Decentralized AI and Architectures for Massive Wireless Network Slicing Scalability and Sustainability in 6G

ELASTIC

Grant No. TSI-063000-2021-54

E5: Final report on AI for the 6G DAWN AE/MS/DE

Financiado por la Unión Europea NextGenerationEU

Plan de Recuperación, Transformación y Resiliencia

UNICO I+D

GOBIERNO DE ESPAÑA

Centre Tecnològic de Telecomunicacions de Catalunya

Abstract

This report describes the implementation of decentralized 6G DAWN AI (MS-AE-DE) in ELASTIC PoCs in detail, their KPI measurements according to the defined ones in E3 and their lessons learned during the implementation and experiments.

Document properties

Document number	E5
Document title	Final report on AI for the 6G DAWN AE/MS/DE
Document responsible	Sarang Kahvazadeh (CTTC)
Document editor	Sarang Kahvazadeh (CTTC), Farhad Rezazadeh (CTTC), Selva Via (CTTC)
Authors	Farhad Rezazadeh (CTTC), Engin Zeydan (CTTC), Albert Bel (CTTC), Luis Blanco (CTTC), Fatemehsadat Tabatabaeimehr (CTTC) Farhana Javed (CTTC), Jorge Baranda (CTTC), Manuel Requena (CTTC), Josep Mangues-Bafalluy (CTTC), Oriol Font-Bach (SRS), Ismael Gomez (SRS), Manuel Lorenzo (Ericsson), Saravanan Kalimuthu (Ericsson), José Luis Jimenez (Ericsson), Marc Molla (Ericsson), Inmaculada Rafael (Ericsson), Miguel Ángel López Serrano (Ericsson), Álvaro Vlad (Ericsson), Carlos Javier Soleto Ramos (Ericsson), Rubén Cerezo (Ericsson), Diego San Cristobal Epalza (Ericsson), Fernando Beltrán González (Ericsson), Alejandro Ramiro Muñoz (Ericsson), Hristo Koshutanski (ATOS), Ignacio Labrador (ATOS), Manuel Jiménez (ATOS), Sonia Castro (ATOS), Jaime Azcorra (Telcaria), Aitor Zabala (Telcaria)
Target dissemination level	Public
Status of the document	Final
Version	1.0
Delivery date	31 December 2025
Actual delivery date	31 December 2025

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

Disclaimer

Contents

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

List of Figures

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

List of Tables

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

List of Acronyms

3GPP – 3rd Generation Partnership Project

5G-ACIA - 5G Alliance for Connected Industries and Automation

ACT- Actuator

AD – Anomaly Detection

AE – Analytics Engine

AF – Application Function

AI – Artificial Intelligence

B5G – Beyond 5G

CaaS – Container as a Service

CQI – Channel Quality Indicator

CNF – Containerized Network Function

COTS – Commercial Off-The-Shelf

CPU - Central Processing Unit

CSP – Communications Service Provider

CSI – Channel State Information

CU- Central Unit

DE – Decision Engine

DL – Downlink

DMO - Domain Manager and Orchestrator

DU- Distributed Unit

E2E – End To End

EC – Energy Consumption

eMBB – Enhanced Mobile Broadband

ETSI – European Telecommunications Standards Institute

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

GNSS – Global Navigation Satellite system

GPS – Global Positioning System

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

ICMP – Internet Control Message Protocol

IDMO- Inter-Domain Manager and Orchestrator

ILE- Infrastructure Layer Emulator

IRU – Indoor Radio Unit

ISPM – Infrastructure Status Prediction Module

KPI – Key Performance Indicator

KPM- Key Performance Measurement

MCData – Mission Critical Data

MCPTT – Mission Critical Push To Talk

MCVideo – Mission Critical Video

MCX – Mission Critical services

ML – Machine Learning

MPLS – Multi-Protocol Label Switching

M&O- Management and Orchestration

MS – Monitoring System

NDT – Network Digital Twin

NEF – Network Exposure Function (3gpp)

NETCONF – Network Configuration Protocol

NF – Network Function

NPN- Non-Public Networks

NR – New Radio

NWDAF – Network Data Analytics Function

OAM – Operations, Administration and Maintenance

O-RAN – Open Radio Access Network

PoC- Proof of Concept

OTA – Over-the-Air

OWD – One Way Delay

PCF – Policy Control Function

PN – Public Network (3gpp)

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

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

PPDR – Public Protection & Disaster Relief

PT-Paquete de Trabajo (Work package)

