. We define the near-RT RIC time scale as a medium time scale with a time step $\Delta_m \; with \; \Delta_m \ll \Delta_h$. Similarly, near-RT policies or control actions can be forwarded to E2 nodes through the E2 interface to enable efficient real-time control policies over them. It should be noted that E2 nodes require real-time control. Therefore, we define a fast time scale with a much smaller timestep for the E2 control process. In return, the E2 node can provide real-time RAN performance feedback to near-RT RIC, which can be used to drive policies of near/non-RT RICs. The E2 timestep $\Delta_l \; with \; (\Delta_l \ll \Delta_m \ll \Delta_h)$ is much smaller than the other two steps.

3 END-TO-END QOS MANAGEMENT

3.1 INTRODUCTION

With the widespread adoption of 5G as a communication standard, mega-constellations have emerged as a viable alternative to terrestrial networks, offering extensive and reliable communication services across a broad spectrum of users and applications. These constellations, organized as grid networks, composed of multiple satellite layers and intersatellite links, provide different options for long-distance 5G communication services. However, their continuously changing architecture makes transport variable, requiring that new adaptation mechanism is developed to ensure the end-to-end QoS. To reach a complete QoS solution, we first present the different QoS segments of an end-to-end connection and then introduce a dynamic QoS assurance mechanism for mega-constellation both as concept and as high-level architecture. We propose to implement the end-to-end QoS by adapting the mega-constellation routing level and the 5G service path depending on the momentary conditions as well as introducing a new cross-layer between the 5G Non-Terrestrial Networks (NTN) system and its mega-constellation transport. The feasibility of the proposed QoS architecture is analyzed through various QoS scenarios, demonstrating its adaptability and implementation potential as extensions of current 5G NTN systems. With this, we provide the foundation for developing a comprehensive 5G NTN QoS solution.

The forthcoming 5G NTN mega constellations will be able to achieve global, ubiquitous connectivity by delivery seamless communication services to a wide range of users and applications, effectively overcoming the geographical and infrastructural limitations inherent to terrestrial networks. Mega-constellations, due to their extensive inter-satellite links provide a multitude of data path options on top of which a robust and reliable end-to-end communication can be established ([14]).

Despite their potential, NTNs present unique challenges for maintaining a similar Quality of Service (QoS) for the communication services compared to current 5G deployments. Unlike their counterpart stable terrestrial backhaul connections, mega-constellations consist of constantly moving satellites, leading to frequently changing links and of variable ground-to-space links impacted by the weather conditions, requiring a new adaptability level. Furthermore, due to their grid-like structure and the limited areas where the data traffic is generated e.g. mostly on the land, populated areas, the data traffic is not uniform in the mega-constellation, being aggregated to a limited set of ground stations, resulting in potential congestions of some links while others remain not utilized ([15]).

Mechanisms designed for assuring QoS over traditional optical based transport networks are fully optimized for static topologies, struggle to adapt to the mobility and variability inherent in NTNs ([16]). Instead, NTNs resemble complex mesh networks with predictable satellite positions, requiring a dynamic approach to manage QoS for the established sessions, by adapting the transport network and the 5G data paths to the momentary conditions.

A fully deterministic solution where the resources are separately allocated for each bearer across the system is not scalable as it requires an intensive routing signaling and data path selection adapted for the very large number of users. This is particularly problematic as it necessitates an intensive communication with the 5G control plane to be able to deploy the end-to-end data paths and to reselect them, highly extending the functionality which should be maintained in the limited resources of the intermediary space nodes. Similarly, an integrated services approach proves too static to effectively account for the specific topology changes,

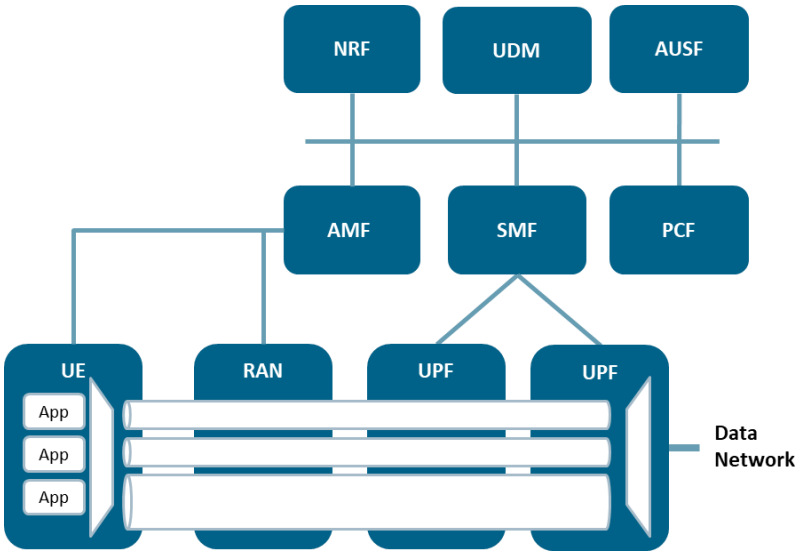

Figure 3-1: 5G Network Architecture and QoS Concept

often resulting in congestion of links, especially when feeder link capacity is adversely impacted by weather conditions ([17]). In such scenarios, rerouting data traffic across multiple potential paths is an obvious solution.

