6GBLUR-JOINT: The Blurring RAN joint RAN and transport network control/orchestration mechanisms

JOINT

Grant No. TSI-063000-2021-57

E3: Final E2E 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 will describe 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), Albert Bel (CTTC), Josep Mangues-Bafalluy (CTTC), Rasoul Nikbakht (CTTC), Ana Larrañaga (CTTC), Sandra Lagén (CTTC), Engin Zeydan (CTTC), Miquel Payaró (CTTC), Biljana Bojovic (CTTC), Manuel Requena (CTTC), Selva Vía (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), Ramón Agüero Calvo (Unican), Luis Francisco Diez Fernandez (Unican), 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-063000-2021-57.

All information in this document is provided “as is" and no guarantee or warranty is given that the information is fit for any particular purpose. The user thereof uses the information at its sole risk and liability.

For the avoidance of all doubts, the MINECO has no liability in respect of this document, which is merely representing the authors view.

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

SLA: Service Level Agreement

SLO: Service Level Objectives

SMF: Session Management Function

SoA: State-of-the-art

SW: Software

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

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

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

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-**JOINT** aims (i) to design the 6G disaggregated mobile network, from RAN to core, according to service-based approaches in a way that it can be treated as any other virtual service that requires some specific hardware support, and (ii) to design and to validate of joint RAN and transport orchestration mechanisms that adapt to the needs of the disaggregated 6G network.

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, among other advancements, KPM monitoring for the E2 interface based on xApps running in the near-RT RIC, and the deployment of Digital Twin and NPN, which supports both autonomic close-loop strategies (embedded) and open-ended (co-operative) approaches.

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-JOINT:

- **K1.1 Data plane reconfiguration and Zero-Touch**: dynamic deployment of cloud-native mobile 5g core.
- **K1.2 Distributed deployment of e2e slices, including O-RAN**: dynamic deployment of cloud-native of the whole e2e mobile network.
- **K1.3 Non-public-network**: improve adoption of NPNs by means of using NDT to improve the deployment complexity and achieve and efficient management and optimization.
- **K1.4 Network Digital Twin**: design and implementation of digital model for orchestrating an NPN.
- **K2.1 Transport network optimization**: design and implementation of a transport slice controller to facilitate the provisioning of connectivity services in slice form.
- **K2.2 Entities' placement decisions**: GNN-based solutions for defining the different virtualized entities of the RAN.
- **K2.3 QoS policies and scheduling**: study of performance of QoS aware scheduling policies at the fronthaul network.
- **K3.1 Monitoring platform**: define novel cost-efficient monitoring solutions for RAN parameters in an O-RAN environment.
- **K3.2 AI/ML Platform (AIMLP)**: design an implement new features in an NDT to optimize NPNs, by means of using AI/ML techniques, applying closed-loop optimization strategies and offering open and accessible APIs to monitor and optimize the performance of NPNs.

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

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

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-**JOINT** aims to (i) design the 6G disaggregated mobile network, from RAN to core, using service-based approaches, enabling it to be treated as any other virtual service requiring specific hardware support, and (ii) design and validate joint RAN and transport orchestration mechanisms that adapt to the needs of the disaggregated 6G network. To achieve these objectives, multiple UCs and PoCs have been developed to experimentally validate the project's KCs within various frameworks, including the integration

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

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 KC 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-JOINT'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-JOINT's KCs related to WP2 and WP3, respectively.

1.1 JOINT Key Concepts

The 9 KCs in 6GBLUR-JOINT are summarized below. These KCs, under study in both WP2 and WP3, are aligned with the principal objective of the 6GBLUR-SMART project: to design and characterize mobile network components from RAN to Core, along with the orchestration mechanisms required to adapt to their disaggregation.

TABLE 1. JOINT KEY CONCEPTS.

Category	KC	Project	WP	Leader
K1. End-to-end orchestration for multi-domain disaggregated mobile networks	**K1.1** Data plane reconfiguration (UPF) and Zero-Touch	JOINT	2	CTTC
	K1.2 Distributed deployment of e2e slices, including O-RAN	JOINT	2	CTTC
	K1.3 Non-public-network	JOINT	2	ERICSSON
	K1.4 Network Digital Twin	JOINT	2	ERICSSON
K2. Backhaul and end-to-end transport management	**K2.1** Transport network optimization	JOINT	2	TID
	K2.2 Entities' placement decisions	JOINT	3	UNICAN
	K2.3 QoS policies and scheduling	JOINT	3	UNICAN
K3. Monitoring and AIMLP	**K3.1** Monitoring platform	JOINT	2	KEY
	K3.2 AI/ML Platform (AIMLP)	JOINT	2	ERICSSON

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

1.2 Mapping of Joint Key Concepts and PoCs

Following the primary objective of the 6GBLUR-JOINT project, the planned PoCs concentrate on designing disaggregated components and orchestration mechanisms aimed at optimizing the performance of 6G networks. Table 2 illustrates the relationship between PoCs and KCs, with a notable focus on optimizing RAN and transport components.

Most of these KCs will feature prominently in UC2 PoCs, which are dedicated to characterizing and analysing the feasibility of the proposed architecture across various network topologies under different scenarios (e.g., low, medium, high-capacity traffic). This characterization aims to delineate the boundaries of network deployments, with the objective of pushing theoretical limits. Such advancements are anticipated to impact network capital and operational expenditures by enhancing hardware and software elements and processes, achieving higher levels of centralization, and reducing the necessity for human intervention in network operations, all while maintaining Quality of Service (QoS) standards for users.

Additionally, the KCs relevant to UC3 PoC1 focus on investigating, designing, implementing, and experimentally verifying smart methods for optimizing RAN resource utilization. Specifically, these efforts target radio and computational resources essential for ensuring desired QoS levels in the studied Non-Public Networks (NPNs).

TABLE 2. MAPPING OF KEY CONCEPTS AND POCS

KC	PoC				
	UC1PoC1	UC1PoC2	UC2PoC1	UC2PoC2	UC3PoC1
K1.1 Data plane reconfiguration		x			
K1.2 Distributed deployment of e2e slices		x			
K1.3 NPN	x			x	x
K1.4 NDT	x			x	x
K2.1 Transport network optimization		x	x		
K2.2 Entities' placement decision			x		
K2.3 QoS policies and scheduling			x		
K3.1 Monitoring platform		x			
K3.2 AI/ML Platform	x			x	x

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

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.

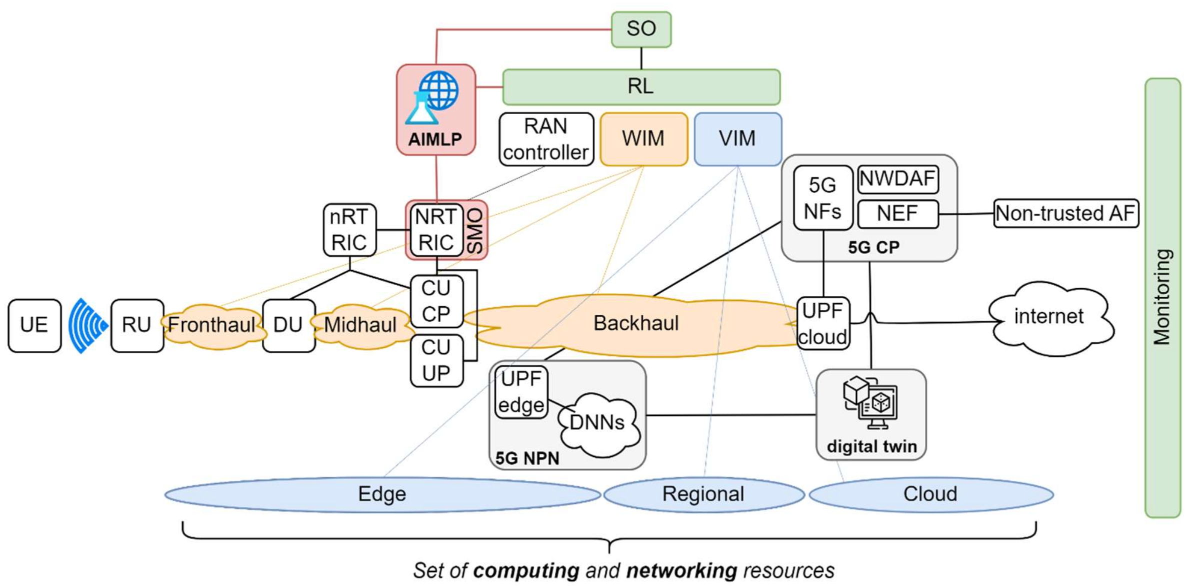

FIGURE 1. 6GBLUR ARCHITECTURE.

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**

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

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

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

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.

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

3 WP2-related Key Concepts

3.1 JOINT-K1.1: Data plane reconfiguration and Zero-Touch

Introduction

In today's rapidly evolving telecommunications landscape, the challenge of achieving real-time and on-demand creation, deletion, and configuration of User Plane Function (UPF) instances, alongside corresponding Packet Data Unit (PDU) sessions in cloud-native setups, looms large. As the demand for diverse and demanding use cases in 5G and beyond continues to grow, effective management of data planes becomes increasingly critical. Additionally, the integration of edge computing with 5G/6G, backed by standards such as ETSI Multi-access Edge Computing (MEC), emerges as a vital necessity, particularly for ensuring end-to-end support for latency-sensitive vertical applications. Consequently, it becomes imperative to thoroughly explore infrastructure and network considerations, including the relocation of server workloads and optimization of user data planes. An important aspect to note is 3GPP's facilitation of third-party vertical providers' access to User Equipment (UE) metrics and data plane capacities through the Network Exposure Function (NEF), highlighting the collaborative nature of these advancements.

SoA

In the literature, there are several recent examples dealing with the deployment of cloud-native 5G infrastructure using open-source software [2] [3] [4] [5]. In [2], the 5G core network is deployed as a single container in a Kubernetes cluster using Open Air Interface software. In [3] [4], the 5G core network, based on Open5GS software, is deployed as multiple containers in multiple nodes of the same Kubernetes cluster. However, the distribution of the different NFs of the mobile core is predefined and static for the given Kubernetes cluster, requiring adaptations for each environment. The same approach, a manual deployment with Docker of a Free5GC open-source cloud-native mobile core, is considered in [5]. Although [3] [5] deal with mobile core network deployments considering multiple User-Plane Functions (UPFs), which is not the case of [4], they are also deployed as a whole with the rest of control-plane NFs of the mobile core.

Besides, to enable edge services, two critical aspects must be addressed concerning infrastructure and network considerations. Firstly, from an infrastructure standpoint, it is essential to support workload relocation from cloud to edge, deploying servers closer to users when required. Secondly, from a network perspective, optimizing the user data plane to reach the edge server swiftly is paramount. As for the former, cloud-native solutions like Kubernetes facilitate the orchestration and reallocation of pods [6] according to service demands. As for the latter, in terms of network optimization, employing techniques such as user plane function (UPF) on-demand deployment [7] and selection is crucial. This ensures that user equipment (UE) sends and receives data via the most suitable interface based on the nature of the flow: latency-sensitive traffic utilizes dedicated UPFs situated near the edge and close to the edge application, while non-latency-critical data may utilize best-effort UPFs located in the cloud. Effective life-cycle management of both the infrastructure and networking contexts is then imperative to ensure optimal service delivery and performance.

Beyond SoA

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

As the first main contribution within this Key Concept, we have developed what we believe to be the first practical implementation of a cloud-native mobile core using Open5GS. This innovative setup connects seamlessly with two radio access networks, forming a versatile experimental platform for over-the-air transmission. By strategically centralizing control-plane functions and allowing user-plane functions to be distributed across various locations, we have achieved a flexible and automated deployment process. This flexibility not only allows us to leverage the benefits of edge locations, such as low latency and data privacy, but also enables us to quickly adapt to changing network needs or new locations. Our efforts have culminated in a tangible demonstration of progress, exemplified by the illustration of our experimental setup at CTTC (see Figure 2).

Furthermore, our research within the 6G-BLUR project extends beyond the development of a cloud-native mobile core. We have also explored the dynamic deployment of different parts of the network to support diverse services in various locations, particularly in scenarios involving private network usage. This capability, achieved through meticulous network design and the utilization of tools like OSM and Kubernetes, marks a significant step towards creating more flexible and adaptable 5G and 6G networks. These advancements not only lay the foundation for more complex network setups but also pave the way for smarter network management strategies in the future.

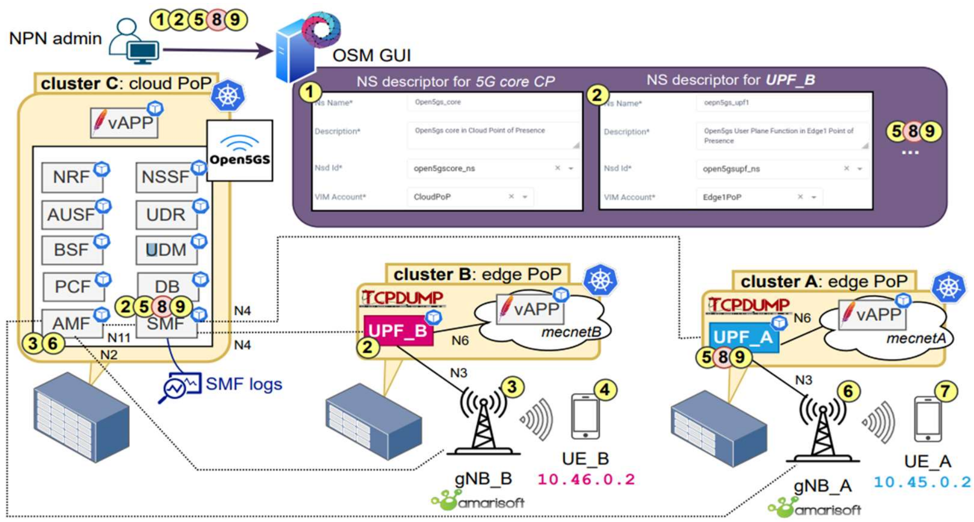

FIGURE 2. DISTRIBUTED 5G MOBILE CORE DEPLOYMENT DEMOSTRATED IN [7].

As for the second main contribution within JOINT-K1.1, our research efforts delve into the complexities of combining containerized workload relocation and UPF selection within 5G networks, with a focus on ensuring low-latency vertical applications. We have developed a proof of concept that integrates critical components, including over-the-air 5G New Radio (NR) transmissions, a containerized open-source 5G core, and a cloud-native platform powered by Kubernetes. This comprehensive setup enables end-to-end monitoring and orchestration of infrastructure, networking, and vertical lifecycle management. By emphasizing automation and interoperability, our work anticipates the evolving 6G landscape and seeks to address the challenges posed by dynamic network requirements.

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

Through our demonstration in [8], we demonstrate the feasibility of catering to low-latency vertical applications for User Equipments (UEs). Leveraging critical components such as over-the-air 5G NR transmissions, an open-source 5G core, and a Kubernetes-based cloud-native platform, our demonstration showcases comprehensive end-to-end orchestration. As we anticipate the transition to 6G networks, our work lays a solid foundation for future advancements in network management and service delivery.

3.2 JOINT-K1.2 Distributed deployment of e2e slices, including O-RAN

Introduction

Recent networking paradigms based on virtualization/cloudification and softwarization are key enabling concepts to provide flexibility and a great degree of automation to the deployment and configuration of e2e mobile networks. This can be applied from the radio access network segment to the core segment to accommodate and plug new stakeholders in the telco ecosystem, like the vertical industries. They are providing innovative use cases with stringent requirement, where a one-fit all deployment is no longer enough and seek for on-demand e2e deployment able to satisfy their requirements.

