6GBLUR-SMART: The Blurring RAN smart decision-making algorithms for efficient end-to-end resource management

SMART

Grant No. TSI-063000-2021-56

E3: Final RAN architecture design

Financiado por la Unión Europea NextGenerationEU
Plan de Recuperación, Transformación y Resiliencia
UNICO I+D
GOBIERNO DE ESPAÑA
cttc Centre Tecnològic de Telecomunicacions de Catalunya

Abstract

This report describes the final 6GBLUR RAN and end-to-end architecture design, as initially developed by CTTC, and later refined by considering the inputs from the operator, vendors, SME and university.

Document properties

Document number	E3
Document title	6GBLUR architecture and Key Concepts
Document responsible	Josep Mangues-Bafalluy (CTTC)
Document editor	Albert Bel, Sergio Barrachina-Muñoz, Josep Mangues-Bafalluy, Katerina Koutlia (CTTC)
Authors	Jordi Baranda (CTTC), Katerina Koutlia (CTTC), Sergio Barrachina (CTTC), Josep Mangues-Bafalluy (CTTC), Albert Bel (CTTC), Nikos Bartzoudis (CTTC), Rasoul Nikbakht (CTTC), Ana Larrañaga (CTTC), Sandra Lagén (CTTC), Jose Rubio Fernandez (CTTC), Antonio Roman (CTTC), Engin Zeydan (CTTC), Miquel Payaró (CTTC), Biljana Bojovic (CTTC), Manuel Requena (CTTC), Manuel Lorenzo (Ericsson), Saravanan Kalimuthu (Ericsson), Gokhan Turhan (Ericsson), Marc Molla (Ericsson), Suraj Rajan Sirwani (Ericsson), Luis Miguel Contreras Murillo (Telefonica), Javier Velázquez Martínez (Telefonica), Maria Teresa Aparicio Peña (Telefonica), Elena Serna Santiago (Telefonica), Óscar Gil Lucía (Telefonica), David Gregoratti (SRS), Ismael Gomez (SRS), Carles Navarro Manchon (KEY), Francisco Garcia (KEY), Jose Ramon Suarez (KEY), Germán Corrales Madueño (KEY)
Target dissemination level	Confidential
Status of the document	Final
Version	1.1
Delivery date	30 June 2024
Actual delivery date	30 June 2024

Disclaimer

This document has been produced in the context of the 6G BLUR Project. The research leading to these results has received funding from the Ministerio Español de Asuntos Económicos y Transformación Digital (MINECO), under grant TSI-063000-2021-56.

Contents

List of Figures

List of Tables

List of Acronyms

ADC: Analog to Digital Converter

AF: Application Function

AI: Artificial Intelligence

API: Application Programming Interface

APU: Application processing Unit

AR: Augmented Reality

B5G: Beyond 5G

BH: BackHaul

BWP: Bandwidth Part

CA: Carrier Aggregation

CAPEX: Capital Expenditures

CC: Component Carrier

CNF: Core Network Function

CoMP: Coordinated Multi-Point

COTS: Commercial-Off-The-Shelf

CPRI: Common Public Radio Interface

CRF: Crest Reduction Factor

CSP: Communications Service Provider

CU: Centralised Unit

DAC: Digital to Analog Converter

DC: DataCenter

DCI: Downlink Control Information

DDC: Digital Down-conversion

DL: Downlink

DNN: Data Network Name

DPD: Digital Pre-distortion

DRL: Deep Reinforcement Learning

DT: Digital Twin

DU: Distributed Unit

DUC: Digital Up-conversion

eCPRI: enhanced Common Public Radio Interface

E2E: End to End

eMBB: Enhanced Mobile Broadband

FDD: Frequency Division Duplexing

FH: FrontHaul

FFT: Fast Fourier Transform

gNB: next Generation Node B (5G node B)

GPU: Graphics Processing Unit

HARQ: Hybrid Automatic Repeat Request

HW: Hardware

ICIC: Inter-Cell Interference Coordination (ICIC)

KPI: Key Performance Indicator

MCS: Modulation and Coding Scheme

MH: MidHaul

MIMO: Multiple-Input, Multiple-Output

ML: Machine Learning

MNO: Mobile Network Operator

MPSoC: Multi-processing System-on-chip

MR: Mixed Reality

NDT: Network Digital Twin

NEF: Network Exposure Function (3GPP)

NFV: Network Function Virtualization

nGRG: Next Generation Research Group

NIC: Network Interface Card

NN: Neural Network

NPN: Non-Public Network (3GPP)

NSSAI: Network Slice Selection Assistance Information

NTN: Non-Terrestrial Network

NWDAF: Network Data Analytics Function

OAI: Open Air Interface

OAM: Operations, Administration and Maintenance

OFDMA: Orthogonal Frequency-Division Multiple Access

OPEX: Operational Expenditures

OSC: O-RAN Software Community

PDV: Packet Delay Variation

PLR: Packet Loss Rate

PN: Public Network (3gpp)

PNI-NPN: Public Network Integrated- Non-Public Network (3gpp)

PoC: Proof of Concept

PON: Passive Optical Network

PRB: Physical Resource Block

QoS: Quality of Service

QUIC: Quick User Datagram Protocol

RAN: Radio Access Network

RF: Radio Frequency

RFSoC: Radio Frequency System-on-Chip

RIC: Radio Intelligent Controller

RL: Reinforcement Learning

RLC: Radio Link Control

RTT: Round-Trip Time

RTU: Real-time processing unit

RU: Radio Unit

SBA: Service-based Architecture

SDN: Software Defined Networking

SoA: State-of-the-art

SLA: Service Level Agreement

SLO: Service Level Objectives

SMF: Session Management Function

SW: Software

TDD: Time Division Duplex

TDMA: Time-Division Multiple Access

TR: Technical Report

TTI: Transmission Time Interval

UE: User Equipment

UL: Uplink

UPF: User Plane Function

URLLC: Ultra-reliable Low-Latency Communication

VPN: Virtual Private Network

VR: Virtual Reality

XR: eXtended Reality

ZSM: Zero-Touch Service Management

Executive Summary and Key Contributions

The *Blurring RAN (6GBLUR)* is a coordinated project funded by UNICO MINECO program. 6GBLUR consists of two subprojects, 6GBLUR-SMART (smart decision-making algorithms for efficient end-to-end resource management) and 6GBLUR-JOINT (joint RAN and transport network control/orchestration mechanisms).

As part of 6GBLUR, 6GBLUR-**SMART** aims (i) to design and to validate a RAN control architecture adapting to all the required decision timescales, from policy definition to scheduling, and (ii) to design and validate smart algorithms for efficient end-to-end resource management, including decision-making at all timescales, with real-time and non-real time decisions.

In this deliverable (*E3: Final RAN and E2E Architecture Design*), we present the final architecture as envisioned by the entire consortium and elaborate on the key concepts (KCs). The **architecture**, initially defined in our previous deliverable (*E1: Project Plan*), has been further consolidated in this document. This refinement encompasses the integration of conventional mobile network architectures with various advancements, including streamlining O-RAN through explainable AI with LLM integration and tailored QoS marking strategies for fronthaul traffic, aiming for increased generality.

Furthermore, this deliverable provides expanded details on the **KCs** that will serve as the foundation for realizing 6GBLUR's use cases (UCs) and proof-of-concepts (PoCs). This includes an in-depth analysis of such key research concepts, their study in existing literature, and the progress 6GBLUR proposes beyond the current state-of-the-art. In particular, the following KCs have been identified in 6GBLUR-SMART:

- **K1.1 CU and DU implementation**: open-source implementation of and ORAN platform of the E2 nodes CU and DU
- **K1.2 RAN orchestration**: focusing on the Open Fronthaul interface and the implementation of strategies to improve its capabilities.
- **K1.3 Dynamic configuration of topologies/architectures for MH distribution**: the modular architecture of O-RAN offers the possibility to exploit the appearance pf the FH and MH concepts and exploit them to optimize the network conditions.
- **K1.4 Micro-orchestration of virtual networks functions at DU-RU level**: Introduction of FPGAs and FPGA-base System-on-Chip to accelerate the DU and RU functions, the high and low-PHY layer functions.
- **K1.5 Dynamic configuration of FH for distributed and centralized architectures**: applying the novel concept of dynamic architectures of the O-RAN functions RU, DU and CU.
- **K2.1 Development and characterization of FH and MH**: implementation of an open-source solution of the FH and MH for O-RAN testbeds.
- **K2.2 Characterization of the FH and MH transport**: characterization of the FH in order to achieve more decentralized architectures in O-RAN.
- **K2.3 Flexible functional split and physical layer**: optimizing the best functional split in O-RAN by enhancing the implementation of them in the ns3 5G-LENA simulator.
- **K2.4 Radio Stack optimization**: optimization of MAC layer protocols on more centralized architectures.
- **K2.5 QoS Management for time-critical immersive applications**: design of effective QoS scheduling algorithms.

- **K2.6 QoS strategies the transport for multi-service aggregation**: characterizing the system behaviour from different QoS strategies for mapping the fronthaul traffic in an O-RAN-based 6G disaggregated mobile network.
- **K3.1 AI/ML Worflows in O-RAN**: leverage advanced ML techniques to enhance the intelligence of RAN, deployed as an xApp.

Finally, let us emphasize that, although some KCs are categorized under 6GBLUR-SMART and others under 6GBLUR-JOINT, there is a strong interrelationship between the two. Concepts explored in one subproject frequently appear in the PoCs of the other subproject, and vice versa. This separation allows for focused research and development while ensuring that ideas and solutions are shared across the project.

1 Introduction

The blurring RAN (6GBLUR) is a project funded by UNICO MINECO program. 6GBLUR consists of two subprojects, 6GBLUR-SMART (smart decision-making algorithms for efficient end-to-end resource management) and 6GBLUR-JOINT (joint RAN and transport network control/orchestration mechanisms).

From a technical perspective, next-generation RANs must overcome numerous challenges, including achieving high spectral efficiency, minimizing power consumption, enabling resource pooling, ensuring scalability, and facilitating cross-layer interworking. To address these challenges, leading standardization bodies like 3GPP and O-RAN are advocating for virtualized architectures. In these architectures, traditional base station functions, historically located near antenna towers, are virtualized and distributed across various logical nodes connected via fronthaul and midhaul links, and further connected to the core network through backhaul links. This virtualization disaggregates baseband processing of individual cells across different physical entities using specific functional splits, while also allowing the consolidation of baseband processing for multiple cells in centralized locations for efficient resource sharing. This dynamic architecture adapts based on network conditions and provider policies, effectively ***blurring*** traditional RAN boundaries.

Equipment vendors and Mobile Network Operators (MNOs) have adopted centralized virtualized RAN architectures as an advanced system paradigm, aiming to reduce both capital and operational expenditures while enhancing energy-efficient data rates in the radio access system. However, deploying such flexible architectures presents five primary challenges: (i) stringent fronthaul capacity and latency requirements, which will be exacerbated in 6G due to wider channel bandwidths, massive MIMO, higher modulation orders, and increased carrier aggregation; (ii) the necessity for a novel RAN control architecture capable of managing the disaggregated mobile network with real-time and non-real-time controllers; (iii) the selection and implementation of potentially multiple and dynamic functional splits across multiple cells, considering shared transport, computing, and baseband processing resources; (iv) variable latency requirements dictated by Key Performance Indicators (KPIs); and (v) establishing trust in virtualized and open networks.

The overarching goal of 6GBLUR is to design an end-to-end architecture with efficient resource management procedures and intelligent control mechanisms for virtualized, adaptive, and integrated mobile network architectures. Managed resources span the distributed and disaggregated mobile network, including spectrum allocation, network capacity, fronthaul resources, baseband processing functions, energy consumption, computing resources, and storage. To effectively manage these resources and orchestrate network operations intelligently in both real-time and non-real-time scenarios, AI/ML processes are leveraged. This approach targets specific UCs such as deploying 6G disaggregated zero-touch mobile networks as a service within O-RAN architectures and non-public networks (NPNs), implementing control mechanisms for optimizing network performance with a focus on fronthaul capacity, centralization strategies, QoS management, flexible functional splits, and digital twins for NPN networks, and developing RAN control architectures and smart algorithms specifically for O-RAN and NPN environments.

As part of 6GBLUR, 6GBLUR-**SMART** aims to (i) design and validate a RAN control architecture that adapts to all required decision timescales, from policy definition to scheduling, and (ii) develop and validate intelligent algorithms for efficient end-to-end resource management, encompassing real-time and non-real-time decision-making across all timescales. To achieve these objectives, multiple UCs and PoCs have been developed to experimentally validate the project's KCs within various frameworks, including the integration

of open RANs and 3GPP-defined non-public networks in an end-to-end architecture. The KCs for WP2 and WP3 were initially discussed during the November face-to-face meeting (2023-11-20/21), where a leader was assigned to each key concept based on their contributions in deliverable *E1: Project Plan*.