To address these challenges, we propose a comprehensive approach to ensure end-to-end QoS in 5G NTNs through a novel dynamic QoS assurance mechanism tailored for mega-constellations. It involves adapting the routing level within the mega-constellation and the 5G service path based on momentary network conditions. Additionally, we propose a new cross-layer interface between the 5G NTN system and its mega-constellation transport layer to enhance QoS management.

The functionality of the proposed solution is specified starting from a generic mega-constellation model that enables a 5G NTN service on top. This model gives us the possibility to assess the necessity and the complexity of the additional features introduced, underlining a path for the further implementation of the solution.

Moreover, we compare our dynamic QoS solution to a classic QoS approach as considered in the current 3GPP standard. The proposed solution is evaluated against a set of significant QoS scenarios, demonstrating its adaptability and potential for implementation as extensions of current 5G NTN systems. This comparison underlines the superior flexibility and efficiency of our approach in maintaining the QoS across dynamic and variable network conditions.

With this, we provide the foundation for developing a comprehensive 5G NTN QoS solution, capable of integrating space and terrestrial subsystems in order to improve overall system performance and usability.

3.2 BACKGROUND

In the 5G networks, the 5G User Equipment (UE) is connecting though the 5G Radio Access Network (RAN) and a set of User Plane Functions (UPFs) to the data network as illustrated in Figure 3-1. To be able to establish this data path, the core network includes a set of dedicated network functions, the main ones being the Access and Mobility Function (AMF) handling the authentication, authorization, radio access control and mobility management, the Session Management Function (SMF) handling the selection of the data path and the establishment of the data bearers in the core network, the Policy Control Function (PCF) providing policy based decisions on the QoS reservations for the data bearers, the Authentication Server Function

(AUSF) providing authentication algorithms and Network Repository Function (NRF) enabling the selection of network functions ([18]) .

Within the 5G network, the establishment of a session with dedicated QoS levels is triggered by the UE initiating a Packet Data Unit (PDU) session establishment request to the 5G core network, specifying the type of service and the traffic requirements. Based on this request, and on the consulting the access control policies of the UE from the PCF, the SMF will decide by selecting the appropriate UPFs representing the data path in the 5G system, allocating the UE IP addresses if needed or not yet done and applying the QoS rules based on the service needs by enforcing them on the selected UPFs as well as through the AMF to the 5G RAN.

QoS is realized through dynamic resource allocation, packet scheduling, traffic shaping and prioritization mechanisms enforced by the 5G RAN for the radio part as well as by the UPFs on the data path, with a critical role of the last UPF towards the data network on which the downlink flows are shaped to fit the allocated resources.

In a typical 5G deployment, the transport network is fixed and provisioned with a stable level of resources, connecting the RAN to the data network through the core network. Being usually realized with optical fibres or point-to-point wireless links, the transport network has a constant capacity, being properly dimensioned to be able to transport all the data traffic in the 5G system, the bottlenecks being considered only for the shared 5G RAN part of the end-to-end connection.

However, this is not the case for a mega-constellation where the transport resources provided to support the bearers do not have a directly a guaranteed level of resources, being composed of multiple inter-satellite link segments and at least a feeder link influenced by the weather. In this situation, optimizing the QoS according to the transport resources and the optimization of the transport according to the QoS service requirements should be considered, as underlined in the next requirements section.

3.3 REQUIREMENTS

Regardless of where the 5G RAN and core network UPFs are placed, the data path of a 5G NTN bearer will be composed of the following segments: a 5G user link with resources reserved and managed by the 5G RAN, potentially multiple inter-satellite links (ISLs) spanning one or more orbital planes within the mega-constellation, high capacity free-space optical links enabling data exchanges between the user and the feeder link satellites, an optical or high-capacity radio feeder link between the feeder link satellite and a ground station, and potentially multiple dedicated/dependable or undependable terrestrial networks interconnecting the ground stations between each other and linking them to the internet where the core network may be placed.

Over these links an end-to-end connection is realized between the UE and the UPFs which should provide the expected QoS in terms of guaranteed bandwidth, acceptable packet loss rate and delay budget. A simplified example is illustrated in Figure 3-2.

While most current literature concentrates on the connectivity establishment and resources allocation of the new 5G NTN link, for a complete end-to-end data path it is essential to select the appropriate ISLs and optimal feeder link at any given moment to provide a coherent end-to-end QoS.