QoS – Quality of Service

RAN – Radio Access Network

RAN WG3 – RAN (Radio Access Network) Workgroup 3

R- RESILIENT

RB – Resource Block

RC – Radio Controller

RDI – Radio Dot Interface

RDS – Radio Dot System

RIC - RAN Intelligent Controller

RF – Radio Frequency

RLC – Radio Link Control

RT- Real Time

RTT - Round-Trip Time

SA1 – System Aspects Workgroup 1

SDN – Software Defined Networking

SDR-Software Define Radio

TDD – Time Division Duplex

TETRA – Terrestrial Trunked Radio

TS – Technical Specification

TR – Technical Report

UC- Use Case

UE – User Equipment

UL – Uplink

UPF- User Plane Function

URLLC – Ultra-reliable Low-Latency Communication

USA – United States of America

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

USRP - Universal Software Radio Peripheral

VPN – Virtual Private Network

WG- Working Group

xApp-Cross Application

YANG – Yet Another Next Generation

ZDM – Zero-Defect Manufacturing

ZSM – Zero-touch Service Management

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

Executive Summary

This document presents the implemented 6GDAWN AI closed-loop (MS-AE-DE) as defined in ELASTIC PoCs and their requirements according to Deliverable E3. This deliverable illustrates the AI closed-loop (MS-AE-DE) implementation, KPI measurement methods, and lessons learned across ELASTIC PoCs. The energy efficiency vs. QoS use case targets energy efficiency while preserving Quality of Service (QoS), and the end-to-end optimization use case focuses on holistic resource optimization. We conclude deliverable with successful implementation of decentralized AI (MS-AE-DE) in ELASTIC PoCs while satisfying KPI measurements defined in E3.

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

1 Introduction

This deliverable reports the final results of 6G DAWN's decentralized AI closed loop—Monitoring System, Analytics Engine, and Decision Engine (MS–AE–DE)—validated through Proofs of Concept (PoCs) targeting elasticity and end-to-end optimization. The work is grounded in cross-domain, real-world deployments with measurable KPIs and repeatable control loops that manage energy and performance. The energy efficiency vs. QoS use case shows that AI-driven adaptation can lower energy consumption while upholding Quality of Service (QoS), while the E2E optimization use case demonstrates how proactive prediction sustains efficient service continuity across the cloud–edge continuum.

Concretely, elasticity and network optimization are demonstrated through: (i) a dynamic RAN xApp that adjusts operative parameters to match traffic load and user populations, improving resource utilization and reducing energy use; (ii) a Network Digital Twin within the 3GPP PNI-NPN framework that applies advanced ML to recommend energy-aware NPN configurations, balancing computational efficiency with interpretability; and (iii) a Proactive Infrastructure Status Prediction Module aligned with continuum orchestration, bridging volatile extreme-edge domains with stable cloud and edge resources.

The mentioned PoCs operationalize the decentralized AI (MS–AE–DE) loop over the defined interfaces, turning telemetry into timely, auditable action.

In the following sections, we describe ELASTIC PoCs implementation details, KPIs measurements and lessons learned according to E3. In table 1, a summary of Use case and PoC mapping with decentralized AI (MS-AE-DE) is illustrated.

TABLE 1. ARCHITECTURAL OVERVIEW

PoC	MS	AE	DE	ACT
E-UC1-PoC1	xApp together with near-RT RIC	xApp	xApp	xApp
E-UC1-PoC2	AF, NDT platform/MS)	NDT platform/AE (network-level) AF (application-level)	AF	NDT platform/ACT
E-UC2-PoC1	MS	ISPM	Service Mobility Orchestrator	Service Mobility Orchestrator

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 ELASTIC Use Cases

Two different use cases, which comprise a total of three PoCs, are defined under 6G DAWN ELASTIC project:

- Use Case E1 Energy Efficiency vs QoS
  - E UC1 PoC1 - Dynamic RAN xApp
  - E UC1 PoC2 - Energy optimization system in NPNs
- Use Case E2 E2E network resources optimization
  - E UC2 PoC1- Proactive infrastructure status prediction module

The following table presents the mapping of Key concepts and PoCs that apply for the ELASTIC project:

TABLE 2. MAPPING OF KEY CONCEPTS AND POCS

Key Concepts	**E UC1 PoC1**	**E UC1 PoC2**	**E UC2 PoC1**
NPN Digital Twin System		X	
Extreme Edge			X
AI/ML agent for control loops			X
xApps in O-RAN	X		
Relation of vertical KPIs with the network configuration		X	
Inter(a)-slice reconfiguration and massive slicing		X	
NEF instance for KPI data and configuration capabilities exposures of NPNs		X	
AI/ML methods for reducing energy consumption	X	X	

2.1 Use Case ELASTIC Energy Efficiency vs QoS

2.1.1 ELASTIC UC1 PoC1 –Dynamic RAN xApp

The goal of the PoC is to show that it is possible to dynamically modify relevant operative aspects of the RAN to fit the current traffic load and number of users, thus optimizing the utilization of resources and, in turn, reducing the resulting energy consumption. Figure 1 shows the proposed PoC evaluation scenario.