This deliverable (*E3: Final RAN and E2E Architecture Design*) provides a comprehensive report on the progress in advancing the state-of-the-art of 6GBLUR-SMART's KCs. It details the final architecture agreed upon by the entire consortium, alongside an in-depth analysis of the various KCs involved in the project and the main advancements to be demonstrated through the UCs and PoCs previously outlined in deliverable E1: Project Plan. The document is organized as follows: Section 1 introduces the scope and motivation of the project; Section 2 presents the final architecture; Sections 3 and 4 discuss 6GBLUR-SMART's KCs related to WP2 and WP3, respectively.

1.1 SMART Key Concepts

The 12 KCs in 6GBLUR-SMART are summarized below. These KCs, under study in both WP2 and WP3, are aligned with the principal objective of the 6GBLUR-SMART project: to efficiently manage a 6G RAN architecture by intelligently orchestrating all critical components.

TABLE 1. SMART KEY CONCEPTS.

Category	KC	Project	WP	Leader
K1. RAN control architecture and orchestration	**K1.1** CU and DU implementation	SMART	2	SRS
	K1.2 RAN Orchestration	SMART	3	CTTC
	K1.3 Dynamic configuration of topologies/architectures for MH distribution	SMART	2	TID
	K1.4 Micro-orchestration of virtual network functions at DU-RU level	SMART	2	CTTC
	K1.5 Dynamic configuration of FH for distributed and centralized architectures	SMART	3	TID
K2. Fronthaul and Midhaul optimization	**K2.1** Development and characterization of FH and MH	SMART	2	SRS
	K2.2 Characterization of the FH and MH transport	SMART	3	TID
	K2.3 Flexible functional split and physical layer	SMART	3	CTTC
	K2.4 Radio Stack optimization	SMART	3	TID
	K2.5 QoS Management for time-critical immersive applications	SMART	3	CTTC
	K2.6 QoS strategies the transport for multi-service aggregation	SMART	3	TID

K3. AI/ML	**K3.1** AI/ML Workflows in O-RAN	SMART	2	CTTC

1.2 Mapping of Smart Key Concepts and PoCs

Following the primary objective of the 6GBLUR-SMART project, the various PoCs developed will focus on optimizing the performance of 6G networks. Table 2 illustrates the relationship between the PoCs and the KCs.

Most of the KCs will be incorporated into UC2PoC1, which aims to characterize and analyse the feasibility of the proposed architecture across different network topologies and scenarios (e.g., low, medium, high-capacity traffic) to support the deployment options explored in UC1. [1] This characterization and analysis will help identify the limits of such network deployments, with the goal of extending theoretical boundaries. This effort will impact the network's capital and operational expenditures by enhancing hardware and software elements and procedures, achieving higher levels of centralization, and reducing the need for human intervention in network operations, all while maintaining the Quality of Service (QoS) for users.

Additionally, the KCs related to two of the PoCs in UC3 are summarized, focusing on smart methods for resource optimization of the RAN in an O-RAN based 6G network.

TABLE 2. MAPPING OF KEY CONCEPTS AND POCS

KC	PoC			
	UC1PoC2	UC2PoC1	UC3 PoC2	UC3PoC3
K1.1 CU and DU implementation	x	x		x
K1.2 RAN Orchestration		x		
K1.3 Dynamic configuration of topologies/architectures for MH distribution		x		
K1.4 Micro-orchestration of virtual network functions at DU-RU level			x	
K1.5 Dynamic configuration of FH for distributed and centralized architectures		x		
K2.1 Development and characterization of FH and MH	x	x		x
K2.2 Characterization of the FH and MH transport		x		
K2.3 Flexible functional split and physical layer		x		

[1] Due to the main concepts worked in 6GBLUR-SMART subproject, this document includes only the analysis of the network deployment of UC1 PoC2. For further details about the analysis of UC1 PoC1, please refer to the 6GBLUR-JOINT E1 document.

K2.4 Radio Stack optimization	x	
K2.5 QoS Management for time-critical immersive applications	x	
K2.6 QoS strategies the transport for multi-service aggregation	x	
K3.1 AI/ML Workflows in O-RAN		x

2 Final RAN and end-to-end architecture

The final architecture has not suffered modifications from the one described in deliverable E1. The primary objective is to construct a 6G disaggregated mobile network that integrates the virtualized RAN and core components. This objective is within **6GBLUR-JOINT**. By adhering to service-based approaches, this innovative architecture (see Figure 1) treats the **end-to-end network as a virtual service**, abstracting specific hardware to meet the distinct requirements of various services and applications. 6GBLUR envisions a future where the mobile network, from RAN to core, operates as a flexible and adaptable entity, ready to cater to the evolving needs of users and services. Through the power of disaggregation and virtualization, 6GBLUR empowers the mobile network to transcend its traditional boundaries by fostering a cohesive ecosystem. The virtualized RAN and core components operate in unison, allowing for dynamic resource allocation, rapid service deployment, and enhanced scalability.

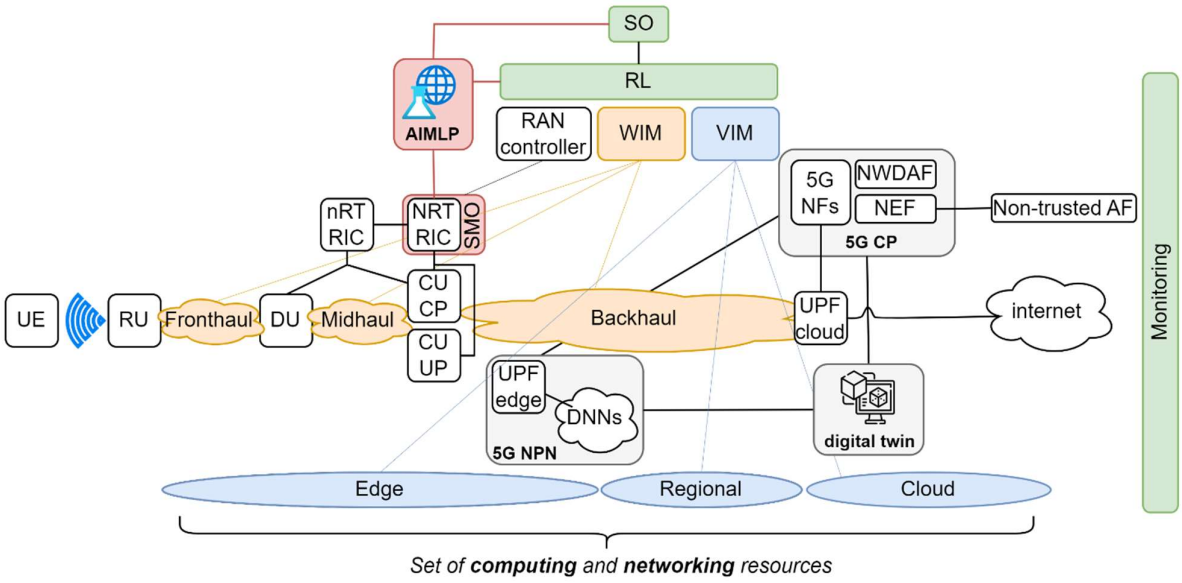

FIGURE 1. 6GBLUR ARCHITECTURE.

Besides, a significant effort will be devoted to the development of a cutting-edge 6G disaggregated and virtualized RAN architecture. The objective is to design a sophisticated **RAN control architecture** capable of efficiently managing multiple flavours of O-RAN deployments. This objective is within the **6GBLUR-SMART** sub-project. Such objective entails adapting to a diverse array of decision timescales, ranging from policy definition to real-time scheduling. By carefully orchestrating the virtualized RAN components through elements like the near real-time Radio Intelligent Controller (RIC), non-real time RIC, and SMO, 6GBLUR aims to create an agile and adaptable network infrastructure that optimizes RAN-related resource allocation.

2.1 Orchestration and end-to-end management

In the pursuit of seamless orchestration and end-to-end management within the network, 6GBLUR introduces a comprehensive set of elements that work in synergy [1]. At the forefront, the **Service Orchestrator (SO)** takes centre stage, offering a robust suite of capabilities for network service and resource orchestration. This empowers the instantiation of network slices across diverse domains, enabling seamless connectivity and collaboration.

Supporting the service orchestrator is the **Resource Layer (RL)**, a vital component that serves as the hosting ground for both physical and virtual compute, storage, and networking resources. Within this layer, network slices and end-to-end services find their execution space, while the infrastructure at vertical sites is diligently managed. It is within this resource layer that the required transport resources are orchestrated, enabling the interconnection of components spanning the network different domains.

To further bolstering the orchestration and management capabilities, 6GBLUR incorporates the **Virtual Infrastructure Manager (VIM)**, encompassing edge, regional, and cloud domains. The VIM takes responsibility for controlling and managing the computation, storage, and network resources within the Network Function Virtualization Infrastructure (NFVI) of a specific operator's domain. This centralized control ensures efficient resource allocation and utilization, optimizing network performance across various infrastructure domains.

To orchestrate and manage the overall transport network infrastructure, the **WAN Infrastructure Manager (WIM)** comes into play. The WIM acts as a central entity that controls and configures the connectivity between different NFVI points of presence (PoPs). It establishes and manages the links, routes, and network resources required for seamless communication between endpoints within the O-RAN architecture. The WIM optimizes network resources, ensures quality of service, and enables efficient utilization of transport infrastructure across the entire O-RAN network.

In addition, 6GBLUR integrates a powerful **monitoring platform** that excels at capturing the intricacies of a heterogeneous set of services and technological domains. Equipped with functionalities like log aggregation, a scalable data distribution system, and dynamic probe reconfiguration, this platform provides comprehensive insights and monitoring capabilities to ensure the network's smooth operation and swift response to any anomalies.

Completing the array of elements is the **AI/ML Platform (AIMLP)**, which takes on the role of managing and deploying AI models. Leveraging data gathered from different network domains, the AIMLP configures and trains these models, offering a catalogue of AI models that can be tuned and chained together to create more complex decision-making frameworks. These AI models are consumed by other 6GBLUR building blocks to enable intelligent decisions at various layers, such as service arbitration and service scaling, further enhancing the network's adaptability and efficiency.

2.2 5G Core and beyond

Looking beyond 5G, the evolution of the core network takes on immense importance in shaping the future of telecommunications. At the heart of this evolution lies the data plane, embodied by the **User Plane**

Function (UPF), which plays a crucial role in facilitating data transmission, inspection, and management within the network. Notably, UPFs can be dynamically deployed across different domains, including edge locations, empowering the core network with edge capabilities. This flexible deployment enables efficient data processing and routing closer to the end-user, enhancing latency-sensitive applications, enabling edge computing functionalities, and delivering seamless user experiences on a distributed and scalable level.

Several mechanisms are being developed that establish trusted and secure interactions between the core network and third-party applications. Among these mechanisms, the **Network Exposure Function (NEF)** stands out, providing a means for third parties to engage with the core network as untrusted **Application Functions (AFs)**. Collaborating with the **Network Data Analytics Function (NWDAF)**, the NEF facilitates on-demand reconfiguration of the User Equipment's (UE) data plane. This dynamic reconfiguration capability unlocks a wealth of opportunities for personalized service delivery, enabling real-time optimization and customization of the network experience for individual users.

A groundbreaking concept emerging within the core network landscape is that of **Non-Public Private Networks (NPNs)**. NPNs revolutionize network provisioning by offering dedicated, private network environments specifically tailored to the requirements of enterprises, industries, and government organizations. By providing enhanced security, reliability, and control, NPNs empower businesses to deploy customized network solutions aligned precisely with their needs. This revolution paves the way for optimized operations, tailored services, and increased operational efficiency.

Another exciting advancement in core network innovation is the concept of a **Network Digital Twin (NDT)**. By creating virtual replicas of the physical network infrastructure, network operators gain comprehensive and dynamic insights into the behaviour and performance of the network. These network digital twins offer significant benefits for the development and deployment of 6G networks, enabling operators to simulate, test, and optimize network configurations in a virtual zero-risk environment. Leveraging these virtual replicas, operators can proactively manage network resources, optimize performance, and seamlessly adapt to evolving user demands, ensuring a network experience that is both adaptive and seamless.

2.3 Transport

Managing the transport infrastructure in O-RAN networks encompasses several key components, including the Fronthaul (FH), Midhaul (MH), and Backhaul (BH) segments. These components, together with the WAN Infrastructure Manager (WIM), play crucial roles in establishing and maintaining connectivity between various endpoints in different NFVI-PoPs.

The **Fronthaul (FH)** segment of the transport network is responsible for carrying time-sensitive and latency-critical data between the Radio Units (RUs) and the Distributed Unit (DU) in the O-RAN architecture. This segment requires high bandwidth and low latency to ensure efficient transmission of real-time traffic. To manage the Fronthaul, advanced techniques such as Common Public Radio Interface (CPRI) and Ethernet-based fronthaul protocols are employed, allowing for high-speed data transfer and synchronization.

The **Midhaul (MH)** segment acts as a bridge between the Fronthaul and Backhaul segments, facilitating the transport of traffic between the DU and CU. The MH segment typically spans a larger geographical area than the Fronthaul, requiring scalability and flexibility to accommodate varying network demands. To manage the

Midhaul, technologies like Ethernet, IP/MPLS, and microwave links are employed, providing efficient connectivity and bandwidth allocation.