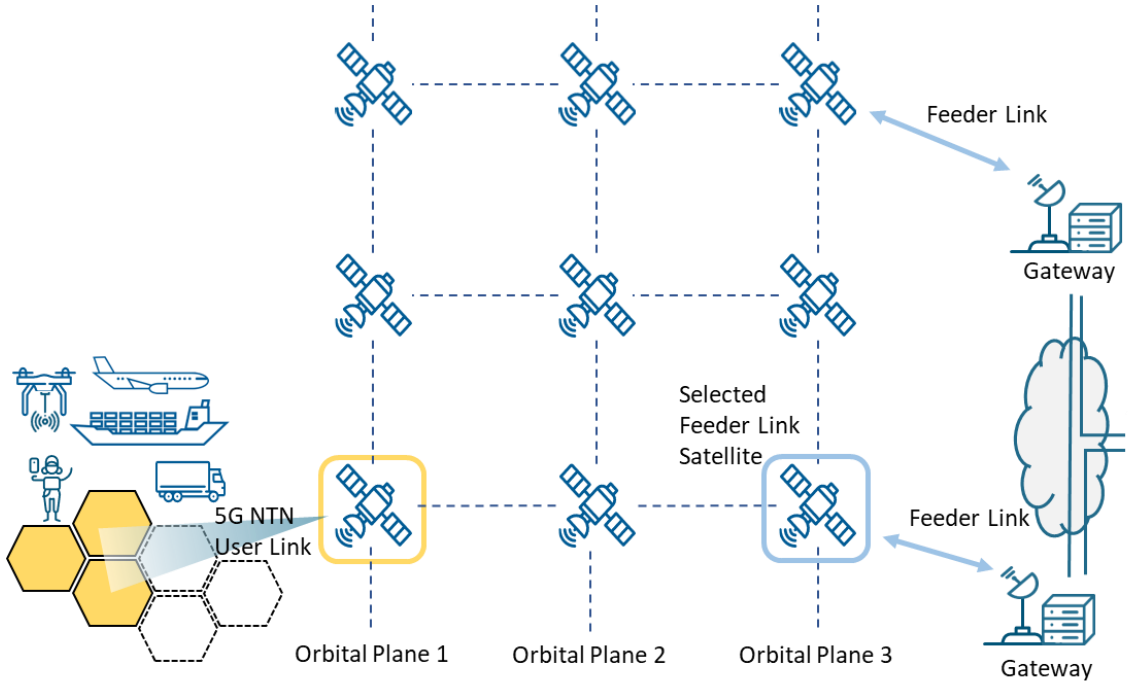

Figure 3-2: End-to-End QoS in a Mega-Constellation

A mega-constellation network has the following characteristics which should be considered when developing a QoS solution. The ISL are usually implemented using free space optical links having a very high capacity and being very stable during the operation times. However, the delay of these links is significant due to the large distances between satellites making an immediate notification of all the nodes not possible. This contrasts with the user and the feeder links which have reduced capacity however relatively short delays when satellites are deployed in low Earth orbits.

The feeder link is particularly influenced by the weather conditions: optical links can be interrupted by clouds, while storms can drastically reduce the capacity of radio links to ensure the data reaches its destination. Thus, QoS depends on an external factor – the weather – which influences a short duration low-capacity link in the system. The most suitable feeder link should be selected based on real-time weather conditions and user requirements with less considerations on the already high capacity available on the ISLs needed to reach the specific feeder satellite.

Once the feeder link is selected, the appropriate ISLs must also be chosen, considering the concurrent data traffic that may cause congestion on some of the nodes, especially towards the feeder link satellite when the overall feeder links number and capacity is reduced. This selection process ensures that the end-to-end data path maintains the required QoS by dynamically adapting to current conditions and data traffic patterns.

3.4 DYNAMIC END-TO-END QOS ASSURANCE MECHANISM

To ensure the end-to-end QoS in the dynamic mega-constellation system, we propose a three-layer adaptation of the data path communication to fit different conditions as illustrated in Figure 3-3:

(1) Adapting of routing at mega-constellation level – after the UPFs are selected for a specific data session, the session remains fixed between the specific RAN and UPF IP

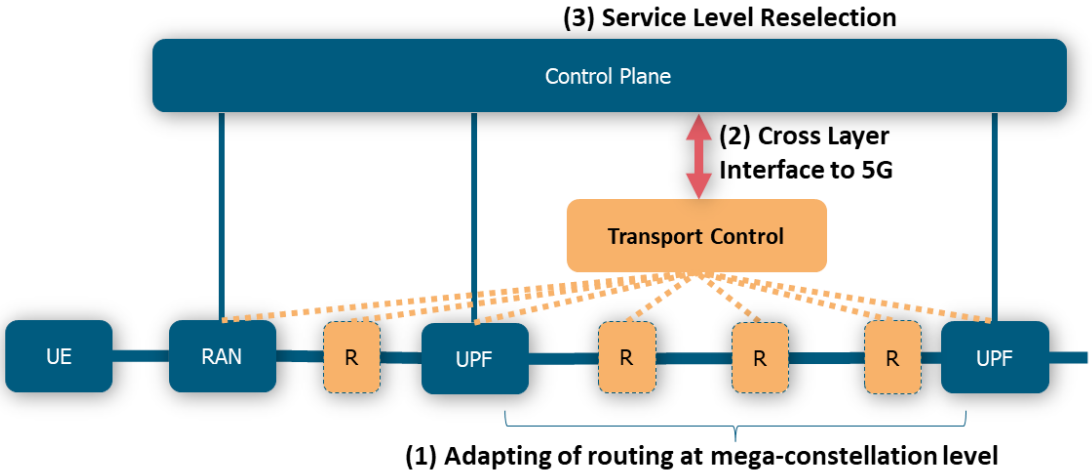

Figure 3-3: Dynamic End-to-End QoS Assurance Mechanism

addresses. However, as illustrated in Figure 3-2, there are multiple routing options between the source and the destination. To optimize this, we propose a new congestion awareness mechanism designed to be implemented on top of the mega-constellation routing and enabling cost-effective and appropriate route of data packets.