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

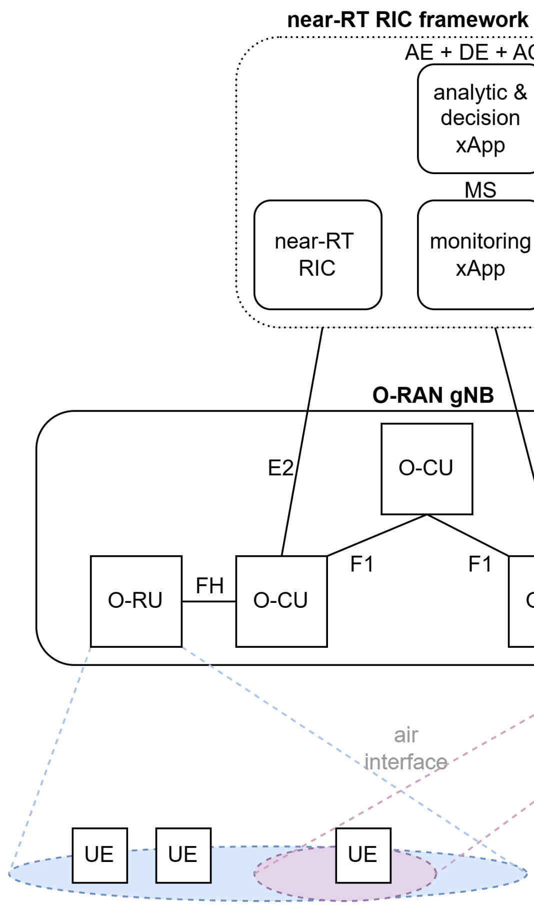


FIGURE 1. E UC1 POC1 EVALUATION SCENARIO

The PoC builds upon O-RAN infrastructure, leveraging its AI-friendly nature, which facilitates the monitoring and dynamic adaptation of the RAN. Towards this end, O-RAN has defined the E2 interface, which enables connecting the gNB in the PoC scenario with the near-real-time (RT) RAN intelligent controller (RIC), which is able to query and control it. In more detail, O-RAN has defined the E2 Key Performance Measurement (KPM) service model (E2SM-KPM), which allows to obtain live measurements from the RAN and provide them to the near-RT RIC. Likewise, an E2 RAN Control (RC) service model (E2SM-RC) has also been defined to modify different RAN aspects at run-time. In the near-RT RIC a predictive AI algorithm will be able to forecast the load of the network by working with the measurements. This enables taking AI-based decisions and implementing the required actions to optimize both the resource usage and energy consumption of the RAN.

As observed in Figure 1, the PoC deploys a RAN with two different cells: a coverage cell and a capacity one. During peak hours, the capacity cell enables those users requiring it to fulfill its high data-rate communication needs, whereas the coverage cell constitutes a wider-reach mid-to-low data-rate link for the rest of users. In off-peak periods, it is expected that both the number of RAN-connected

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

users, as well as its data-consumption requirements, will be dramatically reduced. In this situation, the capacity cell might be switched off, while the connected users hand-over to the coverage one, providing a significant reduction of both RAN resource usage and energy consumption. Likewise, as more users connect back to the RAN and the communication requirements scale up, the capacity cell will be switched on again, providing the required additional RAN resources.

In the evaluation scenario proposed by E UC1 PoC1, the near-RT RIC acts as MS (receives the RAN measures), whereas the AI-based prediction algorithm acts as AE (analyzes the RAN measures), DE (takes a decision based on the evolution of the measures) and ACT (triggers the capacity cell switching on/off).