The **Backhaul (BH)** segment is responsible for carrying aggregated traffic from multiple RAN sites to the core network. It serves as the backbone of the O-RAN transport network, providing connectivity between the RAN sites and the centralized network resources. Managing the Backhaul involves ensuring high-capacity links, efficient routing, and resilience to handle the increasing data demands of modern wireless networks. Technologies like Ethernet, IP/MPLS, microwave, and fiber optic links are utilized in the Backhaul segment to provide reliable and high-speed connectivity.

2.4 RAN

6GBLUR approaches the RAN domain mainly from O-RAN, a groundbreaking approach to mobile communication networks that promotes flexibility, interoperability, and innovation. It disaggregates network components and utilizes (virtualized) open interfaces, and virtualization to enhance performance, reduce costs, and foster competition. In particular, disaggregation allows splitting the functionalities of a base station into different units and deploying them at diverse locations of the network. This collaborative and modular approach drives rapid evolution and benefits network operators and users with improved capabilities and innovative services.

In **Open RAN** networks, the traditional base station is split/divided into three components that can be placed in different geographical locations and can perform different processing functions/operations: the Radio Unit (RU), the Distributed Unit (DU) and the Centralized Unit (CU). RU and DU are connected through a fronthaul network, while DU and CU are connected through the midhaul network. Virtualization enables functional splitting, which determines which protocol layers' functions are implemented in each CU/DU/RU unit. The 3GPP specifications defined functional split Option 2 for CU/DU split, therefore considering a midhaul network with a fixed PDCP/RLC split. The O-RAN Alliance went a step further and defined functional split Option 7.2x (intra-PHY split) for the DU/RU split, thus introducing (beyond the 3GPP vision) the open fronthaul network and the DU/RU split. In addition to the two-layer split and the definition of standardized interfaces to enable open network architectures and interoperability, the Open RAN philosophy has standardized network intelligence through the introduction of **RAN Intelligent Controllers** (RICs) to manage the network, including near-real time RICs and non-real time RICs.

3 WP2-related Key Concepts

3.1 SMART-K1.1: CU and DU implementation

As opposed to most of the other key-concepts (KC), which are heavily research-driven, this key-concept, together with key-concept SMART-K2.1, is mostly focused on providing support for the experimental assessment of the other KCs. Therefore, the main objective is to implement an open-source RAN (Radio Access Network) platform that, although being based on 5G specifications, provides some cutting-edge features that are not currently deployed in commercial networks, which will establish the basis for the aforementioned experiments.

One of the most interesting features of the CU/DU implementation within the 6G-BLUR project will be the support for RAN slicing. RAN slicing [2] [3] refers to the concept that involves the creation of multiple isolated and customizable logical networks within the RAN physical infrastructure. These slices are tailored to accommodate the unique requirements of various services and applications in 5G. RAN slicing is a crucial element within the larger Network slicing framework, which encompasses the end-to-end customization of resources across RAN, transport, and core networks to meet the diverse demands of various services.

RAN slicing enables the segregation and allocation of specific radio resources, such as spectrum, bandwidth, latency, and quality of service (QoS), to different slices based on their distinct service requirements. This segregation ensures that each slice receives dedicated resources to fulfil its particular demands.

RAN slicing focuses on customizing and optimizing the radio access part of the network to meet specific service-level requirements. For instance:

- **Enhanced Mobile Broadband (eMBB)**: High data rates, high throughput, and improved user experience for applications like high-definition video streaming, virtual reality, and online gaming.
- **Massive Machine-Type Communications (mMTC)**: Support a massive number of connected devices with sporadic transmissions, like sensors in smart cities, agriculture, or industrial IoT.
- **Ultra-Reliable Low-Latency Communications (URLLC)**: Ultra-low latency, high reliability, and availability for critical applications like autonomous vehicles, remote surgery, and industrial automation.

RAN Slice-aware Radio Resource Management (RRM)

RAN slicing in the MAC scheduling [4] entails (i) the mapping of UE bearers into given slices, and (ii) the allocation of radio resources to different slices, in such a way that users get served according to the agreed KPMs for the different slices they belong to. The radio resource allocation involves several decision-making processes that determine both the quantities of time/frequency resources and the policies the scheduler should follow when serving the slices.

Slice-aware scheduling can be subdivided into two problems: (i) Inter-slice scheduling, and (ii) Intra-slice scheduling. Inter-slice scheduling primarily consists of allocating radio resources across the active slices in a DU to meet minimal traffic requirements. For instance, slices may be configured with different minimum/maximum PRB ratios per slot, and it will be the role of the inter-slice scheduler to ensure that the

agreed target PRB ratio requirements are met. Intra-slice scheduling, in contrast, consists of allocating radio resources to different UE bearers within a given slice, following an agreed UE scheduling policy. UE scheduling policies define the priorities in which user bearers should be served within a given slice, and the mapping of time/frequency resources to different bearers, based on the radio conditions of the communication channel. Scheduling policies target the optimization of a given metric for the UE bearers it serves, metrics that can be customized to the specific use-case; for example, common scheduling policies are *Guaranteed Bitrate*—that has the objective of serving users with a guaranteed throughput—or *Proportional Fair*—which aims to improve the average throughput, without starving users that suffer from less favorable radio conditions—or also *Minimum Latency*—which aims to minimize the bearers' latency, instead of maximizing the throughput.

Planned RAN slicing features

Within the 6G-BLUR project, SRS aims to implement and support various RAN slicing features, including the following points.

- **APIs for RAN slice configuration** - Implement APIs in the SRS DU for configuring RAN slices. The operator will be able to configure each slice via a configuration file. The implemented APIs should expose the traffic metrics of each active RAN slice to the RIC. Key metrics include the slice minimum, average, and maximum latency, as well as throughput.
- **Mapping of S-NSSAI to RAN slices** - As per the 3GPP standard, each UE Data Radio Bearer (DRB) is assigned a Single-Network Slice Selection Assistance Information (S-NSSAI), which is used to uniquely identify its Network Slice. It is then the responsibility of the DU to map the UE DRBs to the RAN slices that were made available in the DU. The operator will be able to define this mapping via a configuration file.
- **RAN slice-aware scheduler** - The MAC scheduler will need to support the multiplexing of different RAN slices and the allocation of radio resources within a RAN slice in such a manner that it meets the agreed traffic requirements [5]. This capability can be achieved through various techniques, namely through the selection of different scheduling policies on a per slice-basis (e.g. proportional fair, optimizing for maximum throughput and/or lowest latency, hybrid) or the allocation of dedicated radio resources to slices to meet minimal latency or throughput requirements. SRS will primarily target the allocation of dedicated radio resources to slices.

3.2 SMART-K1.3: Dynamic configuration of topologies/architectures for MH distribution

Introduction

Open RAN's modular architecture breaks down base stations into separate components (RU, DU and CU), offering flexibility and cost savings. This is the so-called horizontal disaggregation sought in this disruptive evolution of the radio access networks [6]. Operators can optimize each part independently, fostering competition among vendors and tailoring networks to specific needs. This separation of tasks, together with the possibility of deploying dynamically the units as virtualized functions in different DCs (Data Center), allows to obtain numerous advantages [7].

Also, FH and MH links concepts are born, which can be exploited to optimize the network conditions and, consequently, offer the best conditions depending on the use case. All of this ultimately improves efficiency and reduces costs further. Overall, Open RAN's approach offers a flexible, cost-effective solution for modern networks.

State-of-the-art (SoA)

In recent research and real deployments, Open RAN architectures have embraced distributed topologies as the basis of its design. In these setups, the DU is typically located at the cell site, establishing a direct point-to-point connection with the CU, situated closer to the core network. Also, these entities' locations are fixed, resulting in a static topology that does not adapt to the needs that may arise. This conventional deployment pattern, while functional, presents certain limitations that hinder the realization of its full potential.

Firstly, this distributed approach fails to harness the advantages afforded by DU centralization. But more importantly, the static nature of this static topology restricts flexibility and adaptability, posing challenges in accommodating dynamic network demands and evolving service requirements. With fixed point-to-point connections between DU and CU, operators may encounter limitations in scaling capacity, adjusting coverage, or deploying new services efficiently. This lack of flexibility in the RAN can impede innovation and the ability to respond swiftly to changing market dynamics or technological advancements.

Beyond SoA

In light of these limitations, there is a compelling proposition to establish a dynamic architecture that can intelligently respond to various conditions, including network load, service types, use cases, and other dynamic factors [8]. At the core of this proposal is the concept of network orchestration and automation, where intelligent algorithms and policies governed, for example, by the SMO, control the allocation and optimization of resources in real-time. By leveraging advanced analytics, artificial intelligence, machine learning, and automation technologies, operators can dynamically adjust the locations of the instances of DU and CU or change the routes between these two entities leveraging the MH configuration. This dynamic approach enables the network to optimize the services offered to end-users continually, thereby maximizing both QoS and QoE.

3.3 SMART-K1.4: Micro-orchestration of virtual network functions at DU-RU level

Introduction

Orchestrating cloud computing resources has traditionally focused on how compute workloads could be scaled, migrated, and executed across (highly) distributed computing elements in the cloud continuum. The latter typically comprises computing clusters powered by server-grade multicore processors, or high-end general-purpose processors (GPP) along with graphics processing units (GPU), which are used as function accelerators. With the movement of application and services towards the telco Edge, new considerations have arisen given the heterogeneity of the compute stratum used in Edge computing nodes. While the far edge may still typically look-like a small-scale datacenter, the near and extreme edge typically feature a heterogenous ecosystem of computing elements such as small footprint GPUs, embedded GPPs, micro-controllers, field programmable gate arrays (FPGA), application-specific integrated circuits (ASIC) and

System-on-Chip (SoC) devices that combine some or all the previous computing elements in the same silicon fabric [9].

SoA

Lately, the trend of multiprocessing SoCs devices has been enriched by embedding yet another type of processing elements optimized for artificial intelligence (AI) processing which is denoted as AI engines. Furthermore, with the latest advances in AI, purpose-built neural processors able to carry out intensive AI/ML workloads have been added to the bill [10], [11], along with Deep Learning accelerators employed for training and inference [12] [13]. These processing devices feature edge or cloud-optimized versions in order to address application or service-driven needs. Another striking difference of the computing resources in central clouds, as opposed to those in the far/extreme edge, is that compute resources are scarce and need to be leveraged with the most efficient way [14], exploiting at maximum their compute footprint, while satisfying a number of other KPIs related to energy efficiency and reduced latency.

In such edge environments, the operation granularity of hierarchical multi-site orchestrators needs to be extended or contemplated by engaging smaller-scale micro-orchestrators [15] able to efficiently leverage the computing granularity and address the heterogeneity of Edge computing elements. This is especially challenging in complex SoC devices that embed different general-purpose and accelerator-driven processing elements [16].

FPGA-based SoC devices are used as function accelerators across the fifth generation (5G) radio access network (RAN) and edge/cloud application domains. For instance, having as a reference the open RAN Alliance (O-RAN) architecture, this type of devices are encountered i) in open radio units (O-RU) accelerating low physical layer (PHY) digital signal processing (DSP) functions, ii) in open distributed units (O-DU) accelerating specific high-PHY functions (i.e., channel coding), iii) in open central units (O-CU) accelerating cryptography functions, iv) in the near-real-time RAN intelligent controller (RIC) hosting the inference of different machine learning (ML) models in the form of xApps, v) in network interface cards (NIC) for high-speed network interfacing, vi) in network switches for line-rate packet processing and also lately for in-network computing purposes, vii) in edge nodes accelerating applications that have high-performance, low latency and low power consumption prerequisites, and viii) in cloud computing infrastructures accelerating a wide range of compute intensive applications.

The underlying complex architecture of these multi-processing devices and the heterogeneity of the processing elements, makes challenging and cumbersome the combined virtualization of all the underlying compute resources, the exposure of deep telemetry data and the deployment of such devices in Kubernetes clusters. Equally challenging is the run-time fine grain adaptive management of the computing resources either at task or at function level. Different efforts both from industry and academia have been trying to address these challenges, focusing on specific use cases and offering ad-hoc solutions for specific families of SoC devices, aiming to serve concrete end-applications. For instance, Microsoft's Catapult v2 [17] work focuses on the offloading of network processing from the embedded processor to the FPGA area of the SoC device over Microsoft's Azure framework. Also, the work of IBM Research divides the FPGA spatially into distinct application regions, where hardware accelerated applications are to be programmed; the Service Logic secures access to shared off-chip memory and a dedicated host Processor-based server [18]. Finally, Amazon's AWS F1 instance (FPGA-containing VM instances) offers connectivity to eight FPGA cards which are

connected to a single physical server and a dedicated FPGA-only interconnection network [13]; no sense of abstraction or multi-tenancy is contemplated. Multiple academic works have also explored the deployment of FPGAs in cloud environments, but their thorough review goes beyond the scope of this abstract.