(2) Cross layer interface to 5G – the 5G service layer and the routing layer can exchange information on the expected capabilities between different network nodes. During a session establishment, it would be beneficial to receive a notification if a new session can be served at transport level and potentially a data path for it. However, this may not scale well due to the potential very large number of sessions to be signalled. Additionally, the 5G layer could receive network congestion notifications to trigger the allocation of fewer resources to the users or to prioritize specific data traffic.

(3) 5G Service level re-selection – the 5G layer is able to reselect some of the UPFs on the data path to avoid network congestions at this level. This is especially important when a new feeder link is selected as it can redirect data traffic to be offloaded at another ground station, thereby also adapting the end-to-end service to the change in the data path.

Although the system could function without one of the above proposed layers, the absence would drastically impact the end-to-end service quality and it could lead to unwanted situations where the network cannot serve devices as expected even when having enough available resources due to the inflexibility in the data path leading to longer delays and packet loss or dropping.

3.5 5G NTN END-TO-END QOS MANAGEMENT

In this section, we present the selected options for implementing the dynamic end-to-end QoS followed in the next section by an assessment and a feasibility of implementation study.

3.5.1 Adapting of routing at mega-constellation level

We assume that the system already includes a mega-constellation routing system that accounts for the multiple data paths possible, such as the ones published in [19] or [20].

On top of these routes, 5G sessions are established having fixed RAN and UPF points, representing the end-points for each segment which has to be routed.

To avoid reserving resources at each routing element on the data path, we propose a loose semantic routing level to be added on top of the routing solution in order to manage congestions across different links. The semantic element at each node enables it to select if needed alternative, less optimal data paths to forward the data traffic if the selected path does not have enough resources. Although this approach is not fully optimal, it offers significant advantages in terms of reduced complexity of processing within the nodes compared to complete data path signalling.

In normal low load situations over the different links, no action has to be executed, drastically reducing the need to have an active QoS mechanism. Congested times are expected to be very few and short, due to the proper dimensioning of the system. As such the QoS mechanisms here presented have a minimal impact on the overall system.

While the resources of the inter-satellite links and of the nodes remain constant over time and the data traffic can be statically engineered to prevent node overload, feeder links are dependent on the external factor of weather conditions and can become very fast and unpredictably bottlenecks. While on the terrestrial side, data traffic will be redirected to a new ground station using intra-operator routing on the space nodes a new low resource consuming mechanism has to be deployed. For the feeder downlink, we present three congestion avoidance mechanisms as illustrated in Figure 3-4:

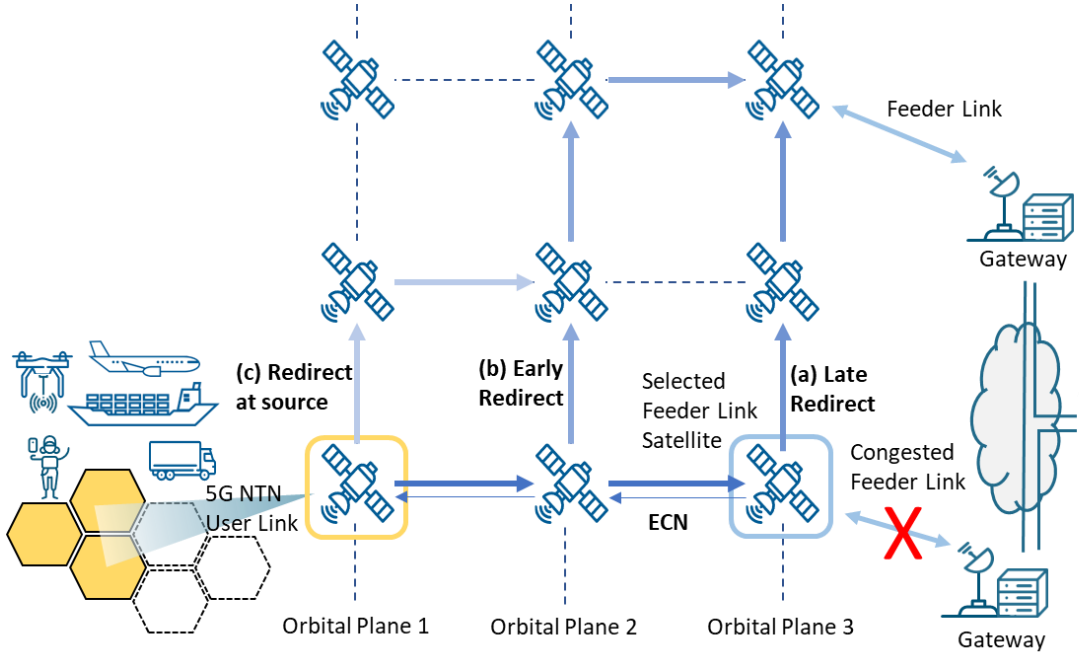

Figure 3-4: Semantic Congestion Management Layer

(i) Late redirect – This mechanism reroutes the data traffic at the last possible point in the data path, the feeder link satellite. When detecting congestion or an interruption of the feeder link, the feeder link satellite attempts to reroute the data traffic to reach the UPF by changing the routing path to another feeder link. The significant advantage of late redirect is its ability to handle sudden interruptions efficiently without requiring extensive changes in the earlier stages of the data path. It reduces the burden on the intermediate nodes and can be implemented with static profiles for load distribution. By managing rerouting at the final point, late redirect ensures that any immediate disruptions are addressed locally, reducing overall network congestion and maintaining service continuity. Note that the capacity of the ISL is very large compared to the user traffic, being based on free space optical links, this redirect very rarely resulting into congestions, which will be fixed with the mechanisms presented underneath. However, data traffic could be sub-optimally routed as other shorter paths might be available that do not include the feeder link satellite.

(ii) Early Redirect – This proactive rerouting mechanism addresses potential disruptions in data traffic as close to the source as possible. When feeder links become available or are interrupted, notifications are sent towards the source nodes of the data packets such as IETF Explicit Congestion Notifications (ECNs) ([21]). These notifications enable the redirection of data traffic before reaching the congested or disrupted link. For instance, if a feeder link failure is detected, the notification is propagated to other satellites in the constellation, enabling them to reroute the downlink data traffic at an intermediary node rather than at the last node. This method optimizes the data paths and minimizes the impact on the entire constellation but requires additional congestion signalling.

(iii) Redirect at source – a special case of early redirect is when the notifications are propagated to all the source nodes which have data traffic for a specific feeder link. In this case, the signalling has multiple intermediary nodes, however it provides fully optimized data paths to the given conditions by notifying only the satellites including active RAN components, a reduced number compared to the complete constellation. Also, the intermediary nodes do not have to have any semantic role, this being taken by the source node.

These mechanisms can also be used in cases of temporary congestions of the inter-satellite links due to misrouting or node failure, signalling unavailability and redirecting the data traffic to avoid the problematic network areas. This ensures end-to-end reliability, another critical QoS characteristic.

With these mechanisms it is unnecessary to fully reserve resources across the data path, unlike QoS Integrated Services ([22]) which were deemed unscalable for a large number of data flows. Also, compared to Differentiated Services (DiffServ) ([23]), this approach utilizes the multiple data path options available to adapt to network conditions without dropping the data packets that cannot be transported through the selected links, thereby providing an overall better service. However, it is to be noted that a fully optimized solution will not be achieved as the nodes are aware of their own routing issues or of the messages received from the feeder link satellites.

Until now, our considerations have been based on a static topology. However, this topology is maintained only for a limited period. Due to the mobility of satellites in lower Earth orbits, at predetermined times, the serving user and feeder link satellites must handover to a next satellite of the same or a different orbit. We assume that QoS semantic information will also be handed over to the next satellite. This implies that congestion information for the feeder links or user links will be transferred along with the 5G context. In this manner, a designated satellite providing user or feeder link services in a specific area will possess the same information as its predecessor, ensuring that QoS awareness remains local to that specific area.