2.1.1.1 PoC Implementation details

Figure 2 shows the components comprising E UC1 PoC1. As it can be observed, the implementation foundations of the PoC are provided by the srsRAN Project[1], an open-source O-RAN native 5G CU and DU SDR-based implementation from SRS. Amongst its many features, the PoC relies on its disaggregated design nature to deploy a single CU and two DUs, that implement L1, L2 and L3 of the coverage and capacity cells. Additionally, the gNB also provides the E2SM-KPM features detailed in Table 3, and the E2SM-RC actions detailed in Table 4. To complete the cells two commercially available radio units (RU) from Benetel are providing the low physical layer, digital and radio frequency (RF) front-end functions of each cell. In more detail, the PoC adopts an O-RAN split 7.2b configuration, where the DUs (hosted in a computing server) are connected to the remote RUs via the front-haul interface (i.e., employing a fiber link). The other key component of E UC1 PoC1 is the AI framework developed by CTTC with a RAN traffic load forecasting algorithm at its core. The forecasting of RAN traffic is based on run-time measurements received from the DUs, through RIC, via the E2 interface. Hence, this AI-based algorithm is executed as an xApp in the RIC. The O-RAN Software Community (SC) Near-Real-time RIC[2] is used in the PoC. Finally, the 5G core network is implemented by the Open5GS[3] open-source solution. All components are interconnected between them using the corresponding O-RAN defined interfaces, independently of their disaggregated deployment (i.e., all components might be hosted in different machines, interconnected by ethernet links). As for the UE and DU communication, it will happen over-the-air, although a cabled setup is also possible and might be utilized to avoid undesired interferences from other nearby laboratory setups operating in the same frequency bands.