In addition to this, such deployments need to be orchestrated to make an efficient use of the underlying computer, memory and transport components based on its geographical distribution and resource availability. In addition to this, these deployments need to be flexible enough to allow the progressive and on-demand deployment of virtualised network entities, thus relying on the automated dimension of the process to reduce the human intervention. For such reasons, this key concept expands the previous one on 5Gcore data plane reconfiguration by applying the same principles to allow the distributed cloud-native deployment of the whole e2e mobile network considering also the RAN segment. More specifically, the new components defined in O-RAN architecture, like CU, DU, Near-RT Radio Intelligent Controller (RIC) and the on-demand addition of xApps.

SoA

Currently, literature shows several examples of open-source projects that is boosting the research in the deployment of e2e mobile network entities to plug vertical applications. Regarding the 5G core, we find Open5GS [9] and Open Air interface (OAI) [10]. Regarding the RAN segment, we can cite SRS RAN Project [11] (developed by a partner of 6G-BLUR) and OAI. There are other alternatives for the RAN segment, like UERANSIM [12], but this is only limited to simulation purposes, and do not align with O-RAN architecture, thus not providing CU/DU split, as the previous mentioned alternatives. Digging into the O-RAN architecture, there are different frameworks providing implementations of the Near-RT RIC and associated xApps like Flexric [13] and the one provided by the O-RAN alliance [14].

Such open-source frameworks are combined to form e2e slices and are progressively starting to embrace containerized deployments employing Docker ([15], [16]) as a first step to progress towards automated deployments. However, such deployments are most likely provided as a single-machine deployment based on a localhost network. As a result, these deployments may not operate accurately in distributed multiple-nodes or cloud environments, which require additional network configuration and the creation of proper network entities artefacts to achieve a proper disaggregation of network entities artefacts to really embrace the benefits of a cloud-native approach enabling flexible and distributed deployments.

Beyond SoA

To achieve the benefits of cloud-native vision, the first step is to automate the disaggregated deployment of an e2e mobile slice. 6G-BLUR JOINT envisages the creation of appropriate entities artefacts based on Kubernetes-based Helm chart definitions to expose the network entities of the mentioned open-source frameworks, like gNB, near-Real time RIC, across distributed Kubernetes clusters by means of Load Balancer types of service definition. Understanding such disaggregated deployments will allow on one hand, to evolve the software to be prepared to run in multiple machines (e.g., announcing proper IP addresses on the different interfaces assuming a deployment using containers) and on the other hand, the definition of complete management and orchestration procedures enabling dynamic configuration, optimization, and management of O-RAN deployments, also taking into account the provision of transport connectivity. The objective is to achieve a deployment scheme as presented in the figure below, which focus on the RAN part and the associated O-RAN entities[1]. This figure represents an example of Scenario C for O-RAN deployments, which are considered beyond current state of the art [17]. The aim through such disaggregated deployment is to increase the flexibility of deployment, allowing for instance the creation/termination of new instances of DUs connected to the RU or the dynamic addition of xApps to a near RT-RIC software. This process will be done progressively, e.g., first automating the deployment of a cloud-native gNB (joint CU/DU) element. Then, the distributed cloud-native deployment of CU and DUs entities as SRS partner is evolving its software towards the split of CU/DUs processes. Regarding O-RAN xApps, first we will work with FlexRic SW and the on-demand deployment of xApps monitoring different available parameters produced by the srsRAN software, like downlink and uplink throughput. Later, the idea is to exploit the O-RAN architectural framework, to perform close-loop operations based on xApps and perform dynamic actions on the O-RAN nodes to adapt to working conditions.

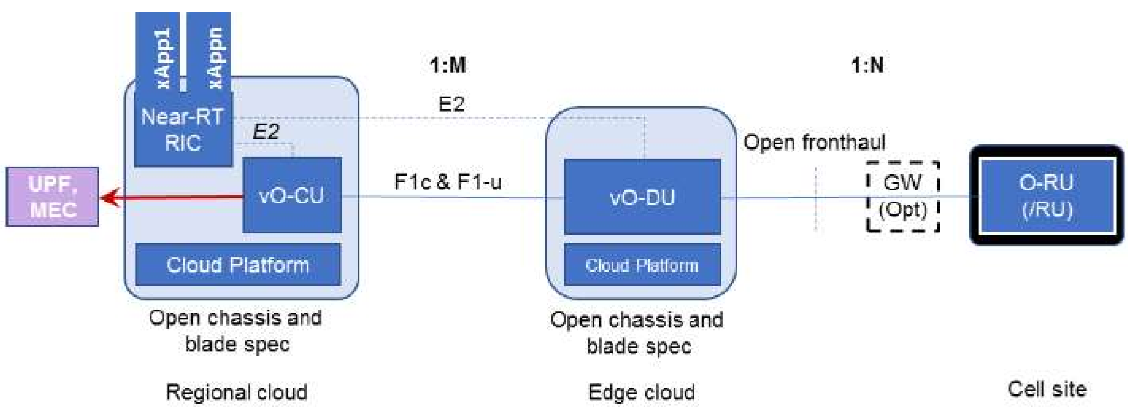

FIGURE 3. DISSAGREGATED AND DISTRIBUTED DEPLOYMENT OF O-RAN ENTITIES. BASED ON [17]

[1] It is worth remembering that the distributed deployment for the 5Gcore has been described in key concept K1.1.

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

3.3 JOINT-K1.3 Non-public-network

Introduction

The 5G system caters to a wider range of use cases than previous generations, including those requiring data privacy and security within an organization's boundaries. These non-traditional use cases often involve organizations that have not traditionally used mobile connectivity. To meet these needs, Non-Public Networks (NPNs) are deployed.

NPNs offer several performance benefits. Firstly, optimized coverage is achieved since the network is designed specifically for the owner's needs, not public use. Secondly, local network entities reduce latency by minimizing physical distance and network hops. Then, local network operation allows for internal management and resolution of disruptions. Beyond performance, NPNs provide the owner with greater control. They can enhance network security by restricting access and improving data privacy. Additionally, traffic prioritization can be implemented based on user types and organizational needs. Similarly, congestion and interference management can be optimized for the specific use case. It's important to note that these benefits may vary depending on the deployment architecture.

An NPN is a network designed for non-public use and 3GPP defines two main distinct categories for NPNs:

- **Standalone Non-Public Network (SNPN):** An NPN that doesn't rely on network functions from a Public Land Mobile Network (PLMN) operator. SNPN deployment requires significant investment and commitment from the owner.
- **Public Network Integrated Non-Public Network (PNI-NPN):** An NPN that leverages a PLMN for support. **PNI-NPNs can be deployed as private network slices** (dedicated portions of the PLMN network) or involve a mix of owner-deployed and PLMN-provided network elements.

A range of configurations exists between fully isolated SNPNs and PNI-NPNs using private network slices. 5G Alliance for Connected Industries and Automation (5G-ACIA) has defined multiple deployment scenarios for PNI-NPNs, however our concept focuses on the deployment option with shared RAN and control plane.

In this deployment scenario, both the Radio Access Network (RAN) and network control planes are handled by the public network. However, the NPN data remains within the organization's defined premises. This scenario can be implemented using Access Point Names (APNs) or network slicing, where a dedicated network slice is provided to the NPN owner.

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

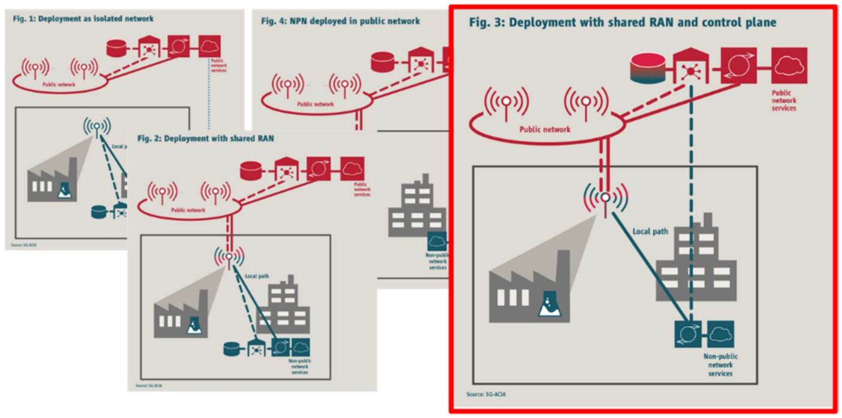

FIGURE 4. DIFFERENT PNI-NPN DEPLOYMENT SCENARIOS DEFINED BY 5G-ACIA

Now 5G network deployments are growing globally, and they are expected to reach increasingly varied types of businesses, users, and devices with a broad range of heterogeneous needs and demands on the network. That is especially the case for the envisaged scenario of the proliferation of PNI-NPNs that rely on 5G Public Networks and pervasive application of Network Slicing. **Network Slicing is, indeed, the key enabler for PNI-NPNs that ensures tailored network services, optimized resource allocation, and enhanced security for diverse use cases and business needs.**

As 5G Network deployments evolve, CSPs who offer and operate NPN services will have to dramatically stretch and scale their Network Slicing capabilities to provide tailored connectivity services to hundreds and then thousands of large, mid-sized, and small enterprises new paradigms will have to emerge for overcoming and managing the intrinsic complexity of **Massive Network Slicing**.

In this project, our aim is leveraging and evolving 5G standards for securing both interoperability and economics of scale of innovative solutions for enabling Massive Network Slicing, pursuing significant enhancements on the levels of **flexibility** and **automation** of Advanced 5G networks and towards their evolution to 6G. In the upcoming sections, we are going to expand these enhancements.

First, a high degree of flexibility for the distribution of network and computing resources is envisioned and encouraged to secure both a best fit of network performance levels to the demands of the subscribers and an optimal utilization of network resources for making Massive Network Slicing a sustainable business. A range of alternative scenarios for optimally distributing 5G Network Functions and leveraging Edge Computing across interconnected 5G Public Networks and Non-Public Networks shall be evaluated.

In the flexibility aspect, **NPN reconfiguration is an important key feature to consider since the demands on the network can change dynamically over time, and networks (or slices) must respond to these changes accordingly**. Imagine a slice for a factory – during peak production, it might need more bandwidth for data transfer. Reconfiguration allows the network to allocate those extra resources within the slice itself, optimizing performance without impacting other slices. **NPN Network Digital Twin holds immense potential to provide such optimization services as it can monitor the NPN, and collect data from it, and after analyzing the collected data it can suggest an optimized NPN configuration that secures to meets the performance objectives of the service that running on NPN**. Moreover, dynamic Quality of Service (QoS) allocation and management mechanisms can be applied to individual slices or users to adapt to fluctuating

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

demands on the network. The Policy Control Function (PCF) acts as the network's key enabler for this functionality. Furthermore, exposing this capability to third-party applications further enhances flexibility, allowing for even more dynamic control of the resources in the network.

Automation is crucial for reducing the lead times for new network deployments, and Zero-Touch approaches shall be embraced. The proposed vision concept for Zero-Touch NPN revolves around four innovation breakthroughs aiming at dramatically accelerating the rate of adoption of 5G in Enterprise domains:

- Self-integration of NPN with Public Network,
- Self-integration of NPN with Edge Computing environments,
- Self-configuration of an NPN 5G System according to intent-based declaration of target KPIs, and
- Autonomous OAM mechanisms for monitoring self-integration status and SLA of the NPN 5G System.

So, in essence, a 5G NPN armed with Zero-Touch superpowers will automatically connect and adapt to the Public Network and Edge environment and will self-configure to meet expected performance levels, all of it happening at NPN start-up time. In support of this ambition, Network Digital Twin technologies shall be leveraged and evolved. Below two different diagrams are given the explain the usage of zero-touch automation procedures:

- Performing initial configuration of NPNs during initial boot-up
- Performing reconfigurations during AI/ML-based continuous optimization

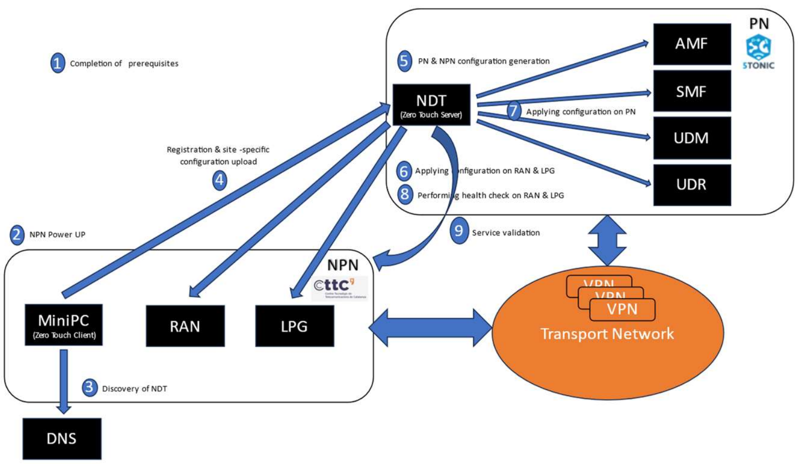

FIGURE 5. ZERO-TOUCH INITIAL CONFIGURATION PROCEDURE

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

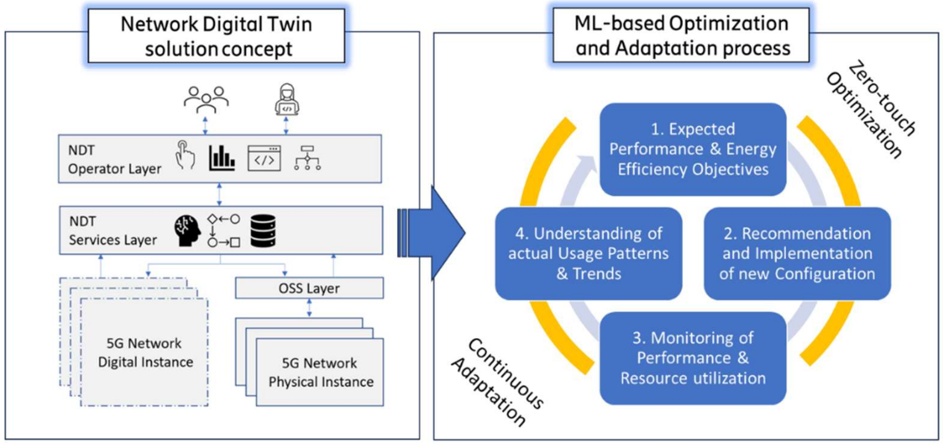

FIGURE 6. ZERO TOUCH RECONFIGURATION OF PNI-NPNS USING NDT

SoA

The standardization of NPN in 3GPP started in Release 15. The SMARTER work item, which aimed to specify the requirements to enable the new services and markets, covered the concept of a private network in its target specification 3GPP TS 22.261 [18].

Release 15 of 3GPP specifications supports only a subset of NPN technology. Specifically, the specifications focused on the concept of private networks and included a few requirements to support private networks. Therefore, NPN based on Release 15 would be deployed as a standalone network for the owner without interaction with the public network.

In Release 16, the concept of NPN was formally introduced into normative specifications. The requirements for NPN, including security and privacy, were specified. Enhancements were introduced to the 5G system to support non-public networks; namely, **network identification, discovery, selection, and access control for NPNs**.

From Release 17, the 3GPP work focused on enhancements to NPN to fully cater to the defined requirements. The first major enhancement was supporting an **UE to select a particular SNPN** when the UE has credentials owned by a separate entity (called Credentials Holder: CH), whether it is a separate SNPN or other PLMN. The second major enhancement was supporting **UE's onboarding and remote provisioning for the NPN**. This enabled supporting onboarding and remote provisioning for PNI-NPNs as well. The third major enhancement came from the support of regulatory requirements for NPNs to define the **support of public safety services such as emergency calling and public safety warnings on NPNs**. Furthermore, the management aspect of NPN was specified. A dedicated specification for the management of NPN was created, which included the requirements for NPN management, deployment scenarios for NPN management, and NPN management models.