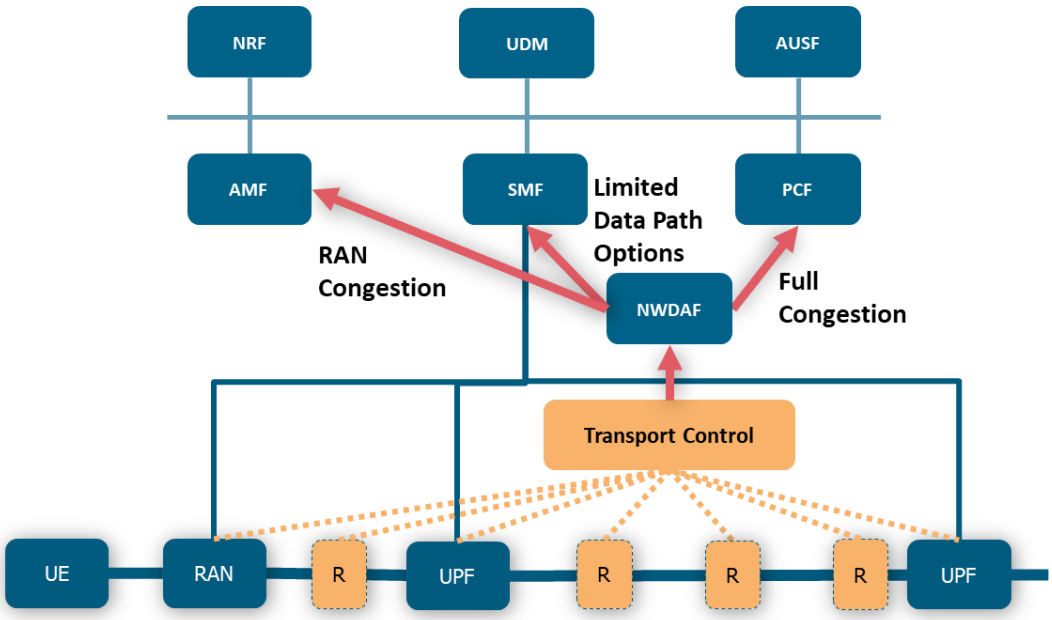

Figure 3-5: Cross-Layer Notifications

3.5.2 Cross-Layer Interface between Routing and 5G

Not in all situations can the routing be adapted transparently for the 5G network. Specifically, in very bad weather conditions, the communication over multiple feeder links may be disturbed or even interrupted. Because of this, the overall capacity of the system to establish end-to-end connections can be drastically reduced. In order not to establish sessions when there are not enough resources available, a new mechanism should be considered to transmit such congestion notification events from the routing level to the packet core level. To achieve this, a cross-layer communication between the network control at the routing level is needed, to transmit to the 5G core PCF a congestion notification. This can be easily implemented as an event transmitted through the Network Data Analytics Function (NWDAF) ([24]) which is able to analyse the state of the network and adapt the communication. PCF is the target entity as it can make decisions based on the subscription profile how much resources to allocate to the different bearers depending on the overall available resources. The architecture for this proposed cross-layer is illustrated in Figure 3-5.

More specific congestions can also be considered such as naming certain ground stations which cannot be anymore reached due to the lack of feeder links even when redirecting sessions. With this type of notifications, the SMF may adapt its session establishment decisions to select other data paths for the sessions. With this notification, new sessions can be established to avoid congested feeder links. This is not only useful in case of bad weather conditions but also in the normal case to fit the communication of the 5G system appropriately in case there are other parallel systems which may occupy resources such as earth observation.

Apart from the QoS along the data path, it is worth noting that also the RAN may benefit from information about the impact of the weather in a given area to allocate lower resources according to the user link momentary capacity.

3.5.3 5G adapted to the transport conditions

Based on the cross-layer notifications, the active sessions may also be adapted to fit more sessions even with reduced resources. This includes the changing of the serving UPF during

specific congestion times triggered by the SMF as well as the changing of the allocated resources for a session within the PCF during a congestion time, as described in the previous subsection.

3.6 FEASIBILITY ASSESSMENT

In this section we demonstrate the feasibility of the proposed solution through a set of scenarios, showcasing how the system would react at different QoS capacity modifications and the implications of the overall system in comparison with a static solution without the here introduced methods as well as the assessing the implementation complexity of such a solution.

3.6.1 Scenarios Assessment

Although not a comprehensive list of scenarios, it will give a general impression of the end-to-end QoS advantages is presented.

Scenario 1) Normal functioning – in the case of normal functioning where all the links are uncongested, the proposed methods are not introducing any additional overhead. This is especially important as we expect that most of the time the system will function in this state.

Scenario 2) Single feeder link congestion – in this situation, the data packets are re-directed at the space routing level towards another feeder link assuring that the data path is receiving enough resources, however, with potential larger delay. Signalling this information to the SMF enables the selection of another 5G data path for the new sessions, with this being able to adapt to the momentary conditions. In the current static system, this would not be possible, resulting into packet loss and communication interruptions.

Scenario 3) Single feeder link decongestion – the same mechanism is used to signal that a link becomes again available for usage, through this enabling the redirection of the suboptimal routed data traffic back to the expected data path. Signalling this event to the SMF will enable the adaptation of the new sessions to the available links.

Scenario 4) Oscillating feeder link congestions - in case the capacity of a feeder link is very fast oscillating, in order to maintain the system stable, after two oscillations, the link should be considered at the worse capacity and the system adapted for this level, through this avoiding unwanted ping-pong decision effects and potential redirects of the data traffic between two feeder link satellites.

Scenario 5) Multiple feeder link congestions – in very bad weather conditions it can happen that multiple feeder links will become congested at the same time. In these situations, very few end-to-end resources remain available. A signalling to the 5G system is needed to make service level decisions which sessions have priority and to reduce the end-to-end reserved resources to the level available.

Please note that as the user links and the feeder link can be generally co-located in the same area e.g., within the continents, these extreme bad weather effects will also impact the user links, reducing also significantly the capacity that can be communicated with the user device.

Scenario 6) Space node or link congestion – due to side effects of the routing it may happen that for very short duration of time, due to the very high capacity of the ISLs, some of the space links become congested. In these situations, the congestion notifications enable the re-routing of the data packets to less optimal routes through the constellation.

Scenario 7) Node or link failure – in the very unlikely cases of node or link failure within the mega-constellation, the same mechanism as the one for avoiding congestions is used. This way, the reliability of the system does not require an additional mechanism, simplifying the overall routing extensions.

As expressed in this section, the proposed solution is relatively less complex than implementing a comprehensive controlled routing solution and at the same time minimally impacts the overall system while being able to adapt the system to the network conditions.

3.6.2 Implementation Feasibility Assessment

To implement this solution within a 5G NTN system, the following additional functional elements must be added to the system at the routing and 5G core network level.

The implementation is highly dependent on the routing solution selected for the mega-constellation. It is assumed that the multiple data paths which can be established between a user and a feeder link satellite are available to be selected at any intermediary node. Specifically, all the different routing options within the satellite grid-like network should be available within the nodes, enabling the load balancing across the multiple links.