[1] srsRAN Project, https://github.com/srsran/srsRAN_Project

[2] O-RAN SC RIC, https://docs.o-ran-sc.org/en/latest/

[3] Open 5GS, https://open5gs.org/

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

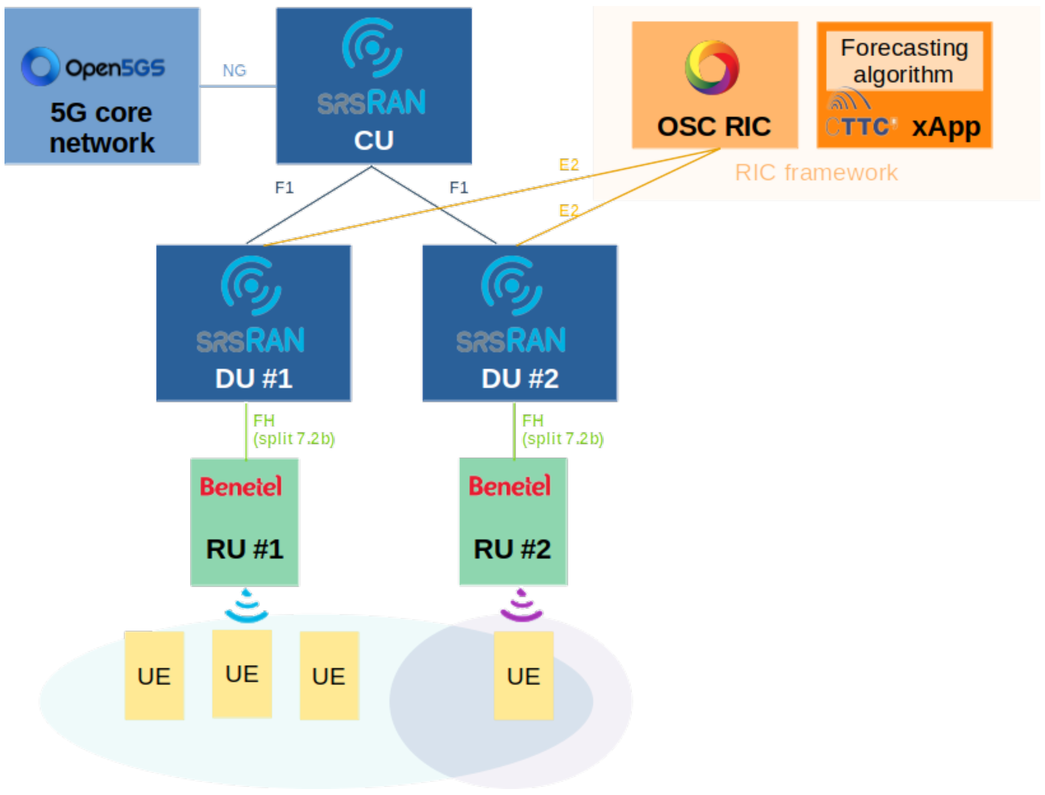


FIGURE 2. KEY COMPONENTS IN E UC1 POC1

TABLE 3. E2 TRAFFIC-LOAD METRICS PROVIDED BY SRSRAN PROJECT IN E UC1 POC1

Category	**Name**	**Description**
Data Radio Bearer	DRB.UEThpDl	Average DL UE throughput in the gNB-DU
	DRB.UEThpUl	Average UL UE throughput in the gNB-DU
Radio Resource Utilization	RRU.PrbAvailDl	DL total available PRB
	RRU.PrbAvailUl	UL total available PRB
	RRU.PrbUsedUl	Mean UL PRB used for data traffic
	RRU.PrbUsedDl	Mean DL PRB used for data traffic

TABLE 4. E2 RAN CONTROL ACTIONS PROVIDED BY SRSRAN PROJECT IN E UC1 POC1

Category	**Name**	**Description**
Connected Mode Mobility Control	Handover Control	Enable traffic steering (i.e., trigger handover of UEs between DUs/cells)

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

In terms of hardware equipment, the PoC comprises the elements listed in Table 5 and set up as shown in Figure 3. E UC1 PoC1 laboratory setup.

TABLE 5. HARDWARE COMPONENTS IN E UC1 POC1

Element	Key specifications	Units	Photo
O-RU Benetel RAN550	• For indoor use • Band n78 • Up to 100 MHz signal bandwidth • Up to 4T4R • Embedded active antennas	2	
General purpose computing server (using the listed components or equivalent ones)	• 24-core CPU (3 GHz) • 64 GB DDR5 (4800 MHz) • 2-port Intel E810 25GE NIC • Unix OS	2	
Falcon-RX/812/G xHaul switch & PTP Grandmaster	• 12x 10GE SFP+ ports • 8x 25GE SFP+ ports • Acts as PTP Grandmaster and boundray clock (e.g., LLS-C1) • GNSS sync support • VLAN tagging support	1	
Amari UE Simbox Series	• NR UE (real-time) UE simulator • Up to 1000x UEs • 4x RF front-end cards (50 MHz, 2T2R)	1	

A disaggregated gNB deployment is employed by E UC1 PoC1, as shown in Figure EUC1POC1-3. The 5G core network is collocated with the O-CU and one of the O-DUs, whereas the other O-DU is collocated with the near-RT RIC framework (which includes the AI-based traffic-forecasting xApp). A 25GE-enabled ethernet switch, with PTP Grand Master features, provides the required interconnection amongst the computing hosts and the O-RUs. Finally, a third computing server provides virtualized UE machinery.