Beyond SoA

Currently, FPGA SoC devices are typically used as monolithic compute resources in the 5G telco cloud continuum; that is a single function/application reserves the entire device per timeslot (i.e., no multi-tenancy is applied), or multiple functions from one or different users reserve the entire device on a permanent basis (i.e., no time division multiplexing of resources is applied). But most importantly, this situation only refers to the FPGA resources of these SoC devices, which means that the variety of their compute resources are underexploited in the spatial and time domain. This deficiency becomes highly critical in edge environments where the processing requirements are elevated due to increasing deployments of edge applications and services offered to end users and vertical industry players. In order to serve the high-performance, latency-sensitive and power efficiency requirements of such edge applications, edge infrastructure owners need to flexibly leverage in a fine grain mode the full capacity and features of such devices. To achieve this, different programming methods like partial reconfiguration, asymmetrical multiprocessing and virtualization among many others need to be combined to create an overlay framework. Similarly, full observability needs to be contemplated in these devices to be able to adaptively manage the resources at run-time. By doing so, edge infrastructure owners could not only efficiently offer sixth generation (6G) services to their clients, but they could reduce their operating expenditures (OPEX) through a more energy-efficient management of the SoC accelerators.

In this work we aim to apply an AI/ML-driven micro-orchestration of application and virtual networking functions (AF and VNF respectively) running on FPGA SoC devices [9], or on similar SoC devices featuring top-up AI engines [9]. The latter will be typically employed in the far and deep edge to address the stringent latency, throughput and energy efficiency budgets of novel 6G applications and services. This type of micro-orchestrator would be able to reconfigure, scale, migrate, or replace functions across the SoC fabric and would also be able to split and place the computational load in different on-chip processing elements (e.g., the layers of a given complex neural network inference model) based on i) non-real-time policies and ii) near-real-time operating scenarios, and iii) real-time extreme Edge applications/services with highly varying operating conditions. To achieve this, the micro-orchestrator will leverage a wide range of on-chip extracted telemetry data for the training process of ML model (e.g., execution performance, fine grain power consumption of all processing elements, throughput of embedded buses, on-chip sensing data), and the reconfiguration framework described in [16]. The micro-orchestrator could be seen as a hierarchical data-driven intelligent controller application that would be devised by training a ML model with RAN traffic conditions combined with context-related semantic information and the mentioned on-chip telemetry data. Using the notions of O-RAN specifications, the micro-orchestrator could take the form of combined rApp and xApp, or a new type of real-time application located at the extreme edge [19]. The micro-orchestration could be also extended to cover multiple FPGA SoC function accelerators residing in the same edge computing cluster, likewise, applying a distributed split computing scheme. Although the AI/ML-driven micro-orchestration could be applicable to any type of AF or VNF accelerated in FPGA SoC devices, the initial focus will be laid on the O-RU and O-DU parts of the RAN. The rather static functionality of accelerated functions residing in the O-DUs and especially O-RUs, is expected to be challenged in 6G Use Cases, to contemplate,

device discovery, parameter exposure and run-time function reconfiguration based on real-time control loops [20]. This will open up new avenues to support a number of different applications and services built around massive MIMO techniques and cell-free architectures.

3.4 SMART-K2.1: Development and characterization of FH and MH

Similarly to key-concept SMART-K1.1, this key-concept is mostly implementation focused. The objective is to provide an open-source implementation of the O-FH interface between O-DU and O-RU. Although the implementation will follow the O-RAN specifications (a necessary requirement to be able to use commercial hardware to build the PoC set-ups), the fact that it is developed inside the project allows for fine tuning and customizations that will be instrumental for the innovations introduced by the other key-concepts.

Introduced in 5G networks, the functional disaggregation paradigm will play a fundamental role in future RAN architectures, including the one proposed by 6G-BLUR. As mentioned before, the idea is to split the gNB into a number of units that are dedicated to specific functionalities, are deployed at different locations of the network and implemented on different hardware platforms. The Open FrontHaul Interface (O-FH) is the communication interface used to link a Distributed Unit (DU), which implements the RLC layer, the MAC layer and (most of) the PHY layer, with one or multiple Radio Units (RU), which are in charge of the OFDM modulation/demodulation and the RF processing [21]. This separation of functionalities is known as *Split 7.2x* and finds a good tradeoff between keeping the RU as simple as possible (size, weight, power, …) and limiting the interface throughput.

The O-FH interface consists of two logical planes, namely the CUS plane for control data, user data and synchronization, and the M plane for the management of the RU. As the name suggests, the CUS plane is further split into different data flows to exchange data between the DU and the RU ([21]- Section 4.3)

1. User plane:
    a. Data flow 1a: flow of IQ data in the frequency domain (before OFDM modulation) on downlink (DL)
    b. Data flow 1b: flow of IQ data in the frequency domain (after OFDM demodulation) on uplink (UL)
    c. Data flow 1c: flow of PRACH IQ data in the frequency domain
2. Control plane:
    a. Data flow 2a: Scheduling commands (DL and UL) and beamforming commands
    b. Data flow 2b: License-assisted access (LAA) listen-before-talk (LBT) configuration commands and requests
    c. Data flow 2c: LAA LBT status and response messages
3. Synchronization plane:
    a. Data flow S: timing and synchronization data

Each O-FH message belongs to a specific data flow that is encapsulated in an eCPRI message which is transported over an Ethernet frame.

Implementation details

The development of the O-FH interface within 6G-BLUR will build on top of the SRS O-FH library. This O-FH

implementation has been split into different components to make it modular, easy to extend, easy to maintain and easy to test. The two main components are mapped to the transmitter and the receiver side, both of which include all the data flows previously mentioned. Individual libraries have been developed for different compression algorithms ([21] - Section 8.1.2 and Appendix A), eCPRI packing and unpacking and raw Ethernet frame management.

Additionally, for performance reasons, both the transmitter and the receiver are executed concurrently in different threads to minimize the latency required to process the IQ samples.

Next steps

In the context of 6G-BLUR, SRS plans to extend the functionality of the O-FH interface with a number of optimizations to improve its efficiency and support the other advances proposed by the project. For instance, the Data Plane Development Kit (DPDK) [22] can be used to reduce the latency introduced when passing network data between the kernel and the user space, an operation that requires data copying and a context switch. On a similar note, IQ sample compression/decompression and packing/unpacking can be further sped up using SIMD instructions. Moreover, several tests and adjustments are needed to meet the requirements of the 6G-BLUR use cases. For instance, even though the SRS O-FH implementation is already compliant with the transmission and reception window parameters specified by the O-FH specification, these parameters have to be tuned to adapt to the specific link and network requirements of UC2 PoC1, which deals with non-dedicated, long-distance fronthaul networks. Also, it is planned to implement vLAN separation for the user plane and the control plane. This will allow using different vLAN IDs for the control plane and the user plane messages, for better orchestration of the network. Relevant configuration options will be added to control this behavior.

3.5 SMART-K3.1: AI/ML Workflows in O-RAN

Introduction-Streamlining O-RAN Through Explainable AI with LLM Integration

The basic objective of this key concept is to leverage advanced ML techniques to enhance the intelligence of RAN, deployed as an xApp of 6GBLUR: Enhancing RAN Operations with LLM-Based Dynamic Dashboard and Full Dynamic RAN Slicing.

Large Language Models (LLMs) have emerged as powerful text processing tools that inject intelligence into various domains, including natural language understanding, generation, and translation. Recent versions of LLMs have expanded beyond text to support multiple modalities such as video, audio, and voice, enhancing their versatility and applicability.

SoA

Our project, 6GBlur, aims to revolutionize RAN operations by introducing full dynamic RAN slicing, underpinned by advanced machine learning (ML) techniques. We focus on creating, adjusting, and deleting preconfigured RAN slices dynamically to optimize network efficiency and user experience. By incorporating explainable AI (XAI) and Large Language Models (LLMs), we ensure transparent decision-making processes and provide network operators with intuitive, adaptive dashboards for insights into network performance

and xApp decisions. The project will leverage ML-ops for robust backend infrastructure, enhancing the O-RAN ecosystem's capabilities for future networks.

LLMs currently lack an understanding of their environment, so deploying them in an actual O-RAN scenario requires a grounded approach. This can be achieved through careful system message design and prompt engineering, incorporating relevant O-RAN deployment details during LLM inference. Additionally, complex tasks cannot be accomplished in a single attempt, even by an intelligent human, necessitating the breakdown of tasks. An agent-based workflow has emerged as a powerful paradigm, decomposing problems into various components, providing tools for the LLM, and offering feedback from the environment [23]. This approach effectively bridges the gap between the LLM's lack of grounding and the complex dynamics of systems like O-RAN. In the case of O-RAN systems, digital twins can be utilized to provide feedback and facilitate the development of LLM-based agent workflows.

Beyond SoA

Main achievements expected to be implemented at the end of the projected are summarized below:

1. **Dynamic RAN Slicing with ML:** Implement ML for dynamic RAN slicing, adapting in real-time to network demands. Ensure efficient resource allocation and fulfillment of user QoS requirements.
2. **LLM-Based Adaptive Dashboard:** Develop an adaptive dashboard powered by LLMs, offering intuitive insights into network performance. Utilize LLM agents for automatic conflict detection among xApps, optimizing network management.
3. **Explainable AI Integration:** Incorporate XAI to make the ML-driven decision-making process transparent and trustworthy. Build confidence among network operators in automated operations.
   ML-Ops Infrastructure: Establish a robust backend infrastructure for deploying and managing xApps and potential rApps. Support the evolution of the O-RAN ecosystem in anticipation of future network demands.
4. **Innovation and Novelty:** Introduce dynamic RAN slicing, LLM-based dashboards, and ML-ops frameworks. Enhance operational efficiency, transparency, and adaptability in RAN operations.
5. **Exploitation Potential:** Offer RAN slicing as a service, enabling new use cases like XR and AR/VR. Improve conflict mitigation among xApps through XAI, ensuring smoother operations.

6GBLUR represents a significant leap forward in RAN operations, harnessing the power of ML, LLMs, and XAI to introduce dynamic and fully adaptive network management. By enhancing the O-RAN ecosystem's capabilities, our project sets the stage for future network innovations, ensuring optimized resource allocation, enhanced user experiences, and a new standard in network operations transparency.

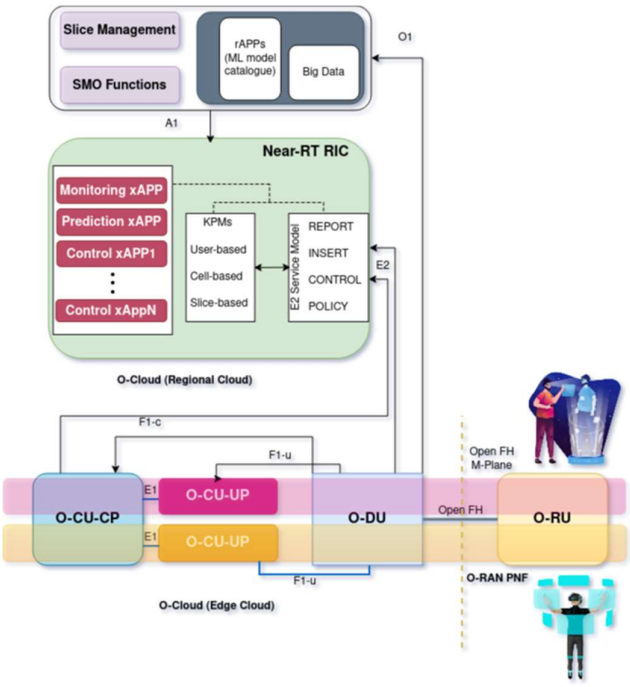

FIGURE 2. AI/ML WORFLOW FOR MONITORING AND CONTROL

The diagram in Figure 2 illustrates the deployment of monitoring, prediction, and control xApps in the Near-Real-Time RIC. The O-RAN system is divided into two slices, each catering to different types of applications or virtual operators

4 WP3-related Key Concepts

4.1 SMART-K1.2: RAN Orchestration

Introduction

Mobile networks are continuously evolving thanks to the significant efforts being made through research and standardization. This evolution has enabled the improvement of the services offered and the support of various novel use cases, such as Industry 4.0, autonomous driving and smart cities. 5G networks have played an important role in the establishment of such use cases, through the introduction of advanced features and technologies. However, despite the improvements brought by 5G, time-critical applications, such as that of the immersive technologies of eXtended Reality (XR) and Cloud Gaming (CG), still remain a challenge due to their stringent requirements in terms of data rates and latency. Open Radio Access Network (RAN) is a technology that leverages disaggregation and intelligence, thus empowering the network with flexibility, scalability, and automation. It is thus envisioned to play a key role in the next generation networks [24]. In fact, Open RAN is considered to be an enabler for XR and CG applications, since it is expected to meet the imposed requirements thanks to the enhanced network performance, the dynamic resource allocation, the adaptability of the network based on traffic changes and the intelligent network management [25].

O-RAN Alliance[2] is a community formed by various Mobile Network Operators (MNOs) in order to provide standardized specifications for the Open RAN architecture. The O-RAN architecture adopts the functional split 7.2x, where the gNB functionality is split into the Central Unit (CU), Distributed Unit (DU) and Radio Unit (RU) (defined as O-CU, O-DU in the O-RAN specifications [26]). The CU is further divided into the Control Plane CU (CP-CU), that includes the RRC and PDCP layers and the User Plane CU (UP-CU) that the SDAP and the PDCP layers. The DU includes the RLC, MAC and high-PHY layers, while the RU includes the low-PHY layer. Moreover, control loops are introduced, known as RAN Intelligent Controllers (RICs) that enable flexible and automated radio resource management through a set of open interfaces. The interconnection of the various functional blocks is achieved through a set of open interfaces. One of the most important specified interfaces in the O-RAN architecture is the Open Fronthaul that interconnects the DU with the RU and that allows for interoperability among equipment of various vendors. The Open Fronthaul, however, can result in a bottleneck and degrade the performance of the air-interface, especially when considering time-critical applications such as the XR and CG and/or various RUs share the same fronthaul link with target the operator's cost reduction.