The ECN mechanism was selected because it is already considered for such scenarios in the terrestrial networks. ECN has the advantage of using unused fields in the IP header eliminating the need for an additional explicit signalling protocol. Additionally, being specifically located in the IP header, it can be easily identified from the packets, significantly reducing the signalling processing in the intermediary nodes. Therefore, no mechanism is needed.

Furthermore, we assume that a single logical SMF and a single logical PCF are deployed in a given area served by multiple user and feeder links. This is in line with the current 5G NTN developments, given that the number of users using 5G NTN is significantly lower compared to a terrestrial mobile network operator, allowing for a centralized behaviour. This simplifies the cross-layer interactions, as there is a single terminating point within the 5G system for different events.

However, lower-level events have to be first gathered at a command centre. Since the command centre is located on the ground, there is a long distance between the constellation nodes and the command centre, causing a significant delay in event notifications. Despite this delay, the system can still adapt in due time, as these notifications are not so late as to be ineffective. Changes at the 5G NTN level should happen only in exceptional situations as they involve update procedures for both the handover of the data paths as well as for the QoS adaptations. Exceptional situations should be notified to the command centre as they are significant for constellation management.

Regarding the implementation of the interface itself, since only event notifications are coming from the system, the NDWAF was chosen. NDWAF is taking data from the system monitoring which can be extended with such link and congestion information received from the command centre.

With this, we have provided a minimal proof that the proposed solution can be easily implemented on top of the existing system without requiring changes to the current standards, only functional additions. This makes our proposed solution viable for further considerations on the implementing of a 5G NTN solution.

Furthermore, as the additions related to the routing are minimal, as well as the processing in the intermediary nodes of the congestion information, it makes this solution highly attractive to be implemented on top of the low-level networking processing resources of the space nodes. Additionally, by using mechanisms already standardized and tested in the terrestrial network environments, a basic trust that our proposed solution will function optimally and would be easy to deploy is achieved, especially because the different elements were already tested into live terrestrial networks.

A cross-layer interface between the transport and the packet core control involves the development of a new exchange interface which may cause unintended side effects on the other parts of the when improperly implemented. However, instead of proposing a complete cross-layer interface, we designed an event notification one from the lower to the higher layers

within the system, aligning with the developments of the NWDAF functionality in 3GPP. This approach ensures that the 5G NTN will only react to the network conditions and accordingly adapt without influencing the behaviour of the routing layer. By eliminating the feedback influence, no means remain for side effects like ping-pong decisions.

3.7 CONCLUSIONS AND FURTHER WORK

In this chapter, we have presented a comprehensive approach for ensuring end-to-end QoS in 5G NTNs through the usage of dynamic QoS assurance mechanisms tailored for mega-constellations at three different layers: QoS aware semantic adaptation of the mega-constellation routing, a cross-layer interface between transport control and 5G system and a service level reselection within the 5G core network. This multi-layer approach enables the addressing of the different specific system requirements at the appropriate levels, bringing a minimal additional complexity while assuring the expected reliability and high-performance connectivity service.

The implementation of the proposed mechanisms requires additional functional elements at both the routing and the 5G core network levels. For this, we proposed to use the semantic mechanism provided by the Explicit Congestion Notification (ECN) mechanism to handle the dynamic transport conditions as well as the interaction through the NDWAF between the transport and the 5G packet core, following already existing mechanism adapted for the specific environment.

Our assessment demonstrates the feasibility and potential of our proposed QoS solution through various scenarios, highlighting its adaptability and simplicity as extensions of the mega-constellation based 5G system, providing the basis for a comprehensive QoS solution.

In the further steps we plan to integrate the congestion mechanism as part of our mega-constellation routing activities ([25]) as well as to integrate advanced algorithms for the redirection able to learn from past experience in order to better adapt to the specific situations.

4 CONCLUSIONS

The deliverable presents a first exploration of AI-driven NTN networking and QoS management strategies to optimize 5G NTNs. It outlines the development of AI-based framework for network management and a dynamic end-to-end QoS assurance mechanism detailing the architecture in D3.2 and setting the base for the further system definition.

Specifically, this deliverable introduced the following novel concepts:

- **AI Data-Driven Network Management:** Development of AI-based optimizations for network management, including resource allocation, network slicing, and multi-connectivity solutions.
- **End-to-End AI-oriented Network Architecture:** Proposal of two architectural schemes for direct and indirect connectivity in NTNs, leveraging AI for dynamic resource management.
- **Controller Framework:** Introduction of a controller managing network elements through traditional optimization methods and AI-based solutions, handling tasks such as RAN function deployment and network slicing.
- **QoS Management Mechanisms:** Introduction of a dynamic end-to-end QoS assurance mechanism tailored for mega-constellations, including adaptation of routing at the mega-constellation level and implementation of a cross-layer interface between the 5G service layer and the routing layer.