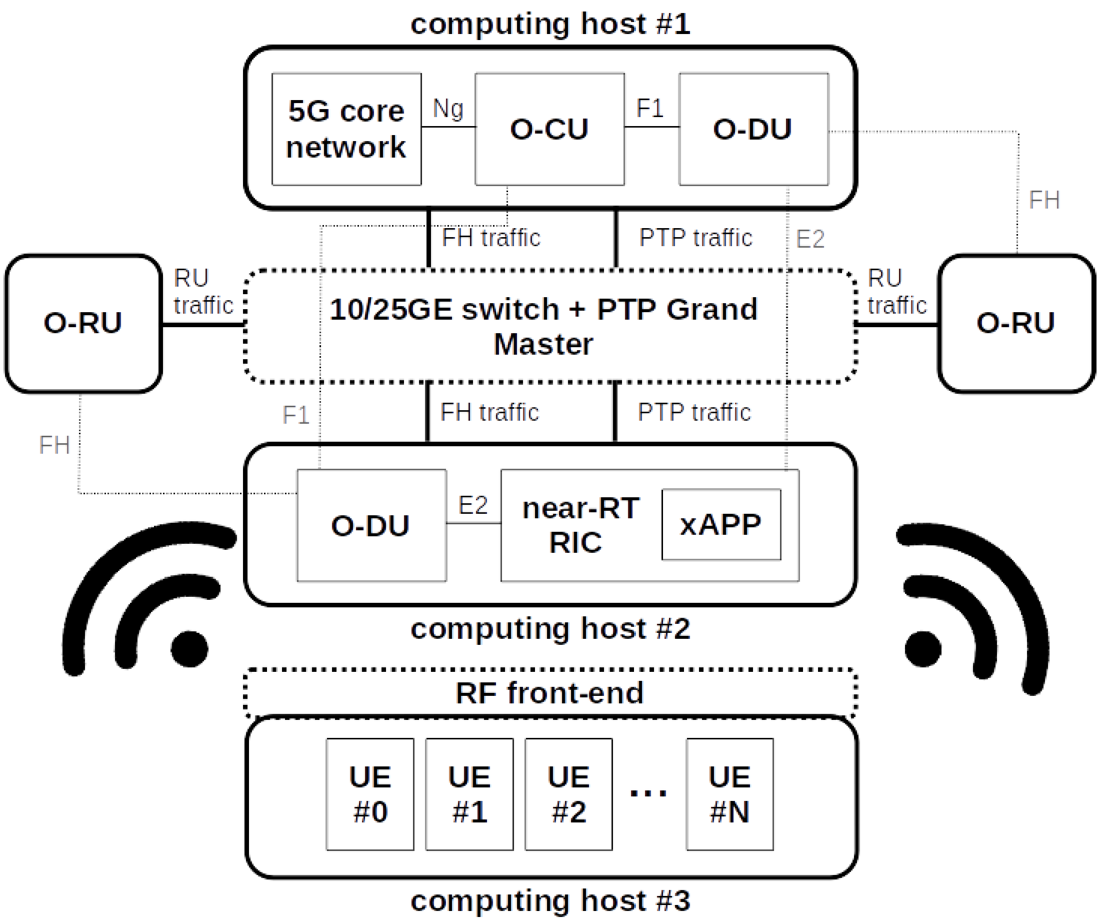


FIGURE 3. E UC1 POC1 LABORATORY SETUP

As detailed above, the goal of the PoC is to demonstrate how AI-based run-time RAN adaptation can help optimizing the energy footprint of 6G networks. Towards that end, a simple two-cell scenario is implemented, allowing to disable one of them when the number of connected users and their compound communication requirements allows it.

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

RF elements (i.e., the O-RUs in E UC1 PoC1) are known to be amongst the most power-hungry components in the RAN[4]. For this reason, powering off the O-RU utilized by the capacity cell provides very relevant energy savings. Moreover, the computational load of the associated O-RAN implementation is also greatly reduced (i.e., only the CU and one DU – the one implementing the coverage cell - need to stay active, whereas the other DU – the one in the capacity cell that is being switched off – will remain in a standby state), which also helps to reduce the resulting energy consumption notably. Note that UEs connected to the capacity cell will be forced to handover to the coverage cell when the first is switched off, to ensure no service disruption takes place.

2.1.1.2 E UC1 PoC1 KPIs Evaluation and Results

We first recall the KPIs defined for Elastic UC1 PoC1. Table 6 shows the KPIs as defined in E3 for the E UC1 PoC1, and their refence points of evaluation defined in the final pilot implementation.