SoA

Various works in the literature have targeted the study of the constraint fronthaul in disaggregated architectures. While a variety of possible solutions have been investigated, one promising strategy is that of fronthaul compression, where the information that is transmitted through the fronthaul link is either reduced partially, or completely. An adaptive compression optimization is presented in [27] that targets to minimize the fronthaul transmission rate, while the work is further extended in [28], where two adaptive compression schemes are proposed that minimize the fronthaul transmission rate considering the BLER constraint.

[2] O-RAN Alliance, Available at https://www.o-ran.org/, 2024. Accessed: 5-02-24.

Another solution is presented in [29], where a linear prediction coding with adjustable scaling is used to perform fronthaul signal compression. The combination of fronthaul compression and BS selection is proposed in [30]. Another scheme that is considered in the specifications of O-RAN [21] is modulation compression that reduces the IQ bitwidth transmitted through the fronthaul link in the downlink direction. The reduction is carried out considering the modulation order in the air, therefore it achieves a trade-off between the fronthaul and the air interface [31]. [32] and [33] presented an interesting approach where modulation compression is combined with fronthaul-aware scheduling methods. Their study has shown that this combination improves significantly the resource utilization, while the impact on the air interface is limited.

Beyond SoA

Despite the fact that fronthaul compression strategies has been demonstrated to be an effective tool for addressing the limitations imposed by the fronthaul, most of the existing works are studying predominantly simplistic traffic models that do not present stringent requirements, such as those inherent in XR and CG traffics. With this work we will target to fill this gap and provide a study related to the impact that the constraint fronthaul link can have on XR and CG traffic, using 3GPP defined realistic scenarios. To perform the evaluation, the strategies presented in [32] and [33] will be implemented and extended using the open-source 5G-LENA network simulator. Various extensions will be performed to the algorithms with target to improve even further the achieved performance. Moreover, various schedulers, including QoS schedulers proposed throughout the project, will be studied in conjunction with the Fronthaul Control and evaluations will be performed focusing on time-critical applications, such as the XR and CG.

4.2 SMART K1.5: Dynamic configuration of FH for distributed and centralized architectures

Introduction

Open RAN heralds a revolutionary shift in base station architecture by adopting a three-tiered processing approach, dividing functions into RU, DU, and CU. Called horizontal disaggregation, this approach not only enhances deployment flexibility but also improves some functionality while simultaneously driving down costs. With RU handling radio transmission and reception, DU managing baseband processing at the cell site, and CU orchestrating network control functions closer to the core, Open RAN topology presents a more versatile, efficient, and cost-effective infrastructure.

SoA

But at present, Open RAN networks primarily depend on static and distributed topologies. This means that the radio access network elements, RU, DU, and CU, are typically deployed in fixed locations, maintaining a consistent configuration over time [26] [34]. Moreover, RU and DU are collocated at the cell site, giving little flexibility to the configuration of the FH according to the needs and therefore not leveraging the numerous advantages of centralized architectures.

Beyond SoA

The spatial diversity in the Open RAN architecture presents an additional opportunity: dynamic architectures.

This would enable the dynamic relocation of network elements to align with the unique requirements of each scenario, ensuring consistent attainment of the expected QoS standards. In this sense, the research is more focused to provide flexibility to offer a network that is more responsive to the different and increasingly wide range of use cases [35]. The aim is to adapt dynamically, through automation and AI/ML algorithms, and controlled for example by the SMO, both the topology of the network, to meet the QoS of the different services according to the variable network loads found in the FH; and also, the split configuration, to set it to the most appropriate configuration for each type of use cases.

The chosen split option depends on the scenario, given that the split configuration is a tradeoff between radio performance and transport capacity. Higher-layer splits relax transport capacity requirements unlike the Lower-layer splits that are more beneficial for the radio performance, but they may require a more expensive and high-capacity transport network [36].

The optimal approach would involve implementing a flexible architecture capable of dynamically selecting the most suitable split configurations based on the prevailing radio performance requirements and the availability of transport resources [34]; and the most suitable topology based on the timing and performance needs. This adaptive architecture would seamlessly adjust the distribution of functions in real-time, ensuring efficient resource utilization and optimal network performance under varying conditions.

4.3 SMART-K2.2: Characterization of the FH and MH transport

Introduction

As it is well known, in traditional RAN networks the BS (Base Station) constitutes a single processing unit, formed by the RU (Radio Unit) and the BBU (Base Band Unit), which is responsible for managing the data of a single cell site [6]. In this type of deployment, all components and interfaces are proprietary and are not interoperable with those of other vendors.

In Open RAN networks, however, both units are disaggregated into DU (Distributed Unit) and CU (Centralized Unit), where each of them is responsible for processing specific functions from RAN protocol stack and are separated by the different split options standardized by the O-RAN Alliance [21], depending on the functions remain in each unit. This separation of tasks, together with the possibility of deploying dynamically the units as virtualized functions in different DCs (Data Center), allows to obtain numerous advantages [37], providing greater flexibility and scalability to the network design, allowing more efficient deployments and greater virtualization, among others. In addition, all components, interfaces and NFs (Network Function) are open and from different suppliers, which gives the system greater flexibility, faster TTM (Time To Market) and better capacity to evolve when new technologies and use cases need to be provided [38].

Within this approach there are two main types of architectures [39]: distributed and centralized. In the former, there are two options: full distributed site (DU and CU located at the cell site) and partial distributed (or centralized, depends on the direction seen), where DU is at the cell site and CU is centralized. The most common deployment is the second one, DU at the cell site and CU centralized, which is typically called “D-RAN”. The centralized topology considers typically full centralization, DU and CU centralized, which is typically named as “C-RAN” and it is an evolution of the previous case, where only CU is centralized. In this case, all the equipment (DU and CU) is moved to be located closer to the core with the intention to serve multiple cell

sites, obtaining numerous advantages such as pooling gains, cost reduction, better scalability, improved efficiency features, better load balancing, improved network operation and maintenance, etc.

SoA

It could be easily concluded then that the network evolution should follow the trend to move progressively to these centralized deployments [40], [41]. This means centralizing the DU from the cell site to a central location deep in the network and closer to the core than the edge. This would mean evolving the networks to meet the specific conditions required by the fronthaul interface so that it could be achieved. For this reason, in order to be able to deploy this architecture (C-RAN), it is first necessary to characterize FH interface trying to derive different types of profiles that will be representative of different RAN scenarios imposing different requirements in time and capacity to see their behavior under those conditions [42].

We require characterizing FH interface in time because of eCPRI properties. eCPRI is the O-RAN FH protocol used between RU and DU to allow open and multi-vendor interoperability between radio and baseband unit. eCPRI is a very latency sensitive protocol, which has some time constraints that needs to be met. The most limiting factor is perhaps the HARQ (Hybrid Automatic Repeat Request) protocol, which allows a maximum time in FH of 250 µs [43]. In the case of the MH, it is more tolerant as it supports a maximum RTT delay of around 10 ms.

FH characterization in capacity is also important due to the expected high amount data expected to be coursed due to the centralization of several cells. Current transport networks are not prepared to support FH traffic along with the other supported services, since they were not dimensioned initially for this double purpose. Unlike, MH imposes less stringent requirements, not only in time but also in capacity, as expected capacity will not be so high, and it could be served by using the existing network.

Thus, the issue with the existing transport networks is that they are limited to a maximum capacity resulting from the actual services offered (coursed though) today, but they cannot accommodate the huge amount of traffic that would require C-RAN deployments, to course FH data rate. Thus, in case to support C-RAN deployments, there are two possible options: (1) upgrade completely all the transport network, dimensioning considering now FH traffic, which will require to a significant investment for the MNO but also complexity and long times until this complete evolution will be a reality; (2) use P2P links, through dark fiber deployment between RU at the cell site and DU need to be deployed ensuring required latency and capacity are met, but this will limit the scalability of these scenarios to just few locations, as it is not practical to deploy in each cell site P2P links. Hence, the work proposed here aims to explore the option of evolving those links to re-use, as much as possible, existing transport network (with the required optimizations and upgrades but seeking for minimal changes) so that they can support these type of centralized scenarios, where DU is deep in the network, and allow centralized DUs to aggregate traffic from multiple sites, adding complexity to the amount of data they can support and to the DU processing capabilities.

Beyond SoA

Consequently, it is necessary to study the FH interface in detail under this centralized topology based on characteristics of the existing transport network. From a theoretical point of view, this means characterizing different link profiles, especially the FH since it is the most crucial and limiting, needing to analyze different parameters to examine the performance of the links under different radio conditions and different level of

cell aggregation (centralization). From this, it will be possible also to extract valuable information to carry out the dimensioning of the transport network elements for different scenarios under different radio load conditions. From a practical perspective, this can be characterized by using network emulators plugged in between DU-RU, for emulating those identified FH profiles, and between DU-CU, for characterizing the MH, in an Open RAN E2E testbed, to validate whether the assumed stress conditions are going to be feasible and ensuring operation of this network topology (C-RAN).

Moreover, the timings observed on the DU side will serve as an input for analyzing and determining necessary adaptations in the DU protocol stack to allow the required level of centralization. Additionally, this study will function as a means to accurately determine the sizing of the network components, as mentioned previously. By considering timing and capacity constraints, it becomes feasible to ascertain the quantity and capacity of elements implicated in the centralized architecture blueprint.

Furthermore, it will provide valuable insights into the extent of optimization required to achieve the desired level of centralization within the network. It will involve implementing various mechanisms aimed at addressing potential impairments that may arise, particularly concerning the competition between FH traffic and other traffics injected in the same routers (network elements) of the transport network.

Finally, as it can be seen, studying the characterization of the links following industry standards, using real equipment and real transport network data allows not only to know which scenarios are potentially valid, but also provides different outcomes for other analyses to be performed so that together, integrating all the work, it is possible to build deployments based on centralized architectures that in the short term can offer to users the numerous advantages that this type of topologies present.

4.4 SMART-K2.3: Flexible functional split

Introduction

5G has developed different technologies to enhance flexibility and efficiency within mobile networks. The new 5G virtualized base station can be divided into three components that are placed in diverse geographical locations: the Radio Unit (RU), the Distributed Unit (DU) and the Centralized Unit (CU). RU and DU are connected through a fronthaul network, while DU and CU are connected through the midhaul network. Virtualization enables functional splitting that determines which protocol layers' functions are implemented in each unit. Depending on the chosen split, protocol layer functions can either be implemented locally in the RU/DU, resulting in a distributed setup, or centrally, in the CU. It is worth mentioning that 3rd Generation Partnership Project (3GPP) specifications just defined functional split Option 2 for CU/DU split [44], and so just the midhaul network was considered with a fixed PDCP/RLC split, although there were studies in 3GPP that analyzed the trade-offs of various functional split options [37] and further studies addressing specifically the low-layer splits [45]. Open Radio Access Network (O-RAN) Alliance, instead, further defined the new RU entity and adopted the so-called functional split Option 7.2x (intra-PHY split) for the DU/RU split [46].

Each functional split offers different trade-offs in fronthaul/midhaul capacity and latency requirements versus implementation and network performance. Centralizing baseband functions offers significant benefits for network operators. For instance, coordinating the same protocol layer across different cells can reduce interferences, enhance reliability, or simplify handover procedure, while reducing costs [47]. However,

achieving maximum centralization is challenging due to the limitations in midhaul network capacity, diverse traffic delay requirements, functional splits constraints and asymmetric computational resources, among others [48] [49]. The study in [50] examines the delay introduced across different functional splits. It is concluded that as centralization increases, induced delay also increases. Therefore, it is observed that as traffic latency requirements become more restrictive, the functions are processed at the edge (DU or RU), so the average power consumption increases by 32.8%, resulting in higher operational costs. This observation is further supported by [51] and [52], which highlight that higher low-latency traffic loads result in lower centralization. Therefore, flexibility in choosing the optimal functional split is crucial to meet the diverse throughput and latency demands of different traffic types and reduce operational costs. This flexibility allows to serve diverse traffic types, optimizing resource utilization and network performance based on specific requirements [53] (see Figure 3).