By laying this foundation work, this deliverable provides a first time AI framework for NTNs as well as a comprehensive end-to-end QoS management structure. These have to be proven through their detailed application within the final system as well as through the development of specific optimization algorithms validating the frameworks and providing their added value to the end-to-end NTN system. These next steps will be reported in the follow-up D5.4 deliverable.

5 REFERENCES

[1] O-RAN Architecture Description 11.0, Release 3, February 2024

[2] J. Moysen and L. Giupponi, “From 4g to 5g: Self-organized network management meets machine learning,” Computer Communications, vol. 129, pp. 248–268, 2018.

[3] X. Lin, “An overview of 5g advanced evolution in 3gpp release 18,” IEEE Communications Standards Magazine, vol. 6, no. 3, pp. 77–83, 2022.

[4] ETSI. "Experiential Networked Intelligence (ENI); System Architecture", ETSI GS ENI 005 V3.1.1 (2023-06).

[5] ETSI. “Zero-touch network and Service Management (ZSM); Cross domain E2E service lifecycle management”, ETSI GS ZSM 008 V1.1.1 (2022-07).

[6] B. Barritt and W. Eddy, "Service Management & Orchestration of 5G and 6G Non-Terrestrial Networks," 2022 IEEE Aerospace Conference (AERO), Big Sky, MT, USA, 2022, pp. 1-11, doi: 10.1109/AERO53065.2022.9843390.

[7] Francisco Muro, Eduardo Baena, Tomaso De Cola, et al. A 5G NTN Emulation Platform for VNF Orchestration: Design, Development, and Evaluation. TechRxiv. January 26, 2024. DOI: 10.22541/au.170628474.46948465/v1.

[8] Nguyen, Cong T., et al. "Emerging Technologies for 6G Non-Terrestrial-Networks: From Academia to Industrial Applications." arXiv preprint arXiv:2403.07763 (2024).

[9] Polese, Michele, et al. "Understanding O-RAN: Architecture, interfaces, algorithms, security, and research challenges." IEEE Communications Surveys & Tutorials 25.2 (2023): 1376-1411.

[10] Palattella, Maria Rita, et al. "5G smart connectivity platform for ubiquitous and automated innovative services." 2021 IEEE 32nd Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC). IEEE, 2021.

[11] Ishizu, Kentaro, et al. "Architecture for Beyond 5G Services Enabling Cross-Industry Orchestration." IEICE Transactions on Communications 106.12 (2023): 1303-1312.

[12] Darwish, Tasneem, et al. "LEO satellites in 5G and beyond networks: A review from a standardization perspective." IEEE Access 10 (2022): 35040-35060.

[13] G. M. Almeida et al., "RIC-O: Efficient Placement of a Disaggregated and Distributed RAN Intelligent Controller With Dynamic Clustering of Radio Nodes," in IEEE Journal on Selected Areas in Communications, vol. 42, no. 2, pp. 446-459, Feb. 2024, doi: 10.1109/JSAC.2023.3336159.

[14] Höyhtyä, Marko, et al. "5G and beyond for new space: Vision and research challenges." Advances in Communications Satellite Systems. Proceedings of the 37th International Communications Satellite Systems Conference (ICSSC-2019). IET, 2019.

[15] Ravishankar, Channasandra, et al. "Next-generation global satellite system with mega-constellations." International journal of satellite communications and networking 39.1 (2021): 6-28.

[16] 3GP TS23.503, “Policy and charging control framework for the 5G System (5GS); Stage 2”, version 17.6.0, https://www.3gpp.org;

[17] Viasat. (2023). How Weather Can Affect Satellite Communications. Viasat News. https://news.viasat.com/blog/scn/how-weather-can-affect-satellite-communications, last visited on 25.05.2024;

[18] 3GPP TS23.501, “5G System Overview”, v18.5.0, March 2024, portal.3GPP.org;

[19] Elbehiry, Eman Adel, et al. "Survey on Routing Algorithms for LEO Constellations Network." Fayoum University Journal of Engineering 7.2 (2024): 89-99;

[20] Li, Rui, et al. "LEO Mega-Constellations Routing Algorithm Based on Area Segmentation." 2023 IEEE Wireless Communications and Networking Conference (WCNC). IEEE, 2023;

[21] De Schepper, K. "RFC 9331: The Explicit Congestion Notification (ECN) Protocol for Low Latency, Low Loss, and Scalable Throughput (L4S)." (2023);

[22] Farrel, A., ed. "RFC 9522: Overview and Principles of Internet Traffic Engineering.", Internet Engineering Task Force, 2024;

[23] Black, D. L., & Jones, P. Differentiated Services (Diffserv) and Real-Time Communication. Internet Engineering Task Force. https://datatracker.ietf.org/doc/html/rfc7657, 2015;

[24] 3GPP TS 29.554, “Background Data Transfer Policy Control Service; Stage 3”. Version 18.2.0, December 2023;

[25] Buhr H. et al., SENDIT: A Satellite Network Routing Digital Twin Solution for Mega-Constellations” 40th International Communications Satellite Systems Conference (ICSSC), October 2023;

[26] 3GPP TS 23.288 Architecture enhancements for 5G System (5GS) to support network data analytics services

[27] 3GPP TS 28.530 Management and orchestration; Concepts, use cases and requirements (Release 18)