With Release 18 and further, standards continue to evolve to investigate potential requirements and solutions. Some of the studies are to support non-3GPP access (e.g., WLAN), to support mobility between NPNs without new network selection and to provide access to local services. Another important aspect being studied is charging. Various charging scenarios and requirements are being studied, along with potential solutions to resolve these. Finally, continuous study is ongoing to enhance the management and security of NPNs. The management study includes end-to-end network management, including RAN and CN (Core Network) domains, in NPN scenarios, cooperation between 3GPP and non-3GPP management systems, and management of capability exposure for various types of NPNs. From the security perspective, the security implications of the new NPN features are being studied. For example, support for **mobility between SNPNs, support for non-3GPP, and enabling of localized services for NPNs**.

Apart from the 3GPP standardization activities, the following emerging use cases and applications would benefit from the mass adoption of NPNs:

- **Industry 4.0**: Smart manufacturing, automated logistics, and real-time machine monitoring benefit from NPNs' high bandwidth and low latency.
- **Smart Cities:** Traffic management, connected infrastructure, and environmental monitoring can leverage NPNs for efficient data collection and control.
- **Energy and Utilities:** NPNs can optimize energy distribution, enable remote monitoring of grids, and support smart metering solutions.
- **Healthcare:** Remote surgeries, real-time patient monitoring, and connected medical devices can benefit from NPNs' reliability and security.

While NPNs, and PNI-NPNs, offer exciting possibilities, they are not without their challenges. The following challenges need to be addressed to allow mass adoption.

- **Deployment complexity**: Integrating automation into PNI-NPN relevant procedures is the key to simplifying the deployment, operation, and management of PNI-NPNs. Defining and using zero-touch network automation tools holds a great deal of potential to address this challenge.
- **Efficient management & optimization**: Rather than following traditional manual network optimization procedures, using close-loop AI/ML-based optimization techniques might provide better and more effective optimization results. **Bringing and applying the digital twinning concept to PNI-NPNs is a great example and it is an innovative approach** (see the diagram below). NPN Network digital twin can continuously monitor and collect data from NPN, then use this data to train its AI/ML integrated optimization engine and finally generate alternative optimized configurations. Zero-touch reconfiguration methods can be used to apply the generated configuration. Closed-loop iteration can continue until the most optimized configuration is applied.

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

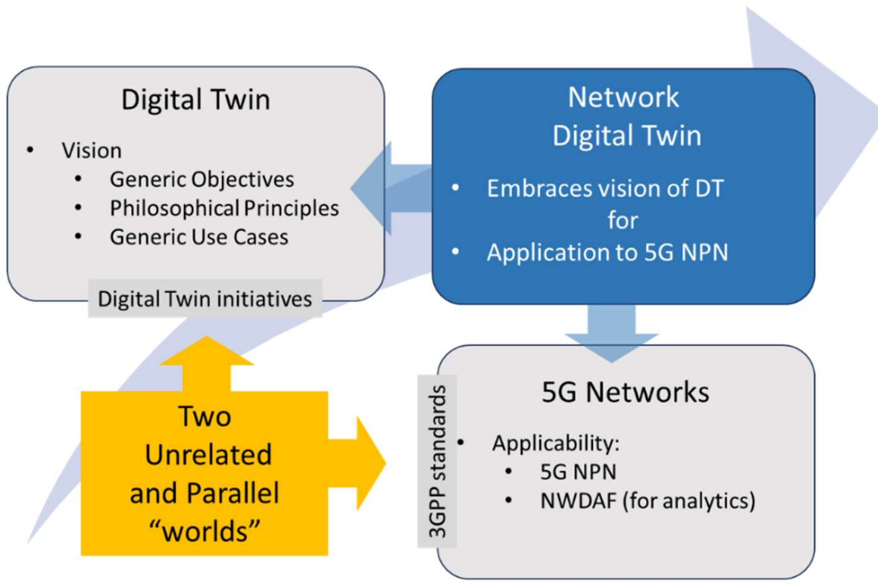

FIGURE 7. APPLYING DIGITAL TWIN CONCEPTS TO NPNS

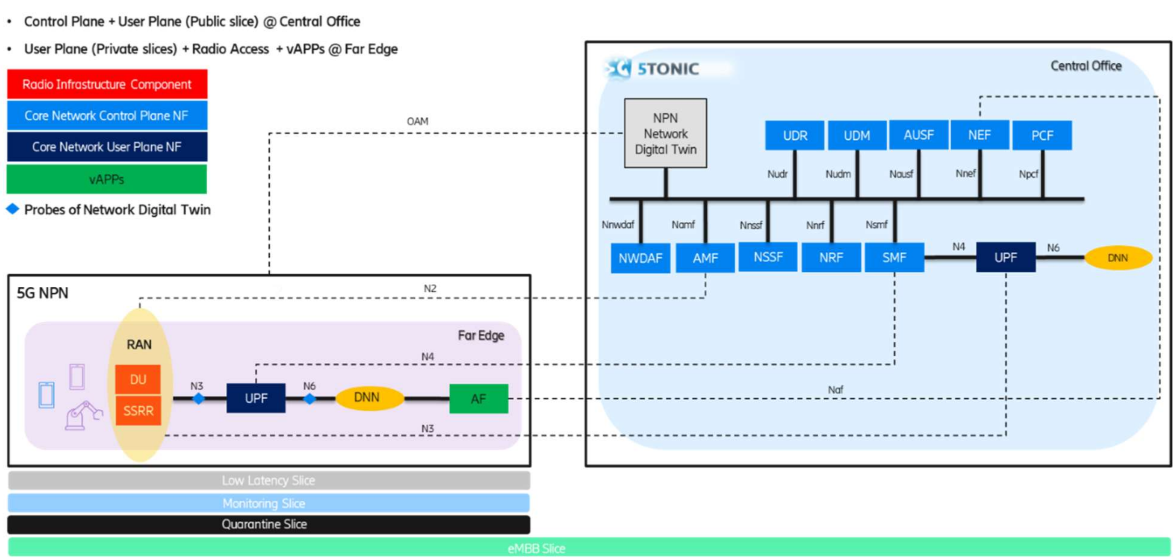

FIGURE 8. PNI-NPN BLUEPRINT ARCHITECTURE WITH NETWORK DIGITAL TWIN

Figure 8 shows the blueprint 3GPP PNI-NPN architecture:

- The central office will be constituting the "Public Network (PN)" part. It holds:
  - All 5GC control plane NFs.
  - One 5GC user plane NF for serving eMBB public slice.
- Enterprise Premises (Far Edge) will be constituting the "Non-Public Network (NPN)" part. It holds:
  - 5GC user plane NF for serving private slices (Low latency, Monitoring, and Quarantine).

- RAN for serving all slices including the eMBB public slice.

It should be noted that this architecture is defined as a blueprint to cover most relevant NPN scenarios by using generic types of slices, therefore the slices depicted in the architecture can be customized based on vertical use cases or enterprise requirements. **PNI-NPNs can be thought of as extensions to the public CSP network and network slicing mechanisms are the main enabler to deploy PNI-NPNs.** PNI-NPNs might have multiple private network slices that provide isolation, service characteristic differentiation, and resource allocation/distribution. There are a couple of deployment models for PNI-NPNs, and the figure above shows one of them in which RAN and Control Plane network functions are shared between public and non-public networks whereas a dedicated UPF is used on non-public networks.

Network Digital Twin comes into play to provide services to fulfill the need of managing and operating PNI-NPNs in an efficient, automatized thus cost-optimized way. The NDT platform described here is precisely and entirely focused on twining the gNB of, specifically, PNI-NPNs. NDT, in that specific applied context, can be a valuable tool for understanding and managing key aspects of performance and configuration of PNI-NPNs, such as:

- **Simulating network scenarios:** NDT can be used to test different PNI-NPN configurations and predict their impact on performance before real-world deployment.
- **Monitoring and optimizing network performance:** NDT can continuously monitor the PNI-NPN, providing insights for optimizing resource allocation, identifying potential issues, and ensuring service quality.

NDT is not a standardized element within 3GPP PNI-NPN specifications, and there is no standardization work ongoing or expected in that regard. So, to leverage the DT technology trend for bringing the services and benefits exposed to PNI-NPN scenarios a strategy for integration of NDT in PNI-NPN scenarios is needed.

The proposed strategy, as represented in Figure 8, involves the NDT platform being i) connected to the NPN instances (on the left-hand side of the picture) to be twinned, via OAM interfaces, and ii) integrated into the PN domain (on the right-hand side of the picture) as part of the 5GC Service Based Architecture. That duality enables, accordingly, in the first place, privileged access to the NPN systems for remote monitoring and actuation, and, in the second place, smooth integration of a cost-effective central NDT platform within the PN instance, and the rest of the centrally deployed NFs, as needed.

The specific use cases that will be implemented as part of the 6GBLUR project in the context of the concept are:

- Initial configuration of PNI-NPN using zero-touch procedures of NDT
- NPN performance optimization using NDT

Implementation details of each scenario are provided in the sub-sections below.

Initial configuration of PNI-NPN using zero-touch procedures of NDT

As the number of NPN deployments increases in the Communications Service Provider (CSP) networks, there shall be a need to have automated and cost-effective procedures for efficient integration, operation, and maintenance activities of NPNs. Embracing Zero-Touch automation approaches shall be crucial for reducing the lead times for launching new NPN services and operating them after their initial launch. Zero-touch automation can significantly enhance the integration and operation of NPNs in several ways including enabling faster deployment times, cost-efficient management of PNI-NPNs, reducing human errors, and decreasing integration and operational costs. Our ambition is to **demonstrate that an NPN equipped with Zero-Touch capabilities will automatically connect and adapt to the public network and edge environment and will self-configure for meeting expected performance levels – all of it happening at NPN start-up time.**

There are two main components of the zero-touch NPN initial configuration procedure:

- **NDT:** It acts as a zero-touch server and orchestrates the whole process. It generates the NPN configuration by using the inputs from the zero-touch client (explained below) and applies the generated configuration on NPN components (RAN & UPF) so that NPN is attached to the Public Network. Finally, NDT performs basic health check procedures and service validation tests.
- **Mini-PC:** It holds a lightweight zero-touch client that only takes part in the discovery of NDT, registration of NPN on NDT, and providing NPN-specific configuration inputs to NDT.

Assuming that prerequisite steps are completed, the zero-touch NPN initial configuration procedure starts by powering on the NPN. Once all the components are up, the zero-touch client will discover NDT and initiate the registration procedure by providing a unique NPN identifier and inputs for NPN configuration. Then zero touch server on NDT accepts the registration request and generates the necessary configuration of each NPN component (RAN & LPG) by using the inputs provided by the client. In the next step, NDT applies the configurations on NPN components and performs a basic health check and internal service validation test. After completing this step, NPN is ready for service.

NPN performance optimization using NDT

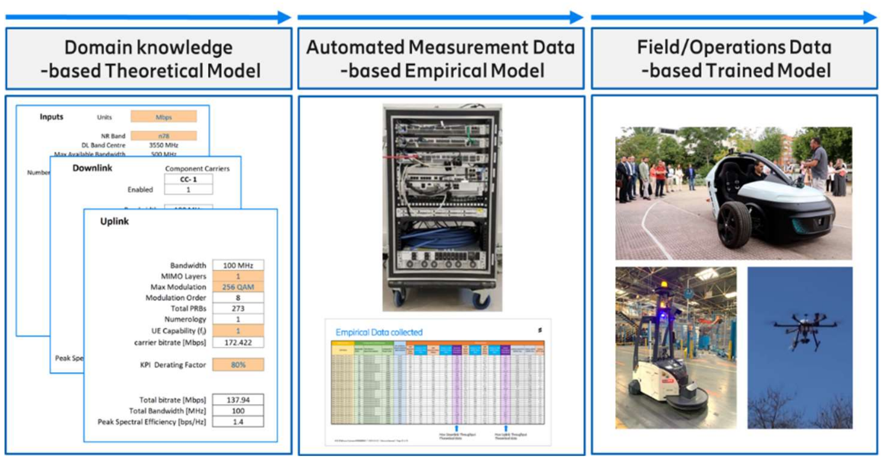

FIGURE 9. FROM DOMAIN KNOWLEDGE MODELLING TO MODEL TRAINING FOR NDT

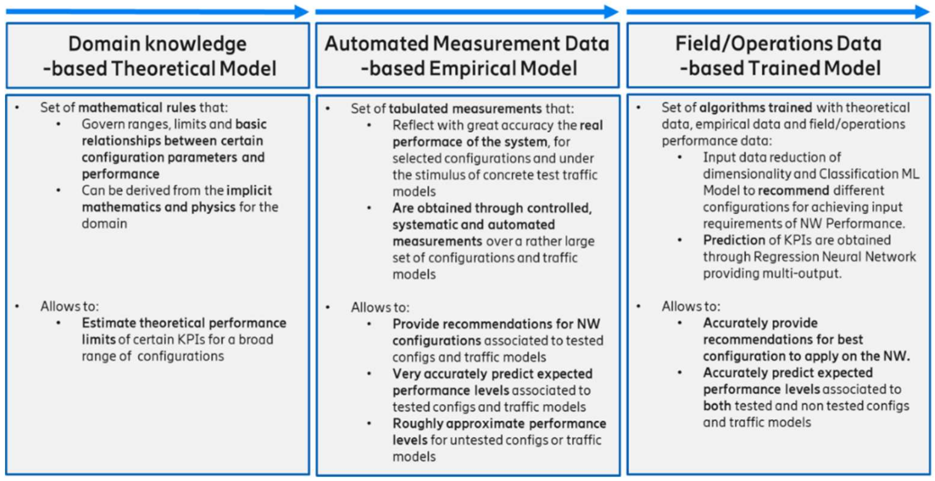

FIGURE 10. THEORETICAL, EMPIRICAL AND TRAINED MODELS CONTINUUM

As illustrated above, the NDT platform is proposed to be fed with i) a domain knowledge model, ii) an empirical model and iii) a data-driven trained model. That combined strategy allows for respectively i) dealing with expectable ranges for performance, ii) predicting performance levels for multiple potential configurations and finally iii) accurately predicting performance for the system in its current configuration, condition, and usage, based on training.

Within this scope and context, the Network Digital Twin overall abstract model represented in Figure 11 is characterized by a set of complementary aspects of the physical system and environment of a specific NPN system, which the corresponding NPN digital instance is aware of, namely, Spectrum Assignment, Spectrum usage technique, Radio Unit type, gNB configurations, air conditions, KPIs and Traffic Models.

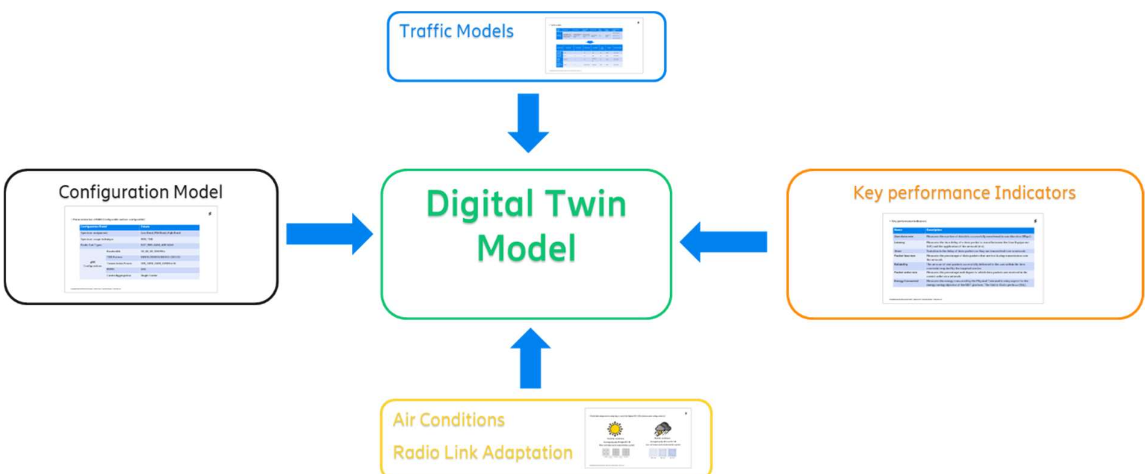

FIGURE 11. THE NETWORK DIGITAL TWIN MODEL

When it comes to performance optimization related use cases supported by the NDT platform they involve i) Configuration proposal for Physical Twin based in service requirements, that delivers rank ordered list of feasible configurations that match the requirements of user service, ii) Configuration implementation in the

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

Physical Twin based in service requirements, securing the actual implementation and necessary reconfiguration of the system accordingly, and iii) Performance comparison between Physical and Digital Twin performance, providing rich information about how the service is running vs expectations.