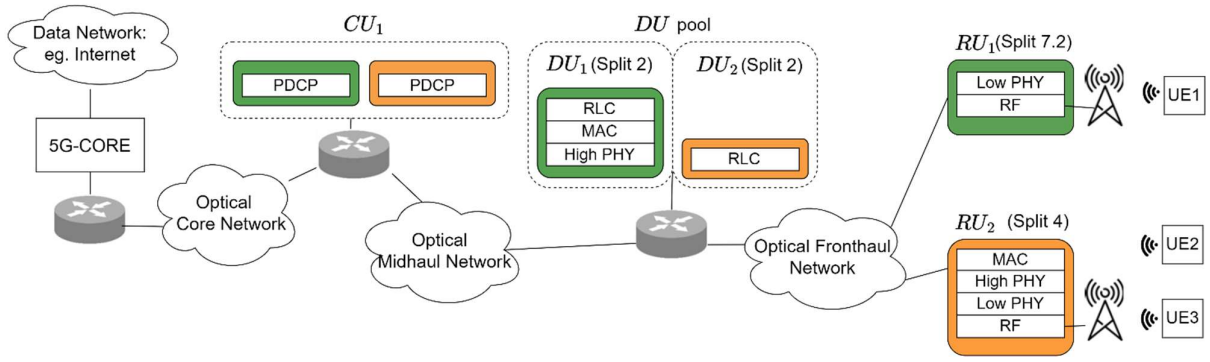

FIGURE 3. CU-DU-RU IMPLEMENTATION EXAMPLE WITH FLEXIBLE FUNCTIONAL SPLITS FOR DIFFERENT TRAFFIC TYPES.

SoA

Some current researchers have studied the selection of the best functional split considering the scenario characteristics, traffic type, the used computational resources or operation monetary cost. The paper [54] analyses different functional split backhauls and RAN implementation constraints and shows the benefits to use different functional splits depending on the transport network characteristics. They highlight that the biggest challenge in flexible functional splits lies within the architectural and implementation domains.

In [55] an optimization problem is proposed to select jointly the optimal functional split and routing path for each RUs and CU pair, looking for a tradeoff between performance and associated monetary costs. The work in [56] also relies on an optimization problem to select dynamically the optimal functional split considering a scenario with heterogenous traffic where the midhaul bandwidth is limited. The optimization problem tries to maximize the centralization of the Cloud Radio Access Network (C-RAN) system, and it concludes that the dynamic functional split selection gives 90% more centralization over the static functional split. In [57], an optimization problem is proposed which calculates the optimal baseband unit (BBU) hotel or centralization placement for cost reduction considering a mixed functional split scheme. To reduce the cost, the optimization goal is to decrease the number of BBU hotels and fibers in the network. They also propose a heuristic algorithm to solve the proposed model for large network scenarios. The results show that a mixed

functional split improves the BBU centralization comparing to each node working in a fixed split. The paper in [58] selects the optimal functional split based on the traffic split, slice (logical network instance operating on a shared network infrastructure [59]) specifications and different placement options, trying to maximize the centralization degree and minimize the active servers. The results show that slices with more stringent requirements do not support high centralized functional splits. Moreover, a heuristic algorithm is proposed, to reduce the high computational complexity when solving the problem for large-scale scenarios. The authors in [60] find the optimal functional split to minimize the operator's total cost and satisfy the user requirements. In order to do this, an optimization and a heuristic problem are proposed which select the base stations that are inactive or sleeping, the functional split and routing path. In [61], the proposed optimization problem seeks to identify the optimal functional split to minimize computational resources usage while maximizing the centralization or aggregation level of functions. The main objective in [17] is to reduce the operational cost. To do this, they select the optimal functional split, slice request admission, CU/DU placement and the traffic routing.

TABLE 3. SUMMARY OF REFERENCES.

	Problem formulation		Monetary cost	Centralization
	Optimiz.	Heuristic		
[61]	✓			Maximize, while minimizing the computational cost
[55]	✓		Selection of optimal functional split, MEC placement and routing path to have a tradeoff between performance benefits and monetary cost	
[56]	✓			Maximize, considering the midhaul bandwidth and the given traffic load at any given time
[57]	✓	✓	Optimize BBU placement	
[58]	✓	✓	Minimize active servers	Maximize
[60]	✓	✓	Reduce active BS for power consumption	Maximize
[62]	✓		Select functional split, DU/CU placement, traffic routing request admission	

The previous papers were focused in reducing operational costs and/or maximizing the centralization based on some constraints such as traffic type, traffic load, computational resources, etc. Despite these mechanisms are needed, there remains a notable scarcity in literature regarding detailed implementation of flexible functional splits. The paper [63]is the first work that addresses flexible functional split implementation challenges, where an adaptive 5G RAN implementation is proposed based on srsLTE [64]

software platform, that provides runtime functions migration in order to support flexible functional splits. They conclude that switching functional splits at runtime is possible at spences of introducing packet losses or additional delay. The work in [65] has develop a testbed that supports four different functional splits using the implementation of srsRAN 4G [66]. The work [32] proposes a FH-aware scheduling methods and uses the ns-3 5G-LENA network simulator to evaluate it. In this case, the selected functional split is 7.2x from O-RAN.

Beyond SoA

In this context, we aim to extend and enhance the functional splits implementation within the ns-3 5G-LENA simulator. Our objective is to analyze which is the best functional split based on factors such as traffic demand and load. Based on the FH control implementation in the ns-3 5G-LENA simulator that is currently being developed, where the capacity constraint set by 7.2x functional split is considered, in this work we will contribute to this code by introducing the 4/6/7.x/8 functional splits capacity and latency requirements. These constraints will limit the transmission of packets through the FH. This contribution is the first step towards enabling the creation of functional split selection algorithms in ns-3 5G-LENA simulator that can dynamically select the most convenient functional split based on the current scenario.

4.5 SMART-K2.4: Radio Stack optimization

Introduction

In modern times, there exists a widespread research and development within the centralization of the Open RAN architecture [44], a strategic shift driven by the manifold benefits and advantages it promises to offer [7]. Now, this progress focuses on the centralization architecture, deploying the DU in closer proximity to the core network to achieve several benefits, for what it will be needed first to perform some modifications in the system to make it a reality.

DU centralization overstresses some of the functions of RAN protocol stack, particularly those associated with the MAC layer. This basically introduces a series of adjustments and considerations aimed at accommodating the longer distances derived from that level of centralization and optimizing processing techniques in the new architecture. As a consequence, the protocols operating within the MAC layer will need to operate under longer latencies which this arises a need to assess and optimize the performance of these protocols under these conditions to ensure efficient operation in these centralized scenarios as it occurred before with the distributed topologies.

SoA

Nowadays, as outlined by the O-RAN Alliance, the maximum distance this link can sustain is approximately 20 km (equivalent to approximately 100 µs in propagation time, one-way) [45], which correspond to the requirements of eCPRI protocol involved between DU and the RU and the type of service. The utilization of Split 7.2x involves splitting the physical layer, with the low-PHY residing in the RU and the high-PHY along with the MAC layer located in the DU. Implementing these layers separately or at a greater distance from the user point of view, results in increasing fronthaul latency which will result, in turn, a reduction of the time budget available for the RU and DU to process L1 and L2 layers that are also highly sensitive to timing.

The main limiting factor relies on the HARQ protocol, which operates at the MAC level [46]. HARQ plays a pivotal role in enhancing the reliability of data transmission through error detection, correction, and retransmission mechanisms triggered in response to packet loss or timeouts. Now, for LTE, the HARQ protocol imposes receiving an ACK/NACK message within 8ms RTT. The DU processing time is usually around 2750 us [47], allowing only a time budget of 250 µs for Round-Trip Time (RTT) between the RU and the DU [51], [49], that is, the budget for the FH. Visually, this can be seen in Figure 4.

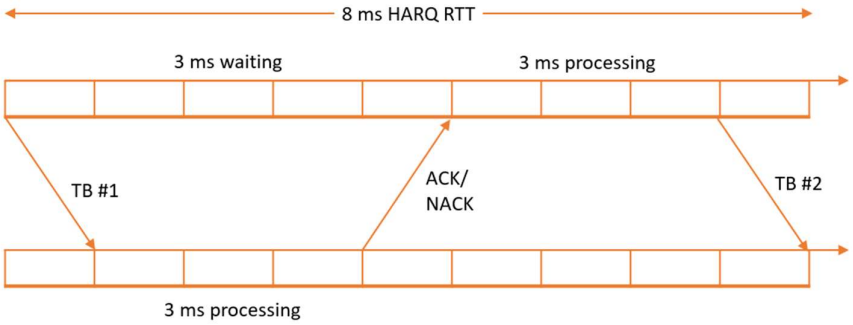

FIGURE 4. HARQ ROUND-TRIP-TIME

As it is shown, it seems maximum allowed FH timing is limited to only 250us, as indicated above and, indeed, we have found that is nearly the actual implementation in RAN networks from different vendors, which imposes several challenges to achieve the presented centralization. Fortunately, there are different proposals suggest it is possible to enhance this process and allow for a better timing, as it proposed in the literature [50]. It has pointed out, for instance, the concept of early feedback. This approach entails providing feedback at an early stage of the decoding process, leveraging the output from the decoder before its completion. A notable example of this approach is outlined by [67], wherein the authors propose a two-step prediction process. Initially, they estimate the bit error rate (BER) based on likelihood ratios, followed by utilizing this estimation to predict the output of a block error indicator predictor using predefined thresholds. This methodology effectively reduces latency and achieves a high accuracy rate of over 90% in generating correct feedback messages. In addition, in [68] researchers forecast the necessity for a re-transmission by analyzing the channel state information. This involves leveraging data on the quality and stability of the communication channel to anticipate whether a message or data packet may require being sent again for reliable delivery.

Beyond SoA

In light of this consideration, it becomes realistic to pursue the introduction of certain adjustments at the DU to counterbalance the increased delays within FH network. This is essential to ensure the seamless operation of all protocols within a network topology where certain functions are placed far away from the central site. The implementation of these modifications involves a meticulous evaluation of the fronthaul latency characteristics and the development of tailored solutions to mitigate any adverse effects on protocols performance.

In this sense, it is planned in this project to study what are the modifications needed. Some of the options could be explored basically relies on three concepts:

- Allowing UL to send a pre-acknowledgment by the receiver even before the data has been sent by the transmitter, so that the feedback arrives in time like in a typical link. In the case of DL, the BS would assume that the packet is successfully received. For the time advance process, it is important to remark that the receiver knows at which time the UE will transmit given that it is stated previously by the scheduler, so also it knows when to send the pre-acknowledgment.
- Increase the number of HARQ processes per UE in both DL and UL as it is already defined by 3GPP in Release 17 [69]. This number is chosen in relation to the RTT to ensure continuous transmission. The more RTT, the more HARQ processes there should be, but minimizing them as much as possible, given that it implies more signalization and processing [53]
- Another function defined in Release 17 could be explored is the option from the RRC layer of disabling semi-statically (process by process) the number of HARQ processes in each moment based on a BLER (Block Error Rate) target. In those cases, the reliability of the channel would rely more on the ARQ of the RLC (Radio Link Control) layer in asynchronous mode, giving more flexibility to the implementation since it allows a longer timeout.

All these solutions, among others, should be considered to make possible those centralized topologies at far edge distances.

4.6 SMART-K2.5: QoS Management for time-critical immersive applications

Introduction

Next generation mobile networks (i.e., 5G and 6G) are designed to support time-critical technologies such as eXtended Reality (XR) - encompassing Virtual Reality (VR), Augmented Reality (AR) and Mixed Reality (MR), and Cloud Gaming (CG) that seamlessly blend the real with the virtual world, enhancing the user experience. This evolution creates the opportunity for various industry sectors to integrate immersive experiences into their services. Examples include the domains of entertainment, healthcare, real-estate, manufacturing and education, among others [70].

Nevertheless, the inherently time-critical nature of these applications remains a challenge. In fact, properly managing XR and CG traffic considering its stringent Quality-of-Service (QoS) requirements, in terms of latency, bitrate and reliability is crucial, resulting in the need for advanced scheduling algorithms. Moreover, XR applications such as VR and AR can generate traffic consisting of various streams (or flows), known as multi-flow traffic, each one with distinct QoS requirements. To address these challenges, 5G New Radio (NR) introduced the QoS model that is a new framework for QoS provisioning that moves the control at the QoS flow level [71]. This architecture can be particularly beneficial for applications such as the XR, since it allows the management of multi-flow traffic with time-critical requirements. However, 3GPP standardization does not specify the technical and implementation details of the MAC scheduling algorithms. As such, the design of effective scheduling algorithms is attracting the attention of researchers.

SoA

Several works in the literature have investigated QoS MAC schedulers for various use cases considering 5G NR networks. For example, a QoS-aware and channel-aware Radio Resource Management (RRM) framework has been presented in [72] considering a multi-numerology 5G NR network. The work is focused on the

effective management of Best Effort (BE) and Guaranteed Bit Rate (GBR) traffic. Solutions based on machine learning and reinforcement learning have also been studied. Work in [73] proposed a knowledge-assisted deep reinforcement learning algorithm for time-sensitive traffic in 5G networks, while in [74] the authors proposed the PreQoS mechanism that is a QoS prediction scheme for Vehicle-to-Everything (V2X) services in beyond 5G systems leveraging machine learning techniques.

Although there is a vast amount of literature covering a broad spectrum of use cases, research on QoS management for XR traffic is currently under way, therefore related work is still in its early stages. A scheduling framework, called 5MART, has been presented in [75] considering heterogeneous traffic of VR, video, VoIP and FTP. In this framework, the authors apply reinforcement learning and neural networks to find the best scheduling decisions. Heterogeneous traffic is also studied in [76], where a resource allocation scheme is proposed employing reinforcement learning for QoS provisioning. However, it's important to acknowledge that while these approaches are addressing QoS provisioning for XR traffic, they may not fully account for all the unique challenges posed by XR traffic, such as the generation of multiple service data flows with different QoS requirements.