TABLE 6: RESILIENT UC1 POC1 KPIS AND EVALUATION POINTS IN FINAL POC IMPLEMENTATION

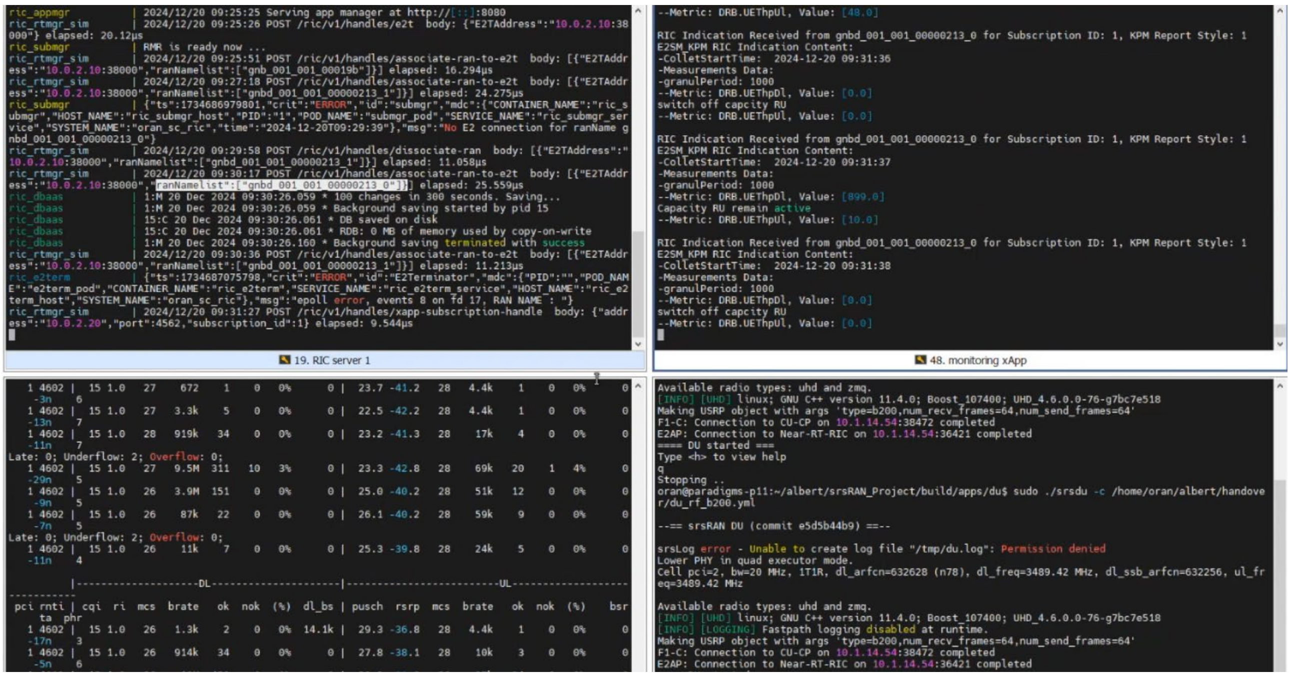
KPI	Unit	Type	Definition	Evaluation Points
DL throughput per UE	Kbit/s	Performance	Number of bits contained in the SDUs (Service Data Units) delivered to Layer 3 that can be transferred from the network to a UE over a certain period.	Throughput extracted thanks to the KPM traffic loads
UL throughput per UE	Kbit/s	Performance	Number of bits contained in the SDUs (Service Data Units) delivered to Layer 3 that can be transferred from the network to a UE over a certain period.	Throughput extracted thanks to the KPM traffic loads
NG-RAN EC	W	Energy	EC of all gNBs. EC at each gNB is computed as the sum of all NFs that constitute the gNB.	Consumption of RUs in watts

[4] Baldini, Gianmarco & Bolla, Raffaele & Bruschi, Roberto & Carrega, Alessandro & Davoli, Franco & Lombardo, C. & Rabbani, Ramin. (2024). Toward Sustainable O-RAN Deployment: An In-Depth Analysis of Power Consumption. IEEE Transactions on Green Communications and Networking. PP. 1-1. 10.1109/TGCN.2024.3426108.

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

This Proof of Concept (PoC) investigates the impact of intelligent RAN control on performance and energy efficiency using an O-RAN–compliant architecture. The deployment consists of two Radio Units (RUs), each connected to an independent Distributed Unit (DU), with both DUs aggregated at a Central Unit (CU).