The performance optimization capabilities of NDT will be showcased through two separate PoCs. In the first PoC, an empirical model built on automated measurement data will be employed on NDT. The optimization will be one-time. For the second PoC, ML trained model based on field or operational data will be used. There will be a semi-close loop implemented between NDT and NPN, so that monitoring and optimization will be a continuous flow. Below use case examples are provided for each PoC scenario to make them easier to understand the concept.

Example NPN performance optimization using empirical model:

- A new service is to be launched on NPN. Service owner wants to make sure that NPN has the right configuration to provide the expected performance. Traffic model and target traffic KPI levels are provided to NDT as an input by the service owner.
- By using its empirical model and inputs provided in the previous step, NDT predicts and provides ranked ordered list of NPN configurations with their respective target KPI fulfillment levels.
- Service owner analyzes the returned NPN configuration alternatives by considering the KPI fulfillment levels vs current service requirements and selects one of the NPN configurations to be applied.
- Service owner uses NDT to activate the selected NPN configuration on the physical NPN system.
- NDT also provides a service to compare the performance of the digital replica and physical NPN system. This service can be used to assess the accuracy of the performance optimization (predicted NPN performance vs actual NPN performance).

Example NPN performance optimization using ML trained model:

- An existing service on NPN is generating traffic continuously.
- NDT monitors the NPN, collects data (traffic data, metrics) from it to train ML model that will adapt existing NPN configuration to improve traffic KPIs (latency, throughput).
- NDT repeats the previous step until reaching the most optimized NPN configuration state (no further optimization on traffic KPIs is possible).

Figure 12 shows NPN performance optimization using ML trained model structure.

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

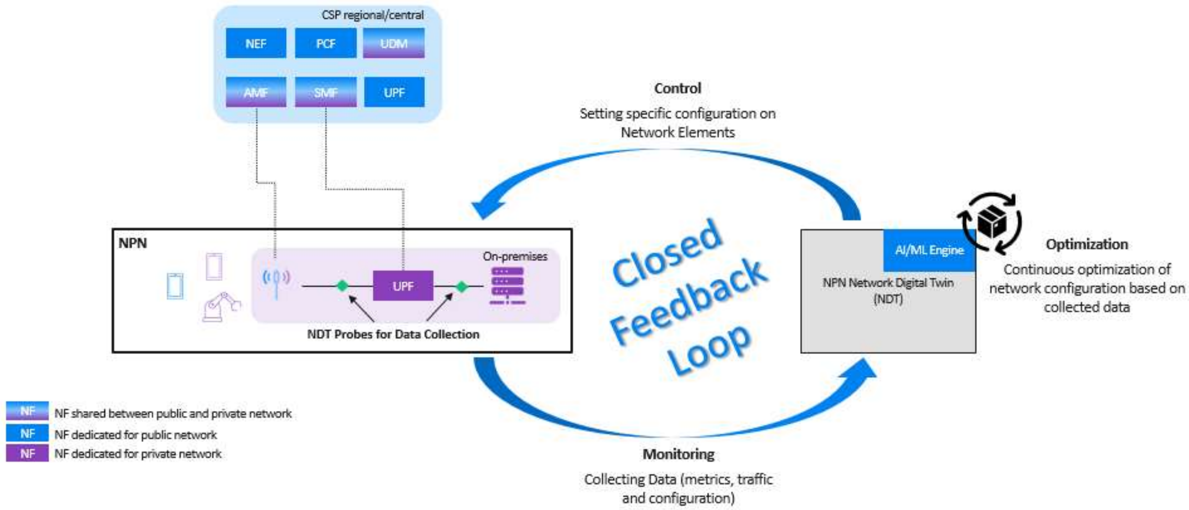

FIGURE 12. NPN PERFORMANCE OPTIMIZATION USING ML-TRAINED MODEL

Beyond SoA

We believe that PNI-NPNs hold great potential to revolutionize how various industries operate. However, there are a couple of challenges that must be addressed to remove the barriers to wide adoption:

- Deployment complexity
- Efficient management & optimization

Our vision is to bring and apply the Digital Twining approach to PNI-NPNs and benefit from its automation and optimization capabilities to address these two challenges.

The key technological aspects enclosed in the NDT concept being researched that go beyond the State-of-The-Art are:

- It acts as the main orchestrator for zero-touch configuration activities and holds the following key features:
    - Discovery and registration of NPNs
    - Generating NPN component initial configurations from a low-level design template
    - Applying the configurations on NPN components
    - Performing basic health checks and service validation tests
- It implements the full set of digital twin functionalities from performance monitoring and analysis to recommendations and actuation -i.e., this NDT platform provides with a **comprehensive one-stop shop innovative Network Digital Twin service developed ahead of eventual standardization** in the mid-term.

- It specifically **addresses PNI-NPN scenarios**, modelling NPN technology and performance vs the typical usage patterns and environmental conditions applying to this type of environments, i.e., this NDT Platform is designed for dealing with, and learning from, specific scenarios and needs **not addressed by generic analytics tools**. Also, given the NPN context, the NDT platform incorporates and exposes special network configuration capabilities for performance optimization purposes -a service that makes full sense for NPNs as opposed as for PNs, where it is neither applicable nor feasible in practice.
- **It leverages three complementary models for NPN performance** (theoretical model, empirical model, and ML model) corresponding respectively to the model of ranges of performance supported by 5G for a broad set of NPN configurations, the model of the actual performance of the system measured in the lab for multiple configurations and usage patterns, and, finally, the ML model derived, and improved continuously, from the actual usage, behavior and performance of the system in the field.
- It **supports deploying both autonomic close-loop strategies (embedded) and open-ended (co-operative) approaches** with upper layer functions and applications (open exposure of APIs) for **network observability and smart ML-based optimization**, i.e., this NDT platform is versatile, modular, and open.

3.4 JOINT-K1.4 Network Digital Twin

Introduction

The Network Digital Twin (NDT) concept we refer to in this project embodies the realization of a Digital Twin for 5G Network systems, and, more specifically, a Digital Twin for a Non-Public Network (NPN) gNB system. NDT sits, thus, in the convergence of Digital Twin and Mobile Networks concepts, where the Digital Twin approach is applied to different use cases of network: from R&D to network operation (management, deployment, and site engineering). The aim is to achieve an accurate real-time representation of a physical network entity creating a virtual image of a real network element and then use it to simulate different scenarios and identify the most suitable one to the requirements and patterns of traffics of the users/devices enjoying mobile network connectivity services in an NPN.

Within this scope and context, the Network Digital Twin overall *abstract model* represented in Figure 11 is characterized by a set of complementary aspects of the physical system and environment of a specific NPN system, which the corresponding NPN digital instance is aware of, namely, Spectrum Assignment, Spectrum usage technique, Radio Unit type, gNB configurations, air conditions, KPIs and Traffic Models.

When it comes to the key use cases supported by the NDT platform they involve i) Configuration proposal for Physical Twin based in service requirements, that delivers rank ordered list of feasible configurations that match the requirements of user service, ii) Configuration implementation in the Physical Twin based in service requirements, securing the actual implementation and necessary reconfiguration of the system accordingly, and iii) Performance comparison between Physical and Digital Twin performance, providing rich information about how the service is running vs expectations.

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

Then, a foundational design principle for the NDT platform is the integration of three sources of knowledge for making up robust and accurate models of the twinned NPN instances.

As illustrated in Figure 9, the NDT platform is proposed to be fed with i) a domain knowledge model, ii) an empirical model and iii) a data-driven trained model. That combined strategy allows for respectively i) dealing with expectable ranges for performance, ii) predicting performance levels for multiple potential configurations and finally iii) accurately predicting performance for the system in its current configuration, condition, and usage, based on training.

Finally, the proposed NDT platform embodies a set of services that may be either directly used by the CSP or invoked, and composed with other services, by several stakeholders of the NPN it serves. This provides the NPN ecosystem with a powerful tool to gain observability of NPN and devise new services for customization, monetization, optimization, and sustainability of the NPN. We propose to i) base the exposure of NDT services on interoperable standard interfaces whenever possible (in practice, whenever a similar service is already defined by 3GPP), and ii) provide open access APIs to access NDT platform services that lay beyond the scope and plans of mobile networks interoperable standards, for dealing with NPN NDT specific services exposure.

SoA

Let's analyze the underlying State-of-The-Art related to the above introduced Network Digital Twin concept.

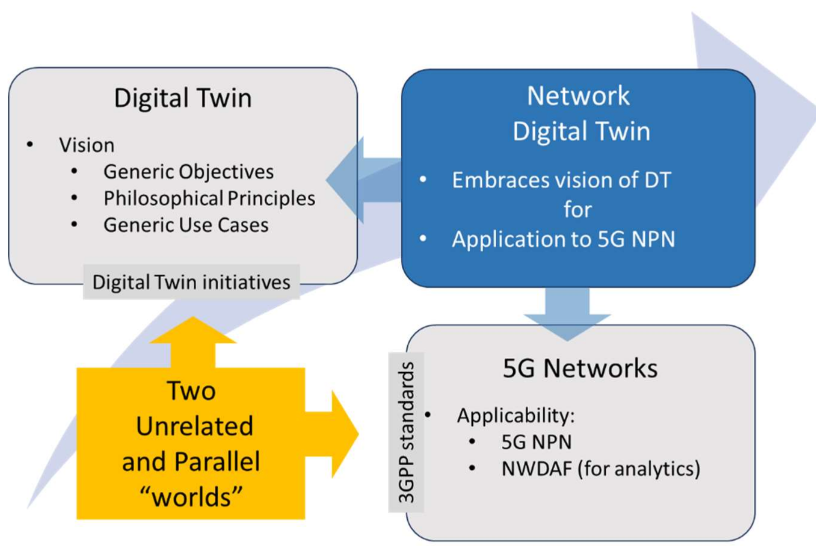

FIGURE 13. SOTA CONTEXT FOR NETWORK DIGITAL TWIN

As represented in Figure 13, the "worlds" of Digital Twin (principles) and Network technologies (standards) are indeed unrelated, evolving in parallel tracks, and governed by different dynamics. So, the Network Digital Twin concept is the first convergence of those two worlds and their trends into an innovative concept that stemming from and embracing the vision of Digital Twin, it is meaningful, feasible and applicable for 5G NPN systems.

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

For a structured analysis, trends in the two complementary worlds (Digital Twin, and Networks) that are instrumental to the overarching concept of Network Digital Twin, shall be analyzed now.

Analysis from the Digital Twin space

The digital twin concept could be defined as a digital representation of a real-world object synchronized at a specified periodicity and fidelity. This original concept where systems were mirrored to monitor unreachable physical spaces was used by NASA in 1970 for the Apollo missions, among others. This potential of virtual and simulated models has been considered as precursors of Digital Twin (DT), however, are not considered proper DTs, as lack of seamless connections and real-time data exchange allowing the periodic "twinning" of the digital to physical [1].

The DT concept evolved and, in 2002, Michael Grieves described an accurate real use case applied to manufacturing [2]. In this white paper the DT concept is defined as containing the following three primary elements: physical object in real space, simulated object in virtual space (with virtual sub-spaces) and the link for data flow between the real and virtual spaces.

Digital Twin is used, today, in industries such as aerospace, automotive, energy and manufacturing helping to improve processes and products. Despite its wide -and fragmented- applicability and momentum, and because of the many diverse ecosystems the DT approach is embraced in, no common industry standards are addressing a specification of this phenomenon. Instead, only general ambition principles are discussed at several associations and institutes such as Digital Twin Consortium [2] and reflected in articles in journals [3].

Analysis from the Mobile Network technologies domain

No 3GPP standards (specs) are defined or planned to be defined soon for a full-fledged all-purpose Network DT platform as such. A possible reason behind could be that an eventual 3gpp specification for a Network Digital Twin platform should cater for creating, maintaining and exposing Digital Twin instances for _any_ type of 3gpp network, so the complexity involved in such type of functionality for complete (physical) 5G networks, with potentially millions of users, multiple traffic models and heterogeneous usage patterns, thousands of apps and many thousands of square kilometers of coverage is simply overwhelming and clearly not addressable for the near future.

That said, within the family of specs of 3gpp the closest one to be supporting some of the basic requirements of Network Digital Twin is that of NWDAF, since the role of this NF in the 5GS is to analyze QoE and provide insights for enabling automation in the Core for reducing CAPEX and OPEX whilst safeguarding QoE target levels. So, in many ways, core processes assumed to be executed by a NWDAF function, including continuous performance measurement collection, and carrying out real-time network data analytics, are also required for a Network Digital Twin implementation. However there's a critical space of requirements that NWDAF specs do not cover vs the intent of NPN Network Digital Twin encompassing i) creating, maintaining and exposing Digital Twin Instances for simulation and planning purposes (Digital Twin principle), ii) the specific focus of application to NPNs (physical system to be twinned), and iii) generating insights, recommendations and enforcing actuations beyond the core and into the RAN, as major influential network segment for NPN performance optimization (domain of analysis and actuation).

Network Digital Twin model for abstracting the NPN gNB system and its environment

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

The model created in the digital twin for abstracting the gNB system of an NPN and its environment encompasses the following aspects: Spectrum Assignment, Spectrum usage technique, Radio Unit type, gNB configurations, air conditions, KPIs and Traffic Models.

Spectrum assignment: 5G technology has standardized a large spectrum for the communications to operate divided into Low-Band (spectrum lower than <1GHz) Mid-Band (spectrum between 1GHz and 6GHz) and High-Band (spectrum higher than 6GHz, normally around GHz). The only band considered for the NDT platform implemented in the 6GEdgeDT project is Mid-Band in n78-B43 (Telefónica Spectrum: 3550MHz).

Spectrum usage technique: Is the strategy used in mobile communication to transmit the data in both forms of duplex. While FDD stands for Frequency-Division Duplex, TDD stands for Time-Division Duplex. 5G technology largely uses TDD strategy and support FDD in some cases. For NDT platform implemented in the 6GEdgeDT projects the model supported is TDD.

Radio Unit types: Based in the different radio types the NPN portable system could support 3 different radio models and are configurable in the Digital Twin:

- Small Cell DOT: Smallest 4x4 cell Ericsson has, that provides indoor coverage for small areas.
- Small Cell RRU4408: Indoor and outdoor 4x4 cell used for small-medium coverage areas.
- Macro Cell AIR6488: First 5G antenna 64x64 that could cover use cases, designed for large areas.

gNB configurations: Based in the configuration of cell and sectors of the NPN portable system supported, the model of the digital twin takes into consideration the following:

- Bandwidth: Width of the band in which the radio communication is performed. The values configurable in the model are: 20, 40, 80 and 100.
- TDD Pattern: In a Time division duplex communications, the pattern defines the number of downlink and uplink slots configured in the communications transmissions, the values configurable in the model are: 4:1 (DDDSU(10:2:2)) and 7:3 (DDDSUDDSUU(10:2:2)), in spite of the fact that the only configuration compatible with the current regulation is 4:1 (DDDSU(10:2:2)).
- Transmission power: Is the power configurable in the NPN portable system so the antenna could propagate the communication signal. The configurable values depend on the specific radio type selected. For the DOTs they would be 500 and 1000 mW, for the Remote Radio Unit (RRU) 4408 is goes from 1000 to 10000 in steps of 1000mW and for the AIR6488 we reduced the possibilities to 2400mW and 10000mW.
- MIMO: Is the method for multiplying the capacity of the radio link using multiple transmissions and receiving antennas to exploit multipath propagation. MIMO stands for multiple-input and multiple-output. Is a configurable that depends of the radio used and to get all capacity also in the end user device used. In the model when small cell radios are used, we can go up to 4x4 value (configuring value of 4), while using AIR6468 we can set 64x64 (configuring value of 0, max). Take in account the use of the maximum antennas also depends on the air conditions.