Beyond SoA

To address the challenges of the time-critical applications, such as the XR and CG, work will be focused on proposing a generic QoS MAC scheduler that will perform QoS provisioning at the flow level and that will properly treat various time-critical traffic types by taking into consideration the QoS requirements of each service data flow. Our vision and novelty for this proposed scheduler is to satisfy the requirements of each flow by considering a combination of QoS flow descriptors and real-time measurements. Moreover, to handle multi-flow traffic, a QoS assignment algorithm at the level of Logical Channels (LCs) will be properly designed and proposed, that will account for the QoS characteristics inside the various flows of the same user and therefore ensure very fine flow-level granularity of the QoS scheduling.

4.7 SMART-K2.6: QoS strategies the transport for multi-service aggregation

Introduction

As mobile networks evolve towards more flexible and efficient architectures, the role of QoS in transport segments becomes a crucial topic. The ORAN architecture, with its disaggregated and modular approach, introduces new opportunities and challenges for QoS management. QoS in transport networks refers to the ability to deliver traffic with varying performance requirements, such as latency, throughput, and reliability, while efficiently utilizing network resources. With the advent of multi-service aggregation, where different types of traffic are consolidated and transported over a unified network infrastructure, the need for sophisticated QoS strategies becomes even more pronounced.

SoA

The state of the art in this field encompasses a variety of approaches aimed at optimizing resource utilization, minimizing latency, and ensuring high reliability across a spectrum of services.

- Traffic Classification and Prioritization: One of the fundamental strategies in QoS management involves the classification and prioritization of traffic based on its characteristics and requirements.

This approach allows network operators to allocate resources more efficiently, ensuring that critical services receive preferential treatment over less time-sensitive applications.

- Service Level Agreements (SLAs) and Traffic Engineering: SLAs define the performance parameters guaranteed by the network operator to the end-users or service providers. Traffic engineering techniques, such as traffic shaping, route optimization, and bandwidth reservation, are employed to enforce SLAs and maintain desired QoS levels. These strategies enable proactive management of network resources, mitigating congestion and ensuring consistent service delivery across diverse traffic types.
- Dynamic Bandwidth Allocation: With multi-service aggregation, the demand for bandwidth varies dynamically across different services and applications. Dynamic bandwidth allocation techniques, such as Quality of Service-aware routing and traffic grooming, adaptively adjust resource allocation based on real-time traffic conditions and QoS requirements. By dynamically allocating bandwidth to accommodate peak demands and fluctuations in traffic patterns, these strategies enhance network efficiency and responsiveness.

Keeping this in mind, the focus of this key concept is on characterizing the system behavior from different QoS strategies for mapping the fronthaul traffic in an ORAN-based 6G disaggregated mobile network. It also encompasses an experimental assessment of these QoS marking strategies and their validation through measurement campaigns utilizing carrier-grade equipment. Additionally, the research explores the impact of combining multiple traffic types in fronthaul scenarios of multi-service aggregation. This key concept will go through several tasks:

- Theoretical Analysis of QoS Marking Strategies, developed by UNICAN: various theoretical models of QoS marking strategies [77] [78] are presented and analyzed. The discussion delves into the benefits and drawbacks of different approaches and their suitability for specific traffic types. The emphasis is on understanding how these strategies affect fronthaul traffic when combined in a multi-service aggregation environment. The analysis of QoS management in the network plays a crucial role, as it will inform the definition of effective strategies and the development of guidelines for implementation.
- System Behavior Characterization, developed by UNICAN: Here, the system behavior resulting from different QoS strategies is characterized through simulations and mathematical models, specifically by using ns-3 [79]. The goal is to identify the impact of QoS strategies on different traffic types and their overall effect on the transport aggregation network. The scenario will include a first node in charge of marking traffic by using the DSCP [80] field in the IP header. Then, a second node will apply QoS strategies based on commercial routers.
- Testing on real equipment: the TID laboratory will be the center of real equipment testing. With access to several carrier-grade equipment, it will prove the theoretical calculations in a real-world environment.
    - Firstly, there will be tests on isolated nodes. These tests will include vendors such as Cisco, Juniper, and others. In this way, it will be possible to demonstrate how the QoS strategies developed affect traffic depending on the router on which they are implemented.
    - Subsequently, the developed QoS strategies will be tested with a topology closer to a production implementation. This topology will be built with the equipment that provided better results in the previous tests, taking into account the limitations of the number of routers available in the laboratory. The reference scenario to be replicated is the one proposed by

Juniper [81] (Figure 5) in which one of the routers applies the marking, and another one applies the QoS strategies.

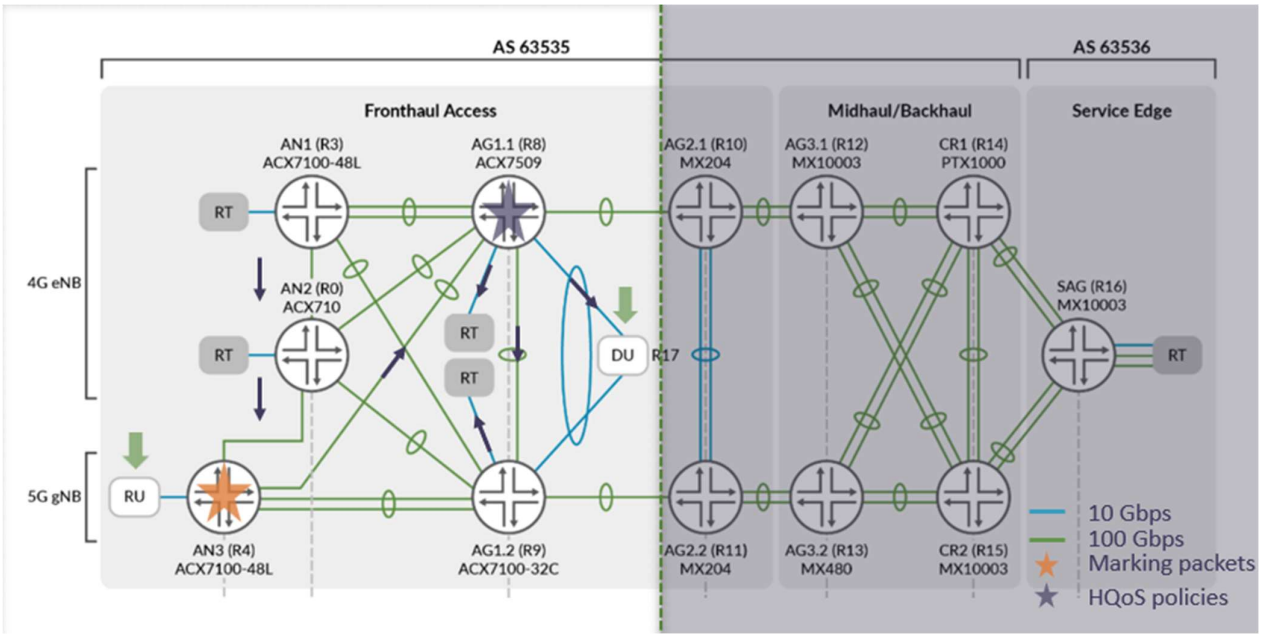

FIGURE 5. JUNIPER VALIDATED DESIGN - 5G FRONTHAUL SERVICES TOPOLOGY

Beyond SoA

Our research extends beyond the state of the art, introducing novel contributions and insights that pave the way for enhanced network performance and efficiency. The key innovations and contributions include:

- Tailored QoS Marking Strategies for Fronthaul Traffic: Building upon existing QoS marking strategies, our research delves into the development of tailored approaches specifically designed for fronthaul traffic in ORAN-based architectures. By leveraging insights from theoretical analysis and system behaviour characterization, we identify optimal QoS marking schemes that strike a balance between resource efficiency and service differentiation.
- Real-world Validation and Testing: Going beyond theoretical analyses, our research conducts extensive validation and testing of QoS strategies on real carrier-grade equipment in the TID laboratory. Through rigorous experimentation and measurement campaigns, we validate the effectiveness and practical applicability of our proposed QoS mechanisms in real-world scenarios, ensuring their readiness for deployment in production environments.

5 References

[1] X. Li, A. Garcia-Saavedra, X. Costa-Perez, C. J. Bernardos, C. Guimarães, K. Antevski and D. R. López, "5Growth: An end-to-end service platform for automated deployment and management of vertical services over 5G networks.," *IEEE Communications Magazine,* pp. 84-90, 2021.

[2] GSMA, "Generic Network Slice Template," 17 June 2022. [Online]. Available: https://www.gsma.com/newsroom/wp-content/uploads//NG.116-v7.0-1.pdf.

[3] ETSI, "Management and orchestration, 5G Network Resource Model (NRM)," [Online]. Available: https://www.etsi.org/deliver/etsi_ts/128500_128599/128541/16.08.00_60/ts_128541v160600p.pdf.

[4] O. S. J. P.-R. ,. R. A. R. Ferrus, "On 5G Radio Access Network Slicing: Radio Interface Protocol Features and Configuration," *IEEE Communications Magazine,* vol. 56, no. 5, pp. 184-192, 2018.

[5] ORAN Test & Integration Focus Group, "O-RAN End-to-end Test Specification," February 2024. [Online]. Available: https://specifications.o-ran.org/specifications.

[6] D. Wypior, M. Klinkowski and I. Michalski, "Open RAN—Radio Access Network Evolution, Benefits and," National Institute of Telecommunications, Warsaw, Poland, 2022.

[7] P. Fetterolf, "The Economic Benefits of Open RAN Technology," AGC Research, 2021.

[8] E. Amiri, N. Wang, M. Shojafar and R. Tafazolli, "Optimizing Virtual Network Function Splitting in Open-RAN Environments," GIC and 6GIC, Institute for Communication Systems (ICS), Guildford, United Kingdom.

[9] AMD, "Zynq UltraScale+ MPSoC Processing System v3.5, LogiCORE IP Product Guide, PG201 (v3.5)," [Online]. Available: https://docs.amd.com/r/en-US/pg201-zynq-ultrascale-plus-processing-system.

[10] graphcore, "Graphcore," [Online]. Available: https://www.graphcore.ai/products/ipu.

[11] Cerebras, "Achieving Industry Best AI Performance Through A Systems Approach," [Online]. Available: https://cerebras.net/wp-content/uploads/2021/04/Cerebras-CS-2-Whitepaper.pdf.

[12] Habana, "Intel®Gaudi® AI accelerator," [Online]. Available: https://habana.ai/products/gaudi/.

[13] Amazon web services, "Trainium," [Online]. Available: https://aws.amazon.com/machine-learning/trainium.

[14] M. Qasaimeh, K. Denolf, J. Lo, K. Vissers, J. Zambreno and P. H. Jones, "Comparing Energy Efficiency of CPU, GPU and FPGA Implementations for Vision Kernels," in *EEE International Conference on Embedded Software and Systems (ICESS)*, 2019.

[15] E. K. e. al, “An Evolutionary Edge Computing Architecture for the Beyond 5G Era,” in *EEE 28th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)*, 2023.

[16] N. Bartzoudis, J. R. Fernández, D. López-Bueno, A. R. Villarroel and A. Antonopoulos, “Agile FPGA Computing at the 5G Edge: Joint Management of Accelerated and Software Functions for Open Radio Access Technologies,” *Electronics,* vol. 13, no. 4, 2024.

[17] D. Chiou, “Heterogeneous Computing and Infrastructure for Energy Efficiency in Microsoft Data Centers,” in *Proceedings of the 2016 International Symposium on Low Power Electronics and Design*, 2016.

[18] F. Chen, Y. Shan, Y. Zhang, Y. Wang, H. Franke, X. Chang and K. Wang, “Enabling FPGAs in the cloud,” in *11th ACM Conference on Computing Frontiers*, 2014.

[19] S. D'Oro, M. Polese, L. B. H. Cheng and T. Melodia, “dApps: Distributed Applications for Real-Time Inference and Control in O-RAN,” *IEEE Communications Magazine,* vol. 60, no. 11, pp. 52-58, 2022.

[20] M. Polese, l. Bonati, S. D'Oro, S. Basagni and T. Melodia, “Understanding O-RAN: Architecture, Interfaces, Algorithms, Security, and Research Challenges,” *IEEE Communications Surveys & Tutorials,* vol. 25, no. 2, pp. 1376-1411, 2023.

[21] O-RAN Alliance WG4, Control, User and Synchronization Plane Specification R003-v12.00, 2023.

[22] Linux Foundation, “DPDK programmer's guide,” [Online]. Available: https://doc.dpdk.org/guides-24.03/prog_guide/index.html.

[23] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan and Y. Cao, “ReAct: Synergizing Reasoning and Acting in Language Models,” [Online]. Available: https://arxiv.org/abs/2210.03629.