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

- Carrier aggregation: Is a technique used to increase the data rate per user, whereby multiple frequency blocks are assigned to the same user using different carriers in the same spectrum. For the NDT platform implemented in the 6GEdgeDT project single carrier is the one allowed.

Air conditions – radio link adaptation

The communication in a mobile network depends strongly on the air conditions. An external source of our system, transmitting in the same radio spectrum, could cause an interference that degrades the signal and affect KPIs as user data rate and latency.

The radio link algorithm is adapting, each millisecond, to the situation of the air interface, trying to stretch the channel at maximum to reach the highest radio signal quality. One of the aspects the radio negotiates with the end user device is the NR modulation and coding scheme also called: modulation and coding scheme (MCS). MCS defines the number of useful bits can be transmitted per resource element in a radio communication.

MCS depends on radio signal quality in wireless link:

- Good air conditions: good signal quality results in that higher MCS could be set and more useful bits can be transmitted with in a symbol.
- Bad air conditions: bad signal quality resulting in that lower MCS could be set and less useful data can be transmitted with in a symbol.

Key performance indicators for model

The KPI refer to a set of quantifiable measurements used to quantify the performance of the NDT platform.

User data rate (Network speed): Measures the number of data bits successfully transferred in one direction between specified reference points per unit time for the user service. Unit: Megabits per second (Mbps).

- Downlink Throughput: Measures the data download rate from the network towards the user.
- Uplink Throughput: Measures the data upload rate from the user towards network.
- Latency: Measures the time delay of a data packet to travel between the User Equipment (UE) and the application of the network and is measured in milliseconds (ms).
- Round-trip Time (RTT): Measures the duration between the transmission of a small data packet from the application layer connected to the UE and the successful reception at the application layer connected to the UPF, plus the equivalent time needed to carry the response back.
- One-Way Delay (OWD): Measures the duration between the transmission of a small data packet from the application layer connected to the UE and the successful reception at the application layer connected to the UPF, without considering the response back.
- Jitter: Jitter measures the variation in the delay of data packets as they are transmitted over a network. Jitter is measured in milliseconds (ms) and represents absolute value of the difference between the One-Way Delay of two consecutive received packets belonging to the same stream.

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

- Packet loss rate: Measures the percentage of data packets that are lost during transmission over the network, including packets dropped, packets lost in transmission and packets received in wrong format.
- Reliability: The amount of sent packets successfully delivered to the user within the time constraint required by the targeted service, divided by the total number of sent packets. NOTE: the reliability rate is evaluated only when the network is available.
- Packet order rate: Measures the percentage and degree to which data packets are received in the correct order on a network.
- Energy consumed: Measures the energy consumed by the Physical Twin and is a key aspect for the energy saving objective of the NDT platform. The Unit is Watts per hour (Wh).

NDT Platform Components

An NPN Network Digital Twin is a digital representation of a physical instance of the Non-public part of a PNI-NPN set-up. Several layers and systems make up the NDT platform propose for supporting this concept.

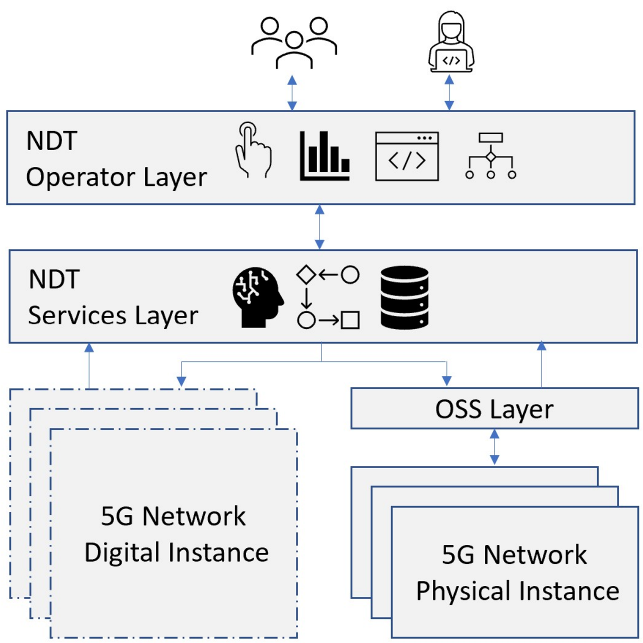

FIGURE 14. NETWORK DIGITAL TWIN PLATFORM

- **NDT Operator Layer**: Is the layer offering the interfaces of the NDT platform to end users for either accessing information or composing customized services, over the platform.
- **NDT Services Layer**: Is the cornerstone of the NDT platform, including the model for NPN performance, and the intelligence for leveraging such model to be delivered through the core Use Cases of the NDT platform. Note that the NDT Services Layer provides (Northbound) the NDT Operator layer with APIs is in communication, leverages (Southbound) the operational support system (OSS)

Layer for interacting with the twined 5G Network Digital Instances, and spawns and manages DT instances for each twined 5G physical instance.

- **5G Network Digital Instance(s):** Are the different 5G Network models created by the NDT platform. These instances are generated with different parametrizations and using both actual performance data collected by the platform and Artificial Intelligence (AI) techniques.
- **OSS Layer**: The OSS is leveraged for the mission of collecting all the relevant data in terms of configuration and actual KPI, from the 5G Network Physical Instances, as well as for applying different configurations on the 5G Network Physical Instance.
- **5G Network Physical Instance**: Consists of the actual part of the NPN that will be replicated by the NDT platform and is the objective part of the network to be monitored and controlled by the end user. Note that there could be more than one instance under the scope of the same NDT platform. For this project, the 5G Network Physical Instance is an NPN Portable System that consists in Radio Access Network (RAN) equipment and User Plane Function (UPF), integrated in a Flight Case system called “carrito”.

NDT Use Cases overview

Having described and explained the components of the NDT platform, it is now necessary to explain how the NDT platform will cover the needs of the user, through the key uses cases outlined here.

First and main concern for a user that will have a service running over a mobile network is to find out how many and which configurations of the NPN may support her needs. So, the first use case taken into consideration is NDT Use Case 1: “Configuration proposal for Physical Twin based in service requirements”, that delivers rank ordered list of feasible configurations that match the requirements of user service.

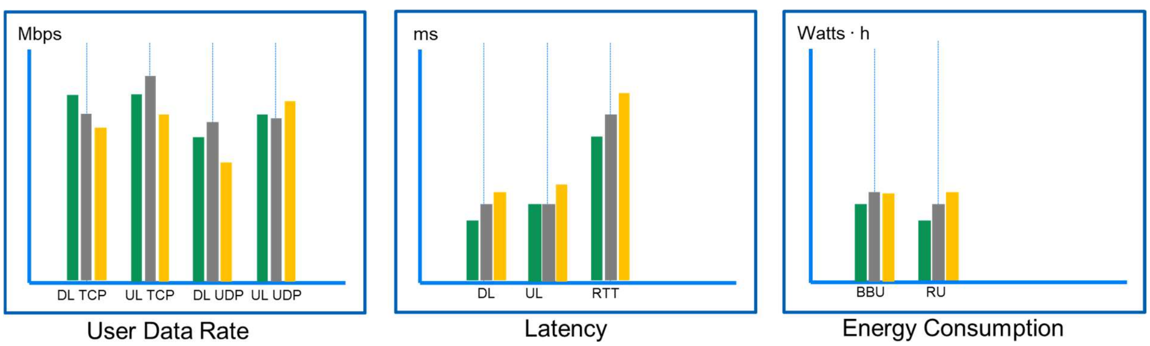

FIGURE 15. CONFIGURATION PROPOSAL FOR PHYSICAL TWIN BASED IN SERVICE REQUIREMENTS

Then, whenever the service demands, according to Use Case 1, the automated implementation is achieved through NDT Use Case 2: “Configuration implementation in the Physical Twin based in service requirements”.

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

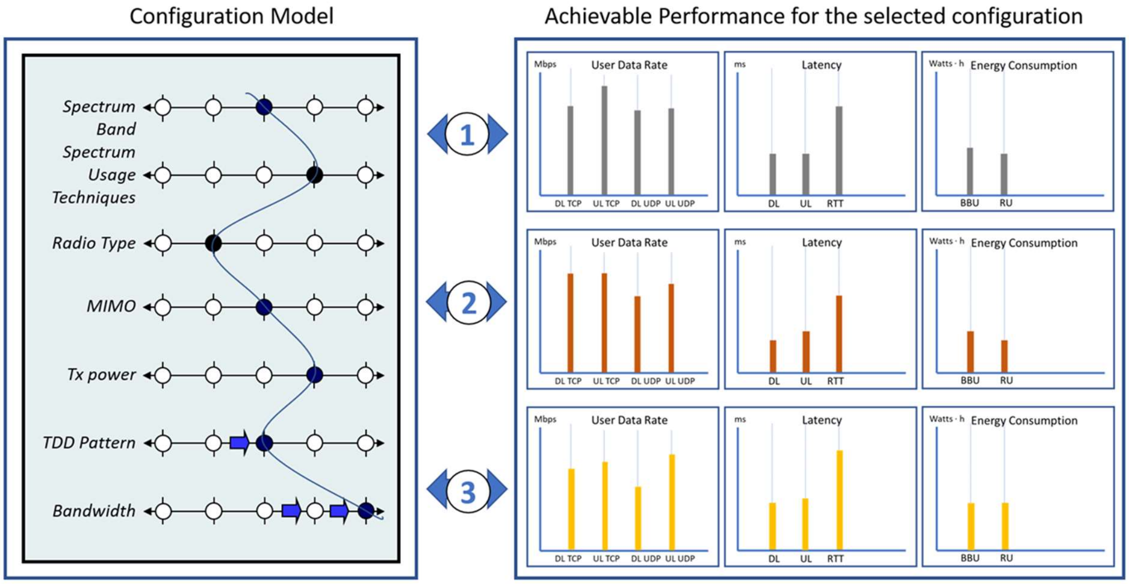

FIGURE 16. CONFIGURATION IMPLEMENTATION IN THE PHYSICAL TWIN BASED IN SERVICE REQUIREMENTS

Finally, once the best/optimal configuration (a priori) is implemented, the continuous comparison of the actual physical performance with the associated Digital Twin instance simulation performance will provide rich information about how the service is running, and that is the scope and goal of the NDT Use Case 3: "Performance comparison between Physical and Digital Twin".

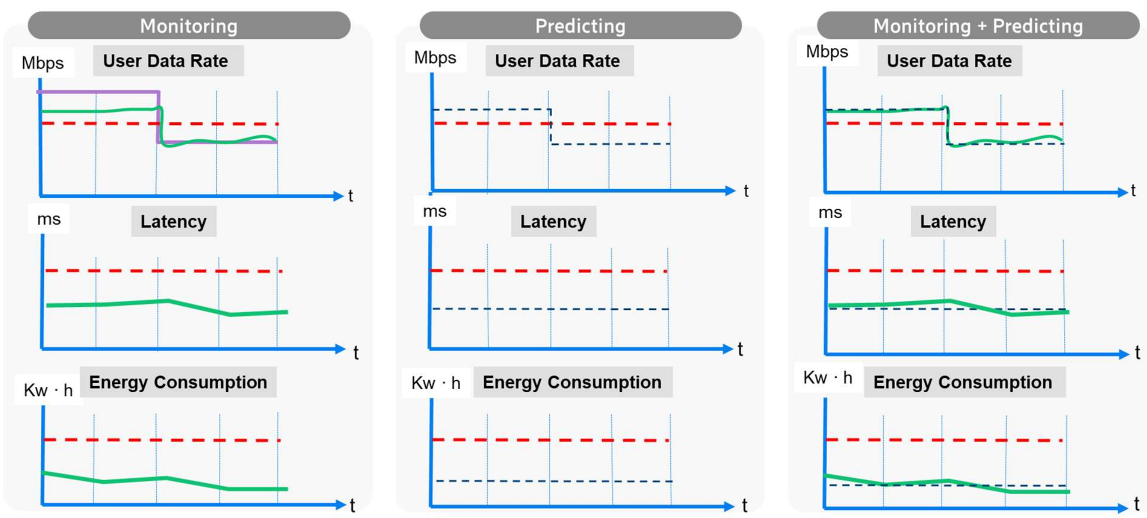

FIGURE 17. PERFORMANCE COMPARISON BETWEEN PHYSICAL AND DIGITAL TWIN

Beyond SoA

The key technological aspects enclosed in the NDT concept being researched that go beyond the State-of-The-Art are:

- It acts the main orchestrator for zero touch configuration activities and holds the following key features:
    - Discovery and registration of NPNs
    - Generating NPN component initial configurations from a low-level design template
    - Applying the configurations on NPN components
    - Performing basic health check and service validation tests
- It implements the full set of digital twin functionalities from performance monitoring and analysis to recommendations and actuation -i.e., this NDT platform provides with a **comprehensive one-stop shop innovative Network Digital Twin service developed ahead of eventual standardization** in the mid-term.
- It specifically **addresses PNI-NPN scenarios**, modelling NPN technology and performance vs the typical usage patterns and environmental conditions applying to this type of environments, i.e., this NDT Platform is designed for dealing with, and learning from, specific scenarios and needs **not addressed by generic analytics tools**. Also, given the NPN context, the NDT platform incorporates and exposes special network configuration capabilities for performance optimization purposes -a service that makes full sense for NPNs as opposed as for PNs, where it is neither applicable nor feasible in practice.
- It **leverages three complementary models for NPN performance** (theoretical model, empirical model and ML model) corresponding respectively to the model of ranges of performance supported by 5G for a broad set of NPN configurations, the model of actual performance of the system measured in the lab for multiple configurations and usage patterns, and, finally, the ML model derived, and improved continuously, from the actual usage, behavior and performance of the system in the field. (See resources).
- It **integrates with 5G network architecture at the 5G Core**, as any other 5GC NF, i.e., it is designed to be deployed in the network itself, thus gaining privileged access to other NFs and resources. This approach is like -and inspired by- that of other standardized 3GPP NFs which must interact with both 5G network and AFs to effectively fulfil its mission.
- It **supports deploying both autonomic close-loop strategies (embedded) and open-ended (co-operative) approaches** with upper layer functions and applications (open exposure of APIs) **for network observability and smart ML-based optimization**, i.e., this NDT platform is versatile, modular, and open.

3.5 JOINT-K2.1 Transport network optimization

Introduction

In recent decades, networks have undergone significant evolution to accommodate a diverse array of services, each with unique Quality of Service (QoS) requirements. Some prioritize high throughput, others emphasize reliability, while some demand low latency or energy efficiency.

To meet these varied demands, Open RAN utilizes split-based base station architecture (RU/DU/CU) to create a flexible network, optimizing the fronthaul, midhaul, and backhaul segments of the radio access network.

SoA

One approach to this optimization, in line with other features in 5G networks, is network slicing [19] [20]. This concept involves partitioning a single physical network infrastructure into multiple virtual networks, or slices, tailored to specific requirements such as bandwidth, latency, reliability, and security. Each virtual network, known as a slice, operates as an independent logical network instance, despite sharing the underlying physical infrastructure.

Network slicing enables operators to efficiently allocate resources to different services and applications, supporting diverse use cases simultaneously. For example, a 5G network can support ultra-reliable low-latency communication (URLLC), enhanced mobile broadband (eMBB), and massive machine-type communication (mMTC) concurrently.

By providing isolated slices with distinct QoS characteristics, network slicing allows operators to meet the diverse requirements of different industries and applications without compromising performance or security. This flexibility is crucial for supporting the wide range of services expected in 5G networks.

To fully harness the capabilities of network slicing technology, it is imperative for telecommunications providers to implement dynamic resource allocation and tailored partitioning across their entire managed network infrastructure. This facilitates the deployment and management of end-to-end (E2E) slices. The integration of E2E requirements into the network slice paradigm necessitates operators to ensure uniform slice behaviour across various network domains, ranging from access networks to external data networks such as the Internet, encompassing all intermediate components like core platforms, cloud computing environments, and transport networks that provide connectivity between them. Ensuring consistency in provisioning and managing E2E slices presents challenges due to the diversity of these network domains, which evolve technologically and incorporate solutions from different vendors. This complexity poses significant operational hurdles for telecom operators [20].

Numerous telecommunications industry forums such as GSMA [21] and NGMN Alliance [22] emphasize the imperative for advancing end-to-end slicing technology. However, existing initiatives, like 3GPP's focus on specific network segments, have overlooked the transport network aspect. Standardized slicing solutions within transport networks are indispensable for seamlessly integrating vendor implementations. The transport network plays a pivotal role in slice construction, linking endpoints to fulfil service level objectives (SLO). Establishing a standardized approach for crafting network slices within the transport network is crucial to accommodate diverse demands and ensure equitable treatment of service requests.

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

Beyond SoA

Presently, the industry lacks a standardized approach for defining a common Transport Slice Controller component capable of facilitating the provision of connectivity services in slice form. This component would be tasked with managing slice requests and associated procedures.

Network Slice Controller

The Network Slice Controller (NSC) is a component defined by the IETF [23] to orchestrate the request, realization, and lifecycle control of network slices. It consists of two main modules: the mapper and the realizer. The mapper processes the customer's request, contextualizing it within the IETF transport network, while the realizer translates this request into a realization of the transport network to fulfil the slicing request, interacting with the associated network controller present in the network. The request received by the NSC originates from a 5G customer requesting an end-to-end slice with specific requirements. This request is handled by the 5G end-to-end orchestrator, which configures the RAN and Core Network elements accordingly and passes the request to the NSC for processing. The NSC then interacts with the relevant network controllers to implement the network slice into the transport network. The whole architecture of the NSC is depicted in Figure 18, highlighting the main modules used in this project. The various modules of the NSC are described in more detail below:

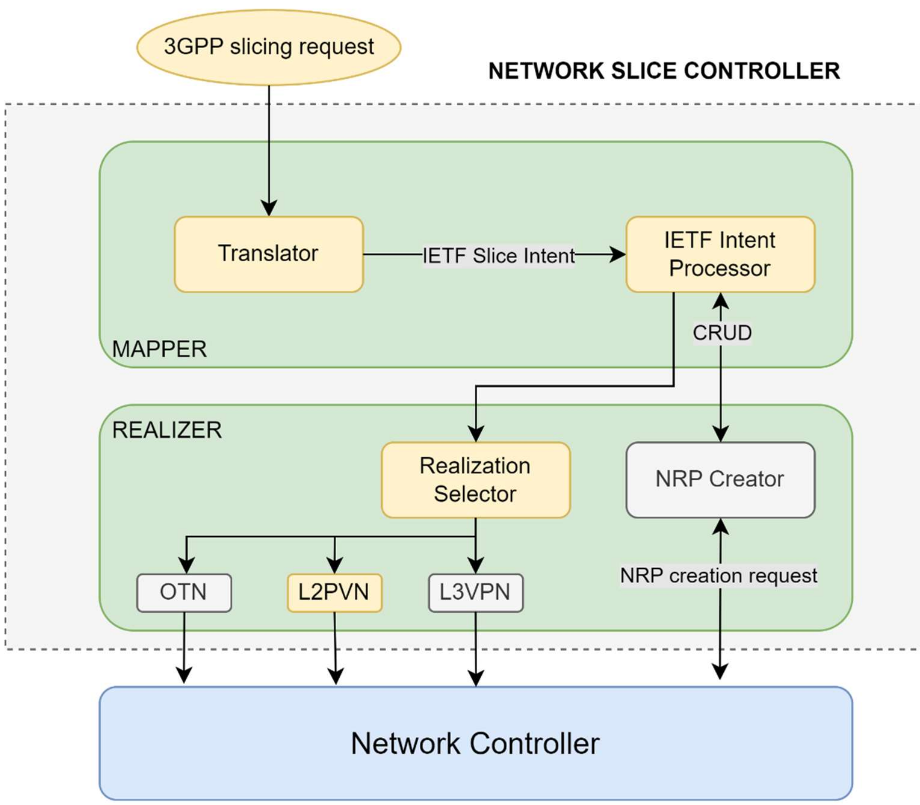

FIGURE 18. NSC ARCHITECTURE

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

Mapper:

The mapper processes client network slice requests and correlates them with existing slices. The workflow is as follows: when a slice request arrives, the mapper translates it by converting the request expressed in 3GPP NRM terms [24] into the IETF NBI data model [25]. This involves identifying the service demarcation points (SDPs) that define the connectivity in the transport network. Once these parameters are identified and mapped into the data model, the next step is to check the feasibility of implementing the slice request. Realizing a slice requires an existing network resource partition (NRP) with the specified slice requirements, which may not be available at the time of the request. This information will be retrieved from an external module, which is beyond the scope of this definition. This module will provide a response regarding the feasibility of realizing the slice. If there are no available NRPs for instantiating the slice, the mapper will request the realizer to create a new NRP. This involves interacting with the network controllers responsible for the transport network handled by the NSC. This process is iterative until the mapper determines that the slice realization is feasible. For the tests conducted in the context of the project, it is assumed that there is only one available NRP corresponding to the entire network, and that it is always accessible to the user.

Realizer:

The realizer module determines the realization of each slice by interacting with specific network controllers, i.e. Teraflow [26]. It receives requests from the mapper and decides on the technologies to be used to instantiate the slice based on the selected NRP associated with the slice. For example, Layer 2 VPN is the technology employed to realize network slices in the tests conducted in the project. To achieve this, the realizer generates a request for the network controller to establish a Layer 2 VPN between two SDPs with the requirements specified in the slice request.

3.6 JOINT-K3.1 Monitoring platform

Introduction

The ongoing trend of softwarization of mobile networks offers several opportunities and challenges to the task of network monitoring. On the one hand, the virtualization of network functions allows to deploy them in commercial off-the-shelf servers, opening up the possibility of flexibly deploying of software monitoring agents that can gather the necessary data for effective monitoring of the network. This avoids costly, tailor-made monitoring tools that need to be designed ad-hoc for the particular hardware running a given function. On the other hand, having network functions running in generic servers and potentially sharing the server resources with each other imposes the need for monitoring, in addition to network parameters related to communication performance, also parameters regarding the computational and memory resources of the servers.

SoA

The disaggregation of the RAN that O-RAN advocates, in particular, opens the door to much more cost-efficient monitoring solutions of RAN parameters than previous architectures. In legacy networks, radio-units (RUs) and the entities controlling their PHY and RRM functions were typically located in the same base-station mast, and connected with proprietary, vendor-specific software and hardware solutions. This made the

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

development of monitoring solutions for the radio parameters infeasible, as ad-hoc solutions needed to be specially designed for each vendor's hardware.

In the disaggregated O-RAN paradigm, on the contrary, the separation of radio units (RUs) from central units (CU)s and distributed units (DUs) and their connection via open and standardized interfaces allows for monitoring solutions that can be software-only, as well as independent of the vendor of RUs, CUs or DUs [27]. Due to this new opportunity, in 6G-BLUR we aim to focus on contributing to a software-based monitoring entity for RAN parameters.

Beyond SoA

An important aspect of such a monitoring platform is to develop the mechanisms to collect the necessary RAN data from the O-RAN interfaces. Along those lines, a solution is to perform KPM monitoring for the E2 interface based on an xApp running in the near-RT RIC. An example of this type of E2-monitoring xApp has been provided by the O-RAN Software Community [28]. Such an xApp, combined with standard monitoring tools for data centers such as InfluxData TICK Stack [29], would allow for a complete view of all relevant network and computational server parameters that need to be tracked in the RAN. Keysight is currently developing such an xApp to be deployed in 6G-SANDBOX testbed on FlexRIC.

The goal is to integrate the monitoring xApp with the rest of tools used in TICK Stack, such that the monitoring data on the E2 interface can be combined with the rest of data collected by the observability framework developed under TICK Stack, as illustrated in the schematic view provided in Figure 19. The proper functioning of this solution can be extensively tested via the use of Keysight's RICtest solution: an emulator of E2 interfaces that allows to generate data to test a RIC on different scenario assumptions. The role of such software in order to test and validate the proposed monitoring platform approach can also be observed in Figure 19.

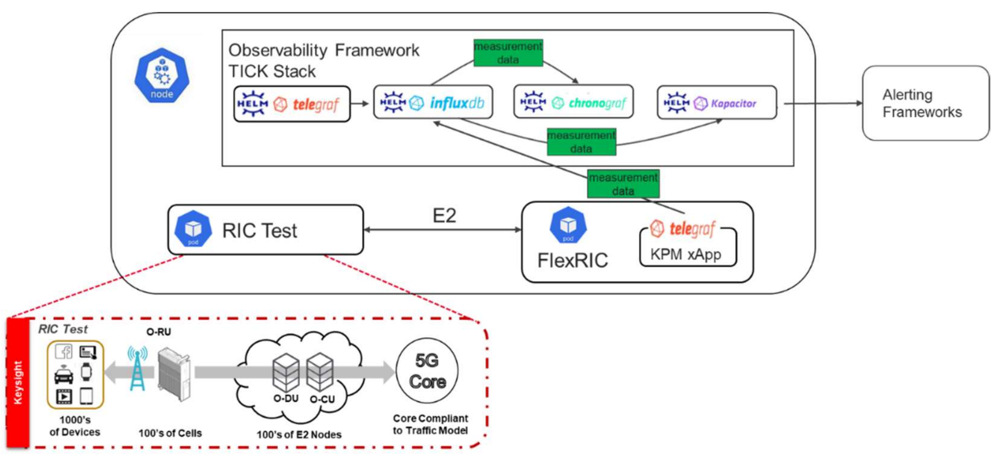

FIGURE 19. ARCHITECTURE OF A TESTING FRAMEWORK FOR AN E2 MONITORING XAPP

Another avenue that can be explored is intent based on adaptive monitoring. We propose an adaptive monitoring system for O-RAN that integrates intelligent query handling mechanisms using Large Language

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

Models (LLMs). The system aims to streamline data visualization and adapt to dynamic network conditions and user requirements. Our approach dynamically converts user intent into low-level O-RAN component APIs and other infrastructure tools, validating the system through deployments with both software-based and RF front ends. The results demonstrate the system's capability to inject advanced intelligence into various network tasks, optimizing performance and efficiency.

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. In the context of O-RAN, the majority of interactions between different components are managed through configurations in YAML files, software specifications, and infrastructure running on cloud-native tools like Kubernetes. By integrating LLMs, O-RAN and future 6G networks can benefit from advanced intelligence, optimizing configurations, automating processes , and improving overall network performance. This infusion of AI at the core of network operations promises to enhance efficiency, adaptability, and the user experience in next-generation mobile 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 [30]. 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.

In this work, we first create a software-based and RF-based deployment environment for the O-RAN network. Next, we develop an LLM-based agent workflow that incorporates tools such as RIC, Grafana, and a Redis database. Finally, we integrate this system and demonstrate how advanced intelligence can be injected into O-RAN networks, showcasing the potential for enhanced network optimization and performance through AI-driven solutions.

Our experiment setup is shown in Figure 20. We utilize the Open5GS core and SRS gNB, which can be deployed on bare metal, Docker, or Kubernetes (K8s). Two sub-scenarios are considered:

- Scenario A involves a ZMQ-based gNB and UE, with all components deployed on a single VM.
- Scenario B uses an RF front end, where the gNB is connected to a USRP RF front end, and we utilize commercial UEs.

For both scenarios, we use the O-RAN Software Community RIC (SC RIC) due to its high compatibility with the SRS gNB. This setup allows us to effectively manage and orchestrate network interactions, facilitating advanced monitoring and intelligence integration.

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

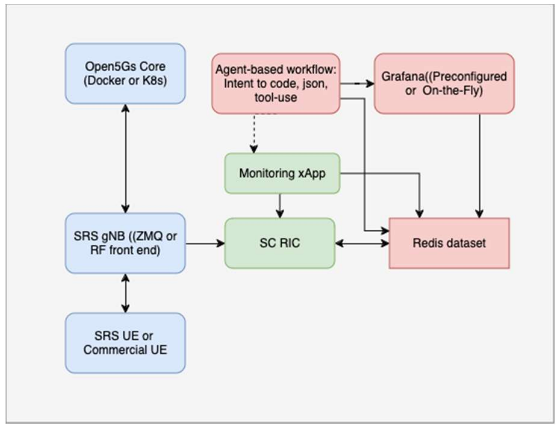

FIGURE 20. PROPOSED ARCHITECTURE OF THE ADAPTIVE MONITORING SYSTEM FOR O-RAN.

3.7 JOINT-K3.2 AI/ML Platform (AIMLP)

Introduction

Aiming at ensuring swift and effective mechanisms for ML optimization of mobile networks performance and resources the approach of embedding analytics engine in Digital Twin is emerging. The proposed approach integrates **two complementary strategies**: **Digitally Twinning the physical system** -and therefore continuously monitoring it for performance and status-, on the one hand, and leveraging that information through **autonomous embedded analytics for gaining insight on how to effectively improve/optimize the performance of *that* system** -taking into account its actual patterns of traffic-. The Digital Twin platform might so enable, or even autonomously enforce, the identified optimization actions derived from such insight generation.

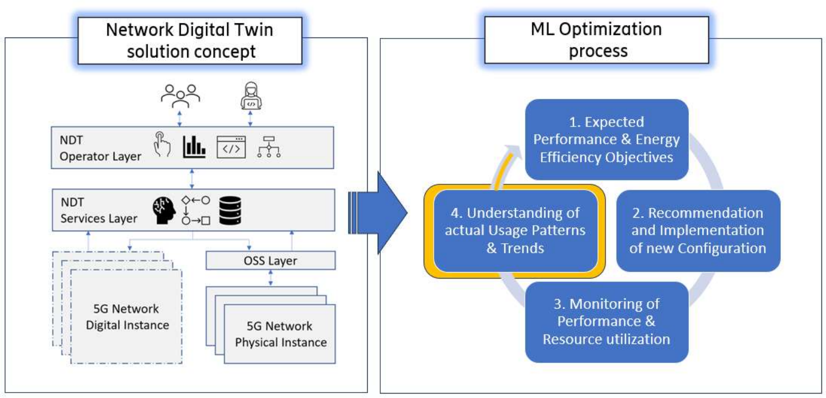

FIGURE 21. AI/ML POWERED NDT PLATFORM

Digital Twin solutions are mainly aimed at timely detecting unexpected divergences between the actual and predicted performance of the twined physical system, so that alarms can be raised for trying to prevent or mitigate critical failures of performance that might follow shortly after. The Network Digital solution concept

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

that we are proposing (represented in Figure 21) extends beyond that objective and into providing support for advanced services of adaptation of the system itself to the requirements on its performance.

The sequence of steps 1-2-3 depicted in Figure 21-be articulated in either manual, semi-automated or fully automated fashions- enables that the desired performance and efficiency *objectives for the expected usage of the physical system* (step 1) are taken into consideration for optimizing the configuration of the system (step 2), which will be concurrently monitored and simulated (accordingly), as per the typical Digital Twin process (step 3). Now, beyond the native goal of step 3, i.e., comparing the actual and predicted behavior and performance it is also possible to incorporate ML techniques for **extracting the *actual patterns of usage of the system* (step 4), and to use them as key inputs for reassessing the original recommendation for the optimal configuration of the physical system** (arc from Step 4 to Step 1). In other words, it could be possible to close (and re-start) the optimization loop that was started from considering the initial -a priori- expectations on performance of the network for the expected usage patterns with a new cycle of recommendation-optimization-monitoring triggered by the extraction of the *actual* conditions and usage patterns on *that* physical system.