[24] T. Taleb, A. Boudi, L. Rosa, L. Cordeiro, T. Theodoropoulos, K. Tserpes, P. Dazzi, A. I. Protopsaltis and R. Li, “Toward Supporting XR Services: Architecture and Enablers,” *IEEE Internet of Things Journal,* vol. 10, no. 4, pp. 3567-3586, 2023.

[25] W. Azariah, F. Bimo, C.-W. Lin, R.-G. Cheng, N. Nikaein and R. Jana, “A Survey on Open Radio Access Networks: Challenges, Research Directions, and Open Source Approaches,” *Sensors,* vol. 24, no. 3, p. 1038, 2025.

[26] O-RAN Alliance, "O-RAN architecture and specifications," [Online]. Available: https://orandownloadsweb.azurewebsites.net/specifications. [Accessed 16 April 2024].

[27] T. X. Vu, H. D. Nguyen, T. Q. Quek and S. Sun, “Fronthaul compression and optimization for cloud radio access networks,” in *IEEE International Conference on Communications*, Kuala Lumpur, 2016.

[28] T. X. Vu, H. D. Nguyen, T. Q. S. Quek and S. Sun, “Adaptive Cloud Radio Access Networks: Compression and Optimization,” *IEEE Transactions on Signal Processing,* vol. 65, no. 1, pp. 228-241, 2017.

[29] L. Ramalho, I. Freire, C. Lu and M. B. a. A. Klautau, “Improved LPC-Based Fronthaul Compression With High Rate Adaptation Resolution,” *IEEE Communications Letters,* vol. 22, no. 3, pp. 458-461, 2018.

[30] S. -H. Park, O. Simeone, O. Sahin and S. Shamai, “Robust and Efficient Distributed Compression for Cloud Radio Access Networks,” *IEEE Transactions on Vehicular Technology,* vol. 62, no. 2, pp. 692-703, 2013.

[31] S. Lagén, L. Giupponi, A. Hansson and X. Gelabert, “Modulation Compression in Next Generation RAN: Air Interface and Fronthaul Trade-offs,” *IEEE Communications Magazine,* vol. 59, no. 1, pp. 89-95, 2021.

[32] S. Lagén, X. Gelabert, L. Giupponi and A. Hansson, “Fronthaul-Aware Scheduling Strategies for Dynamic Modulation Compression in Next Generation RANs,” *IEEE Transactions on Mobile Computing,* vol. 22, no. 5, pp. 2725-2740, 2023.

[33] S. Lagén, X. Gelabert, A. Hansson, M. Requena and L. Giupponi, “Fronthaul Compression Control for Shared Fronthaul Access Networks,” *IEEE Communications Magazine,* vol. 60, no. 10, pp. 36-42, 2022.

[34] 3GPP TR 38.801, “Study on new radio access technology: Radio access architecture and interfaces,” [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3056. [Accessed 16 April 2024].

[35] Frascolla, V., Dominicini, C. K., Paiva, M. H., Caporossi, G., Marotta, M. A., Ribeiro, M. R., ... & Both, C. B., “Optimizing C-RAN Backhaul Topologies: A Resilience-Oriented Approach Using Graph Invariants,” 2019.

[36] Li, Y., Mårtensson, J., Skubic, B., Zhao, Y., Zhang, J., Wosinska, L., & Monti, P., “Flexible RAN: Combining Dynamic Baseband Split Selection and Reconfigurable Optical Transport to Optimize RAN Performance,” 2020.

[37] P. Fetterolf, “The Economic Benefits of Open RAN Technology,” [Online]. Available: https://infohub.delltechnologies.com/section-assets/acg-the-economic-benefits-of-open-ran-technology. [Accessed 30 10 2023].

[38] “The evolution of Open RAN,” 5G Americas Whitepaper, February, 2023.

[39] "What is Open RAN?," Juniper, [Online]. Available: https://www.juniper.net/us/en/research-topics/what-is-open-ran.html. [Accessed 15 January 2024].

[40] V. Quintuna Rodriguez and F. Guillemin, “Towards the deployment of a fully centralized Cloud-RAN architecture”.

[41] Habibi, M. A., Nasimi, M., Han, B., & Schotten, H. D, "A Comprehensive Survey of RAN Architectures Toward 5G Mobile Communication System," Ieee Access, 7, 70371-70421, 2019.

[42] Gonzalez-Diaz, S., Garcia-Saavedra, A., De La Oliva, A., Costa-Perez, X., Gazda, R., Mourad, A., ... & Leithead, P., "Integrating Fronthaul and Backhaul Networks: Transport Challenges and Feasibility Results," IEEE Transactions on Mobile Computing, 20(2), 533-549, 2019.

[43] S. Bjφmstad, D. Chen and R. Veisllari, "Handling Delay in 5G Ethernet Mobile Fronthaul Networks," in *European Conference on Networks and Communications (EuCNC)*, 2018.

[44] Chih-Lin, I., Huang, J., Duan, R., Cui, C., Jiang, J., & Li, L., "Recent Progress on C-RAN Centralization and Cloudification," IEEE Access, 2, 1030-1039., 2014.

[45] O-RAN Alliance, "Open RAN Technical Priority Release 3," April 2023.

[46] ETSI, "TS 38.321 Medium Access Control (MAC) protocol specification," 2019.

[47] Y. Hao, F. Musumeci, J. Zhang, Y. Xiao, M. Tornatore and Y. Ji, "DU/CU Placement for C-RAN over Optical Metro-Aggregation Networks," April 2021.

[48] H. Yu, F. Musumeci, J. Zhang, Y. Xiao, M. Tornatore and Y. Ji, "DU/CU Placement for C-RAN over Optical Metro-Aggregation Networks," in *Optical Network Design and Modeling*, Springer, 2020, pp. 82-93.

[49] ETSI, "Physical layer procedures for data," [Online]. Available: https://www.etsi.org/deliver/etsi_ts/138200_138299/138214/16.02.00_60/ts_138214v160200p.pdf. [Accessed 08 April 2024].

[50] Alhajj, T., "Impact of centralized-radio access network architecture on 5G performance," Doctoral dissertation, Ecole nationale supérieure Mines-Télécom Atlantique Bretagne Pays de la Loire, 2023.

[51] H. Yu, F. Musumeci, J. Zhang, Y. Xiao, M. Tornatore and Y. Ji, "DU/CU Placement for C-RAN over Optical Metro-Aggregation Networks," February 2020.

[52] G. Berardinelli, S. R. Khosravirad, K. I. Pedersen, F. Frederiksen and P. Mogensen, "Enabling Early HARQ Feedback in 5G Networks," in *IEEE 83rd Vehicular Technology Conference (VTC Spring)*, 2016.

[53] Y. Mohammed, "Study of HARQ Techniques for Satellite Network Support of 5G Wireless System," Politecnico di Torino, 2020.

[54] A. Maeder and e. al., "Towards a flexible functional split for cloud-RAN networks," 2014.

[55] A. Garcia-Saavedra, X. Costa-Perez, D. J. Leith and a. G. Iosifidis, "FluidRAN: Optimized vRAN/MEC Orchestration," in *IEEE Conference on Computer Communications*, 2018.

[56] H. Gupta, A. A. Franklin, M. Kumar and a. B. R. Tamma, "Traffic-Aware Dynamic Functional Split for 5G Cloud Radio Access Networks," in *IEEE International Conference on Network Softwarization*, 2022.

[57] M. Ahsan, A. Ahmed, A. Al-Dweik and a. A. Ahmad, "Functional Split-Aware Optimal BBU Placement for 5G Cloud-RAN Over WDM Access/Aggregation Network," *IEEE Syst J,* vol. 17, no. 1, pp. 122-133, 2023.

[58] N. Sen and A. F. A, "Slice Aware Baseband Function Placement in 5G RAN Using Functional and Traffic Split," *IEEE Access,* vol. 11, pp. 35556-35566, 2023.

[59] J. García-Morales, M. C. Lucas-Estañ and J. Gozalvez, "Latency-Sensitive 5G RAN Slicing for Industry 4.0," *IEEE Access,* vol. 7, pp. 143139-143159, 2019.

[60] Y. Xu and e. al, "Joint Optimization of Base Station Sleeping, Functional Split and Routing Selection in Virtualized Radio Access Networks," in *IEEE Wireless Communications and Networking Conference*, 2023.

[61] F. Z. Morais and G. M. F. d. Almeida, "PlaceRAN: Optimal Placement of Virtualized Network Functions in Beyond 5G Radio Access Networks," *IEEE Transactions on Mobile Computing,* vol. 22, no. 9, pp. 5434-5448, 2023.

[62] M. Mushtaq, M. Golkarifard, N. Shahriar, R. Boutaba and a. A. Saleh, "Optimal Functional Splitting, Placement and Routing for Isolation-Aware Network Slicing in NG-RAN," in *International Conference on Network and Service Management*, 2023.

[63] A. M. Alba, J. H. G. Velásquez and W. Kellerer, "An adaptive functional split in 5G networks," in *IEEE Conference on Computer Communications Workshops*, Paris, 2019.

[64] I. Gomez-Miguelez, A. Garcia-Saavedra, P. D. Sutton, P. Serrano, C. Cano, D. J. L. Gomez-Miguelez, A. Garcia-Saavedra, P. D. Sutton, P. Serrano, C. Cano and D. J. Leith, "srsLTE: An open-source platform for LTE evolution and experimentation," in *Tenth ACM International Workshop on Wireless Network Testbeds, Experimental Evaluation, and Characterization*, 2016.

[65] C. C. Erazo-Agredo, L. Diez, R. Agüero, M. Garza-Fabre and J. Rubio-Loyola, "Enabling realistic experimentation of disaggregated RAN: Design, implementation, and validation of a modular multi-split eNodeB," *Computer Networks,* vol. 235, 2023.

[66] Software Radio Systems, "Open Source RAN," [Online]. Available: https://www.srsran.com/.

[67] G. Berardinelli, S. R. Khosravirad, K. I. Pedersen, F. Fred-eriksen and P. Mogensen, "Enabling Early HARQ Feedback in 5G Networks," IEEE 83rd Vehicular Technology Conference (VTC Spring), 2016.

[68] Yuta Imamura, Dairoku Muramatsu, Yoshihisa Kishiyama and Kenichi Higuchi, "Low latency hybrid ARQ method using channel state information before channel decoding," 23rd Asia-Pacific Conference on Communications (APCC), 2017.

[69] 3GPP, "TR 21.917 Summary og Rel-17 work items," 2023.

[70] Huawei, “Empowering Consumer-focused Immersive VR and AR Experiences with Mobile Broadband,” 2016. [Online]. Available: https://www.huawei.com/en/huaweitech/industry-insights/outlook/mobile-broadband/insights-reports/vr-and-ar.

[71] 3GPP, “System architecture for the 5G System,” [Online]. Available: https://www.etsi.org/deliver/etsi_ts/123500_123599/123501/17.05.00_60/ts_123501v170500p.pdf.

[72] L. Miuccio, D. Panno, P. Pisacane and S. Riolo, “A QoS-aware and channel-aware radio resource management framework for multi-numerology systems,” *Comput. Commun,* vol. 191, pp. 299-314, 2022.

[73] Z. Gu, C. She, W. Hardjawana, S. Lumb, D. McKechnie, T. Essery and B. Vucetic, “ Knowledge-assisted deep reinforcement learning in 5G scheduler design: From theoretical framework to implementation,” *IEEE J. Sel. Areas Commun,* vol. 39, no. 7, pp. 2014-2028, 2021.

[74] S. Barmpounakis, N. Maroulis, N. Koursioumpas, A. Kousaridas, A. Kalamari, P. Kontopoulos and N. Alonistioti, “AI-driven, QoS prediction for V2X communications in beyond 5G systems,” *Comput. Netw,* vol. 217, p. 109341, 2022.

[75] I.-S. Coma, R. Trestian, G.-M. Muntean and G. Ghinea, “ 5MART: A 5G SMART scheduling framework for optimizing QoS through reinforcement learning,” *IEEE Trans. Netw. Serv. Manag,* vol. 17, no. 2, pp. 1110-1124, 2020.

[76] A. Kaur and K. Kumar, “A reinforcement learning-based green resource allocation for heterogeneous services in cooperative cognitive radio networks,” *IEEE Trans. Netw. Serv. Manag,* vol. 19, no. 2, pp. 1554-1566, 2022.

[77] X. Xiao, “Commercial, Regulatory and Technical Challenges,” in *Networking,*, Morgan Kaufmann, 2008, pp. 77-136.

[78] NGMN, “Backhaul Evolution - Integrated QoS Management’,” [Online]. Available: https://www.ngmn.org/publications/backhaul-evolution-integrated-qos-management.html. [Accessed 26 10 2023].

[79] nsnam, “ns-3 | a discrete-event network simulator for internet systems,” [Online]. Available: https://www.nsnam.org/.

[80] F. Baker, J. Polk and M. Dolly, “A Differentiated Services Code Point (DSCP),” 2010. [Online]. Available: https://datatracker.ietf.org/doc/html/rfc5865.

[81] Juniper Networks, “5G Fronthaul Class of Service—Juniper Validated Design (JVD),” [Online]. Available: https://www.juniper.net/documentation/us/en/software/jvd/jvd-5g-fh-cos-02-02/about_this_document.html.