The AI/ML Powered NDT Platform takes the complexity out of optimizing 5G Non-Public Networks (NPNs). No human intervention shall be required in this E2E autonomous process other than that of initially setting a range of expectations on network performance levels and resource usage efficiency objectives.

SoA

AI/ML Powered NDT Platform is a new concept that: i) reuses ML advanced methods and tools, ii) embraces Digital Twin vision and iii) is applicable to 3GPP PNI-NPN standard. Each of these levers with baseline technologies defining the underlying State-of-The-Art for this new concept is represented in Figure 22 and further developed in this section.

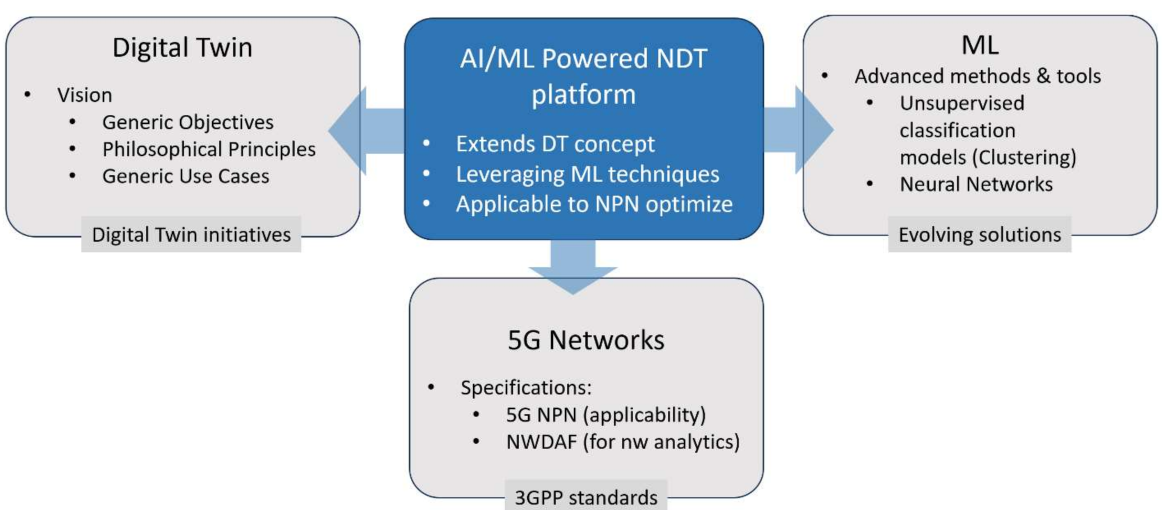

FIGURE 22. STATE-OF-THE-ART MAP FOR AI/ML POWERED NDT PLATFORM

Analysis from ML perspective

By continuously monitoring network performance, we can extract valuable features for machine learning models. These models can then be used to optimize network performance by recommending the optimum configurations for the NPN system. Additionally, the models can predict network performance for real-world NPN configurations, even for scenarios not considered during the training phase.

For that purpose, following techniques are developed and deeply analyzed the output to improve the needs of the real Network once introducing our solution in real scenarios.

There are two phases in that process:

1. Supervised Learning
2. Unsupervised Learning

Firstly, based on the input data collected in the lab, machine learning model is being trained with data collected knowing the output, as we are building the input database based on measurements for each configuration with multiples combinations:

Analysis from 3GPP perspective

No 3GPP standards (specs) are defined or planned to be defined soon.

That said 3GPP has specified NWDAF and some of its principles being applicable or at least a source of inspiration for either the implementation or the exposure of the type of Analytics Services proposed by the solution for ML Optimization with Embedded Analytics in DT.

3GPP NWDAF analyzes QoE and provide insights for enabling automation in the Core for reducing CAPEX and OPEX whilst safeguarding QoE target levels. So, in many ways, core processes assumed to be executed by a NWDAF function, including continuous performance measurement collection, and carrying out real-time network data analytics, are also required for a Network Digital Twin implementation. However there's a critical space of requirements that NWDAF specs do not cover vs the intent of NPN Network Digital Twin encompassing i) creating, maintaining and exposing Digital Twin Instances for simulation and planning purposes (Digital Twin principle), ii) the specific focus of application to NPNs (physical system to be twinned), and iii) generating insights, recommendations and enforcing actuations beyond the core and into the RAN, as major influential network segment for NPN performance optimization (domain of analysis and actuation).

NDT Models

A foundational design principle for the proposed NDT platform is the integration of three sources of knowledge for making up robust and accurate models of the twinned NPN instances.

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

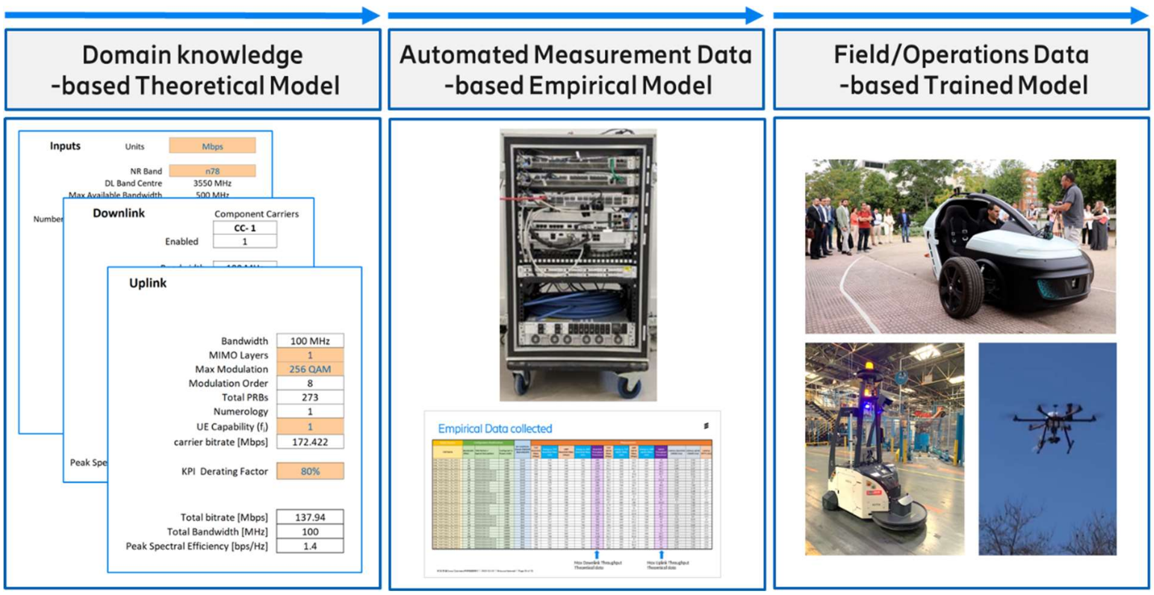

FIGURE 23. FROM DOMAIN KNOWLEDGE MODELLING TO MODEL TRAINING FPOR NDT

As illustrated above, the NDT platform is proposed to be fed with i) a domain knowledge model, ii) an empirical model and iii) a data-driven trained model. That combined strategy allows for respectively i) dealing with expectable ranges for performance, ii) predicting performance levels for multiple potential configurations and finally iii) accurately predicting performance for the system in its current configuration, condition, and usage, based on training.

Domain knowledge -based Theoretical Model	Automated Measurement Data -based Empirical Model	Field/Operations Data -based Trained Model
• Set of **mathematical rules** that: ◦ Govern ranges, limits and **basic relationships between certain configuration parameters and performance** ◦ Can be derived from the **implicit mathematics and physics** for the domain	• Set of **tabulated measurements** that: ◦ Reflect with great accuracy the **real performace of the system**, for selected configurations and under the stimulus of concrete test traffic models ◦ **Are obtained through controlled, systematic and automated measurements** over a rather large set of configurations and traffic models	• Set of **algorithms trained** with theoretical data, empirical data and field/operations performance data: ◦ Input data reduction of dimensionality and Classification ML Model to **recommend** different configurations for achieving input requirements of NW Performance. ◦ **Prediction** of KPIs are obtained through Regression Neural Network providing multi-output.
• Allows to: ◦ **Estimate theoretical performance limits** of certain KPIs for a broad range of configurations	• Allows to: ◦ **Provide recommendations for NW configurations** associated to tested configs and traffic models ◦ **Very accurately predict expected performance levels** associated to tested configs and traffic models ◦ **Roughly approximate performance levels** for untested configs or traffic models	• Allows to: ◦ **Accurately provide recommendations for best configuration to apply on the NW.** ◦ **Accurately predict expected performance levels** associated to **both** tested and non tested configs and traffic models

FIGURE 24. THEORETICAL, EMPIRICAL AND TRAINED MODELS CONTINUUM

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

Closed-loop continuous analysis and adaptation mechanism

The aim of closed-loop continuous analysis and adaptation mechanism is to enable an adaptative configuration from the proposed solution. It doesn't necessary mean that the new proposed configuration is implemented automatically. It will need evaluation from "end user engineer" to implement proposed configuration in the Physical twin unless some custom rules are defined to limit the impact on the Network if a complete autonomous solution is wanted and an implementation process defined.

The feedback of this closed-loop solution is based on:

- Provide optimum configuration based on performance requirements.
- Predict performance based on configuration.
- Identify room for improvement using configuration proposal based on performance comparing expected performance vs measured performance from real network.

The mechanisms implemented for closed loop will be the ones already explained in analysis from Machine Learning perspective section of this document.

Beyond SoA

In today's mobile networks, it is not possible to see an example of Network Digital Twin platform which has built-in AI/ML capabilities to perform different optimizations on the network such as network performance. This project aims to address this gap by developing an innovative, proof-of-concept level Network Digital Twin (NDT) platform which has the following fundamental features:

- Optimize NPN network performance KPIs using AI/ML techniques.
- Offer versatility and modularity by supporting both closed-loop (embedded) and open-loop (cooperative) optimization strategies.
- Open and accessible through APIs for network observability and further ML-based optimization.

By combining these features, the NDT platform has the potential to be a significant advancement in efficient NPN management.

It should be also noted that while developing AI/ML model for the NDT platform, our focus would be finding the most convenient AI/ML model among the existing, well-established models from the research community that fulfills the requirements of performing an accurate performance optimization on the NPN system, rather than developing a novel AI/ML model or algorithm that produces better optimization results than the existing methods.

4 WP3-related Key Concepts

4.1 JOINT-K2.2: Entities' placement decisions

Introduction

Today's networks are filled with a massive and constantly growing variety of network functions that, along with proprietary devices, lead to network ossification and hinder network management and service provisioning. Network function virtualization is a promising paradigm to change this situation by decoupling network functions from hardware and performing them in software form, known as virtualized networks. Such decoupling introduces many benefits including reducing Capital Expenditure (CAPEX) and Operational Expenditure (OPEX) and greater flexibility in service provisioning. On the other hand, the location of the virtualized entities poses new challenges since computation and communication capabilities are to be jointly considered in order to ensure the required performance. In general, the underlying optimization problems are referred to as Virtual Network Embedding (VNE). By enabling the definition of virtual networks over a common physical infrastructure, VNE problems aim to optimize resource utilization and adapt to changing network needs. Additionally, it can help intelligently manage network resources by allocating processing, storage, and bandwidth capacities, as well as traffic demands according to application and service requirements.

In the context of the RAN segment, the recent trend towards a virtualized and dissagregated architecture conveys the necessity to decide the optimum placement of the different entities while considering the effect of such decision over the rest of the network.

SoA

The main challenge of VNE problems is its computational complexity, as it is a well known NP-hard problem. It means that finding the optimal allocation of resources from a virtual network in a physical substrate, ensuring that the allocation of resources meet the requirements without exceeding the capacity of the physical infrastructure, is a computationally expensive challenge and does not have a polynomial solution in a reasonable time for large instances. Due to its NP-hard nature, finding an optimal solution may require a significant amount of computing time, limiting the scalability of traditional algorithms. Additionally, the number of possible combinations of resource allocations in virtual and physical networks can be exponential, making it more difficult to find the optimal solution in a reasonable time.

Another major challenge in VNE is the consideration of multiple criteria and constraints, such as Quality of Service (QoS), processing capacity, bandwidth, latency, and geographic location of resources. Optimizing these criteria simultaneously and efficiently adds complexity to the problem and requires sophisticated algorithms to find viable solutions.

Many type of solutions have been proposed to tackle VNE problems. Heuristic techniques, such as Greedy Algorithm and Harmony Search, aim to find acceptable solutions within a reasonable time, although they do not guarantee the optimal solution. On the other hand, metaheuristic techniques, like Genetic Algorithm, Ant Colony Optimization, and Tabu Search, offer a more intelligent and adaptive search and are more likely to find high-quality solutions but are computationally more expensive. The choice between them depends

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

on the problem's complexity, computational efficiency, and the balance between exploration and exploitation of solutions. For complex problems where escaping local optima is required, metaheuristics may be more suitable, while heuristics may be preferable if computational efficiency is prioritized. Besides, different types of AI/ML approaches have been also proposed.

Beyond SoA

Considering the natural graph-based topology that emerges in networks (both virtual and physical), in recent years the research community has proposed the adoption of Graph Neural Networks (GNN) as a suitable tool for problems in communication systems.

As mentioned in [31], GNN can be applied to different communication-related areas ranging from wireless networks to SDN. Particularly, a set of works have proposed the adoption of GNN-based AI/ML architectures to solve VNE problems. In this sense, both Ullah et al. [32] propose a hierarchical approach for VNE problems in 5G, which combines both GNN and deep reinforcement learning (DRL). Similar approaches can be found in [33], [34] or [35] presenting differences in the AI/ML solutions architecture. Similarly, in [36] Habibi et al. proposes GraphViNE, a high performance GNN general solution for VNE problems which allows parallelization by design.

Nevertheless, the mentioned works do not apply to specific scenarios, so that both the virtual and physical networks are modeled in a generic manner.

Other works have focused on more specific scenarios. In this regard, in [37] a GNN solution is proposed for embedding of fog computing elements in vehicular network scenarios. However, the adoption of GNN to the RAN segment in general, and the O-RAN architecture in particular, has not been addressed yet. We consider the the adoption of GNN-based solutions may bring advantages compared to other approaches for different reasons. First, the placement of O-DU and O-CU would imply extreme delay requirements in the fronthaul network while demanding very high communication capacity. On top of that, it is necessary to properly capture the complex interaction between the fronthaul traffic and other services that may share the same communication infrastructure. In this sense GNN-based solutions permit understanding the complex relationship between topology, routing, and input traffic to estimate the network performance [38]. Besides, it is expected that multiple, and potentially conflicting, heterogeneous criteria are defined to decide the placement of the different virtualized entities of the RAN. Such heterogeneity would make it difficult to define a single utility function. In order to overcome this challenge, a DRL-based architecture combined with GNN would allow is to consider a wide variety of parameters.

4.2 JOINT-K2.3: QoS policies and scheduling

Introduction

As a result of the convergence of network technologies and the requirements imposed by Open Radio Access Networks (O-RAN) architectures, Quality of Service (QoS) policies and scheduling mechanisms play a key role. Since Open Fronthaul Networks are designed within the framework of the Centralized Radio Access Networks (C-RAN) architecture, in particular split 7.2x, which separates legacy radio base stations into two different nodes: Open-Radio Unit (O-RU) and Open-Distributed Unit (O-DU). This approach introduces strict

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

requirements for the transport network in order to connect these nodes. Meanwhile, transport network is shared by traffic flows belonging to the backhaul as well as open-fronthaul traffic.

There is a lack of works looking at QoS-aware scheduling policies at the fronthaul network using realistic traffic models. However, several works in the context of 5G deployment pay special attention to latency challenges and network performance optimization.

SoA

The authors of [39] discuss the challenges of deploying Reinforcement Learning (RL) policies in real-time scenarios, especially in fronthaul networks. The authors propose using policy distillation to extract simpler Decision Tree models. In the same line, the authors of [40] discuss a Tra Traffic Pattern Adaptive Mechanism designed to bound packet delay and its variability in 5G Fronthaul networks. The proposed mechanism, Time-Window with Timeout (TWT), aims to balance packet delay and jitter while ensuring efficient traffic aggregation and management.

Zhang et al. [41] delve into the optimization of routing and packet scheduling in the context of the 5G O-RAN Fronthaul architecture. Particularly, they focus on enhancing communication efficiency between O-DU pools and O-RUs to meet the diverse service requirements of 5G networks. The proposed solution, based on Dynamic Programming, is able to select optimal routes and packet sizes for transmitting traffic across multiple O-DU pools.

Pérez et al. investigate in [42] the performance of CPRI protocol over ring-star topologies used by mobile operators. The study comes up with a number of key findings: benefits of bidirectional ring topologies in reducing average queuing delays, queuing theory models to analyze delay metrics, and packetization strategies in minimizing worst-case aggregated queuing delays. Besides, the work in [43] proposes alternatives for latency planning for both fronthaul and 5G slicing.

These works collectively underscore the significance of optimizing MAC and PHY layers, as well as the underlying network topology configurations, to minimize latency in fronthaul networks. As can be observed, there exists a gap in the evaluation of the configuration of intermediate nodes and HQoS in the transport segment, which integrates both open-fronthaul and backhaul traffic. There are only a few works that touch this aspect.

Budhdev et al. design and implement in [44] a Fronthaul Slicing Architecture (FSA), which uses a wireless scheduler to identify the slice/user for each fronthaul packet. An evaluation over a real testbed evinces FSA's capability to handle up to 80 Gbps of fronthaul traffic with low-latency routing, while yielding a significant reduction in flow completion times.

Balogh et al. compare in [45] the performance of Fair Queuing (FQ) Queue Scheduling Disciplines (QSD) with Weight Round Robin (WRR in a bottleneck topology. The study suggests that FQ algorithms show worse computational performance compared to Round Robin (RR) algorithms. However, the paper hints that an accurate bandwidth allocation in WRR could lead to better performances. 6GBLUR diverges from existing research by focusing on the assessment of performance of the HQoS mechanisms proposed by different vendors, specifically for the fronthaul network. Unlike previous efforts, which mostly concentrated on enhancing fronthaul at the MAC/PHY layers, we will assess the behaviour of various QoS approaches in practical scenarios, where technologies converge. We will also anylyse the impact of prioritizing certain

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

technologies over backhaul services, thus pointing out the circumstances where HQoS would be required to ensure optimal performance.

Beyond SoA

In this context, one of our key contributions is to meet the requirements defined by the ORAN Alliance and to study the impact on the background traffic. To this end, by means of the NS-3 event driver simulator, we pretend to create some utilities in order to accurately mimic the open fronthaul traffic and finally apply different strategies under different circumstances to ensure the requirements of the ORAN Alliance without worsening other traffic slices.

Characterize traffic patterns and load from open fronthaul traffic:

In order to study the performance of the various strategies, it is first necessary to define the input open fronthaul traffic, encompassing both the packet size as well as the rate distribution, according to the bandwidth and numerology per radio port in a Time Division Duplex (TDD) mode. In addition, a double approach is carried out by using the srsRAN framework.

QoS Policies and Strategies:

In the ns-3 event driver simulator, realistic QoS policies are faithfully developed according to the schemes proposed by vendors following the O-RAN specifications, and those with a modified Hierarchical Quality of Service (HQoS) configuration [46] [47] [48], such as a combination of priority queues with Round Robin (RR) fashion, and their effect on backhaul traffic is compared. We have also considered the traffic marking at the edge nodes to mimic a realistic network, as mentioned in Section 4.1.

Define network topologies:

A first set of experiments is carried out over a toy scenario in order to see the effect of the scheduling, to shed light on when they yield benefits. Secondly, we will extend this initial characterization to scenario close the reality, where each DU has a pool of sites, moving the DU over the IP Transport Hierarchy Levels (HL) while increasing the sites, and so traffic load. Figure 25 pictures a spine-leaf topology, where O-DU are settled at spine nodes in HL4. Furthermore, this analysis will also consider events where a link is broken. In this case we will analyse the impact of the scheduling strategies and whether they are sufficient to meet the corresponding requirements.

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

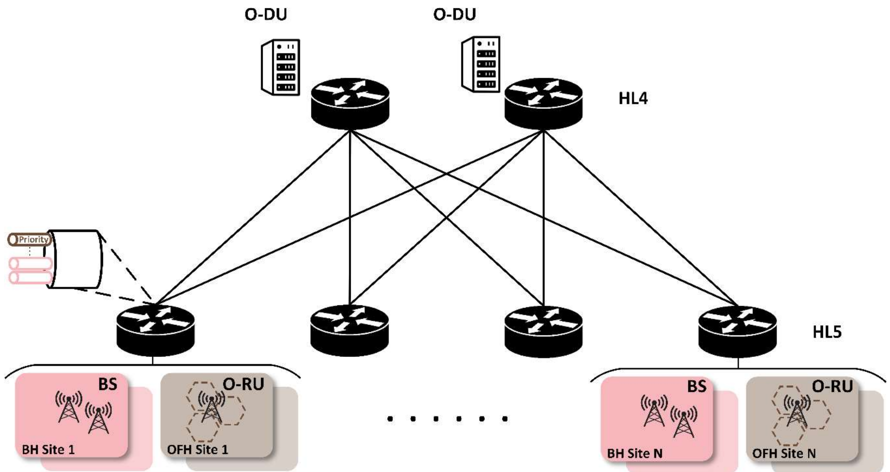

FIGURE 25. EVALUATION PLATFORM SETUP IN NS-3, WITH A DIFFSERV ARCHITECTURE AND FRONTHAUL AND BACKHAUL SERVICE CONVERGENCE.

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

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] O. Arouk and N. Nikaein, "5G Cloud-Native: Network Management & Automation," in *NOMS 2020 - 2020 IEEE/IFIP*, 2020.

[3] S. Barrachina-Muñoz, M. Payaró and J. Mangues-Bafalluy, "Cloud-native 5G experimental platform with over-the-air transmissions and end-to-end monitoring," in *13th International Symposium on Communication Systems, Networks and Digital Signal Processing*, 2022.

[4] N. Apostolakis, M. Gramaglia and P. Serrano, "Design and Validation of an Open Source Cloud Native Mobile Network," *IEEE Communications Magazine,* vol. 60, no. 11, pp. 66-72, 2022.

[5] J. Wu, Y. Gao, L. Wang, J. Zhang and D. O. Wu, "How to Allocate Resources in Cloud-Native Networks Towards 6G," *IEEE Network,* vol. 38, no. 2, pp. 240-246, 2024.

[6] S. Barrachina-Muñoz, R. Nikbakht, J. Baranda, M. Payaró, J. Mangues-Bafalluy, P. Kokkinos, P. Soumplis, A. Kretsis and E. Varvarigos, "Deploying cloud-native experimental platforms for zero-touch management 5G and beyond networks," *IET Networks,* vol. 12, no. 6, pp. 305-315, 2023.

[7] J. Baranda, S. Barrachina-Muñoz, R. Nikbakht, M. Payaró and J. Mangues-Bafalluy, "Disaggregating a 5G Non-Public Network via On-Demand Cloud-Native UPF Deployments," in *19th International Conference on Network and Service Management*, 2023.

[8] S. Barrachina-Muñoz, R. Nikbahkt, A. Bel, J. Baranda, M. Payaró and J. Mangues-Bafalluy, "On-demand 5G/6G edge verticals via third-party UPF selection and cloud-native relocation," in *29th IEEE Symposium on Computers and Communications*, 2024.

[9] OPEN5GS, "Open Source implementation for 5G Core and EPC," [Online]. Available: https://open5gs.org/.

[10] Open Air interface Software Alliance, "OpenAirInterface | 5G software alliance for democratising wireless innovation," [Online]. Available: https://openairinterface.org/.

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

[12] A. Güngör, " open source 5G UE and RAN (gNodeB) simulator," [Online]. Available: https://github.com/aligungr/UERANSIM.

[13] EURECOM, "Flexible RAN Intelligent Controller (FlexRIC) and E2 Agent," [Online]. Available: https://gitlab.eurecom.fr/mosaic5g/flexric.

[14] O-RAN Software Community, "O-RAN SC I Release," [Online]. Available: https://docs.o-ran-sc.org/en/latest/.

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

[15] ORAN SC, "ORAN SC RIC in Docker," [Online]. Available: https://www.google.com/url?sa=t&source=web&rct=j&opi=89978449&url=https://github.com/srsran/oran-sc-ric&ved=2ahUKEwjOnfXR1-SGAxVZSaQEHWStBRMQFnoECA8QAw&usg=AOvVaw2inIOjMw7KxMJIpbz4paJk.

[16] SRS, "srsRAN Project Multi-Container Solution," [Online]. Available: https://github.com/srsran/srsRAN_Project/tree/main/docker.

[17] O-RAN Alliance, "O-RAN Cloud Architecture and Deployment Scenarios for O-RAN Virtualized RAN 6.0," 2024. [Online]. Available: https://specifications.o-ran.org/specifications.

[18] 3GPP, "service requirements for 5G system (stage 1)," [Online]. Available: https://www.3gpp.org/ftp/Email_Discussions/SA5/SA5-level%20discussions/S5-23xxxx%20DP%20Service%20Management_OAM%20rapporteur%20call%202023-03-30.doc.

[19] 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.

[20] X. Foukas, G. Patounas, A. Elmokashfi and M. K. Marina, "Network Slicing in 5G: Survey and Challenges," *IEEE Communications Magazine,* vol. 55, no. 5, pp. 94-100, 2017.

[21] GSMA, "Network Slicing Use Case," 2018. [Online]. Available: https://www.gsma.com/solutions-and-impact/technologies/networks/wp-content/uploads/2018/07/Network-Slicing-Use-Case-Requirements-fixed.pdf.

[22] NGMN Alliance, "5G end-to-end architecture framework," 2017. [Online]. Available: https://www.ngmn.org/publications/5g-end-to-end-architecture-framework-v3-0-8.html.

[23] A. Farrel, J. Drake, R. Rokui, S. Homma, K. Makhijani, L. M. Contreras and J. Tantsura, "A Framework for Network Slices in Networks Built from IETF Technologies," 2024. [Online]. Available: https://www.rfc-editor.org/info/rfc9543.

[24] 3GPP, "5G Network Resource Model (NRM)," 2024. [Online]. Available: https://www.etsi.org/deliver/etsi_ts/128500_128599/128541/16.06.00_60/ts_128541v160600p.pdf.

[25] B. Wu, D. Dhody, R. Rokui, T. Saad and J. Mullooly, "A YANG Data Model for the RFC 9543 Network Slice Service," 2024. [Online]. Available: https://datatracker.ietf.org/doc/draft-ietf-teas-ietf-network-slice-nbi-yang/.

[26] TeraFlow, "TeraFlow SDN," [Online]. Available: https://teraflow-h2020.eu/teraflowsdn.

[27] 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.

[28] ORAN SC, "O-RAN SC KPI Monitoring Documentation," [Online]. Available: https://docs.o-ran-sc.org/projects/o-ran-sc-ric-app-kpimon/en/latest/index.html.

[29] G. Aasen, "Introduction to InfluxData's InfluxDB and TICK Stack," [Online]. Available: https://www.influxdata.com/blog/introduction-to-influxdatas-influxdb-and-tick-stack/.

[30] 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.

[31] W. Jiang, "Graph-based deep learning for communication networks: A survey," *Computer communications,* vol. 185, pp. 40-54, 2022.

[32] I. Ullah, H.-K. Lim, Y.-J. Seok, M. S. Khan and Y.-H. Han, "Efficient Virtual Network Embedding with Hierarchical and Cooperative Multi-Agent Reinforcement Learning," in *International Conference on Communication, Computing and Digital Systems (C-CODE)*, 2023.

[33] Z. Yan, J. Ge, Y. Wu, L. Li and T. Li, "Automatic virtual network embedding: A deep reinforcement learning approach with graph convolutional networks," *IEEE Journal on Selected Areas in Communications,* vol. 38, no. 6, pp. 1040-1057, 2020.

[34] P. Zhang, C. Wang, N. Kumar, W. Zhang and L. Liu, "Dynamic Virtual Network Embedding Algorithm Based on Graph Convolution Neural Network and Reinforcement Learning," *IEEE Internet of Things Journal,* vol. 9, no. 12, pp. 9389-9398, 2022.

[35] A. Rkhami, T. A. Q. Pham, Y. Hadjadj-Aoul, A. Outtagarts and G. Rubino, "On the Use of Graph Neural Networks for Virtual Network Embedding," in *International Symposium on Networks, Computers and Communications (ISNCC)*, 2020.

[36] F. Habibi, M. Dolati, A. Khonsari and M. Ghaderi, "Accelerating Virtual Network Embedding with Graph Neural Networks," in *16th International Conference on Network and Service Management (CNSM)*, 2020.

[37] N. Chen, P. Zhang, N. Kumar, C.-H. Hsu, L. Abualigah and H. Zhu, "Spectral graph theory-based virtual network embedding for vehicular fog computing: A deep reinforcement learning architecture," *Knowledge-Based Systems,* vol. 257, p. 109931, 2022.

[38] K. Rusek, J. Suárez-Varela, A. Mestres, P. Barlet-Ros and A. Cabellos-Aparicio, "Unveiling the potential of Graph Neural Networks for network modeling and optimization in SDN," in *ACM Symposium on SDN Research (SOSR '19)*, 2019.

[39] J. P. Martins, I. Almeida, R. Souza and S. Lins, "Policy Distillation for Real-Time Inference in Fronthaul Congestion Control," *IEEE Access,* vol. 9, pp. 154471-154483, 2021.

[40] F. Tonini, C. Raffaelli, S. Bjørnstad, D. Chen and R. Veisllari, "A Traffic Pattern Adaptive Mechanism to Bound Packet Delay and Delay Variation in 5G Fronthaul," in *European Conference on Networks and Communications (EuCNC)*, 2019.

[41] C. C. Zhang, K. K. Nguyen and M. Cheriet, "Routing and Packet Scheduling For Virtualized Disaggregate Functions in 5G O-RAN Fronthaul," in *IEEE Global Communications Conference (GLOBECOM)*, 2021.

[42] G. O. Pérez, J. A. Hernández and D. L. López, "Delay analysis of fronthaul traffic in 5G transport networks," in *IEEE 17th International Conference on Ubiquitous Wireless Broadband (ICUWB)*, 2017.

[43] L. M. C. David Larrabeiti a, G. O. J. A. Hernández and J. P. Fernandez-Palacios, "Toward end-to-end latency management of 5G network slicing and fronthaul traffic (Invited paper)," *Optical Fiber Technology,* vol. 76, 2023.

[44] N. Budhdev, R. Joshi, P. G. Kannan, M. C. Chan and T. Mitra, "FSA: fronthaul slicing architecture for 5G using dataplane programmable switches," in *27th Annual International Conference on Mobile Computing and Networking*, 2021.

[45] T. Balogh and M. Medvecký, "Performance evaluation of WFQ, WF2Q+ and WRR queue scheduling algorithms," in *34th International Conference on Telecommunications and Signal Processing (TSP)*, 2011.

[46] 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.

[47] ORAN Alliance, "O-RAN Xhaul Packet Switched Architectures and Solutions 7.0," 2024. [Online]. Available: https://specifications.o-ran.org/specifications.

[48] ORAN Alliance, "O-RAN Control, User and Synchronization Plane Specification 14.0," 2024. [Online]. Available: https://specifications.o-ran.org/specifications.

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