A dedicated KPM Collection xApp retrieves near–real-time Key Performance Measurements (KPMs) from the E2 nodes. These KPMs include metrics such as throughput, PRB utilization, and RU power consumption (see Table 3). The collected data is stored in an InfluxDB time-series database and visualized through Grafana dashboards, enabling continuous monitoring of RU behavior and network performance.


```
ric_appmgr          | 2024/12/20 09:25:25 Serving app manager at http://[::]:8080
ric_rtmgr_sim       | 2024/12/20 09:25:26 POST /ric/v1/handles/e2t  body: {"E2TAddress":"10.0.2.10:38
000"} elapsed: 20.12µs
ric_submgr          | RMR is ready now ...
ric_rtmgr_sim       | 2024/12/20 09:25:51 POST /ric/v1/handles/associate-ran-to-e2t  body: [{"E2TAddr
ess":"10.0.2.10:38000","ranNamelist":["gnb_001_001_00019b"]}] elapsed: 16.294µs
ric_rtmgr_sim       | 2024/12/20 09:27:18 POST /ric/v1/handles/associate-ran-to-e2t  body: [{"E2TAddr
ess":"10.0.2.10:38000","ranNamelist":["gnbd_001_001_00000213_1"]}] elapsed: 24.275µs
ric_submgr          | {"ts":1734686979801,"crit":"ERROR","id":"submgr","mdc":{"CONTAINER_NAME":"ric_s
ubmgr","HOST_NAME":"ric_submgr_host","PID":"1","POD_NAME":"submgr_pod","SERVICE_NAME":"ric_submgr_ser
vice","SYSTEM_NAME":"oran_sc_ric","time":"2024-12-20T09:29:39"},"msg":"No E2 connection for ranName g
nbd_001_001_00000213_0"}
ric_rtmgr_sim       | 2024/12/20 09:29:58 POST /ric/v1/handles/dissociate-ran  body: [{"E2TAddress":"
10.0.2.10:38000","ranNamelist":["gnbd_001_001_00000213_1"]}] elapsed: 11.058µs
ric_rtmgr_sim       | 2024/12/20 09:30:17 POST /ric/v1/handles/associate-ran-to-e2t  body: [{"E2TAddr
ess":"10.0.2.10:38000","ranNamelist":["gnbd_001_001_00000213_0"]}] elapsed: 25.559µs
ric_dbaas           | 1:M 20 Dec 2024 09:30:26.059 * 100 changes in 300 seconds. Saving...
ric_dbaas           | 1:M 20 Dec 2024 09:30:26.059 * Background saving started by pid 15
ric_dbaas           | 15:C 20 Dec 2024 09:30:26.061 * DB saved on disk
ric_dbaas           | 15:C 20 Dec 2024 09:30:26.061 * RDB: 0 MB of memory used by copy-on-write
ric_dbaas           | 1:M 20 Dec 2024 09:30:26.160 * Background saving terminated with success
ric_rtmgr_sim       | 2024/12/20 09:30:36 POST /ric/v1/handles/associate-ran-to-e2t  body: [{"E2TAddr
ess":"10.0.2.10:38000","ranNamelist":["gnbd_001_001_00000213_1"]}] elapsed: 11.213µs
ric_e2term          | {"ts":1734687075798,"crit":"ERROR","id":"E2Terminator","mdc":{"PID":"","POD_NAM
E":"e2term_pod","CONTAINER_NAME":"ric_e2term","SERVICE_NAME":"ric_e2term_service","HOST_NAME":"ric_e2
term_host","SYSTEM_NAME":"oran_sc_ric"},"msg":"epoll error, events 8 on fd 17, RAN NAME : "}
ric_rtmgr_sim       | 2024/12/20 09:31:27 POST /ric/v1/handles/xapp-subscription-handle  body: {"addr
ess":"10.0.2.20","port":4562,"subscription_id":1} elapsed: 9.544µs
```

```
--Metric: DRB.UEThpUl, Value: [48.0]

RIC Indication Received from gnbd_001_001_00000213_0 for Subscription ID: 1, KPM Report Style: 1
E2SM_KPM RIC Indication Content:
-ColletStartTime:  2024-12-20 09:31:36
-Measurements Data:
-granulPeriod: 1000
--Metric: DRB.UEThpDl, Value: [0.0]
switch off capcity RU
--Metric: DRB.UEThpUl, Value: [0.0]

RIC Indication Received from gnbd_001_001_00000213_0 for Subscription ID: 1, KPM Report Style: 1
E2SM_KPM RIC Indication Content:
-ColletStartTime:  2024-12-20 09:31:37
-Measurements Data:
-granulPeriod: 1000
--Metric: DRB.UEThpDl, Value: [899.0]
Capacity RU remain active
--Metric: DRB.UEThpUl, Value: [10.0]

RIC Indication Received from gnbd_001_001_00000213_0 for Subscription ID: 1, KPM Report Style: 1
E2SM_KPM RIC Indication Content:
-ColletStartTime:  2024-12-20 09:31:38
-Measurements Data:
-granulPeriod: 1000
--Metric: DRB.UEThpDl, Value: [0.0]
switch off capcity RU
--Metric: DRB.UEThpUl, Value: [0.0]
```

```
   1 4602 |  15 1.0   27    672    1    0   0%      0 |  23.7 -41.2   28   4.4k    1    0   0%      0
    -3n    6
   1 4602 |  15 1.0   27   3.3k    5    0   0%      0 |  22.5 -42.2   28   4.4k    1    0   0%      0
   -13n    7
   1 4602 |  15 1.0   28   919k   34    0   0%      0 |  23.2 -41.3   28    17k    4    0   0%      0
   -11n    7
Late: 0; Underflow: 2; Overflow: 0;
   1 4602 |  15 1.0   27   9.5M  311   10   3%      0 |  23.3 -42.8   28    69k   20    1   4%      0
   -29n    5
   1 4602 |  15 1.0   26   3.9M  151    0   0%      0 |  25.0 -40.2   28    51k   12    0   0%      0
    -9n    5
   1 4602 |  15 1.0   26    87k   22    0   0%      0 |  26.1 -40.2   28    59k    9    0   0%      0
    -7n    5
Late: 0; Underflow: 2; Overflow: 0;
   1 4602 |  15 1.0   26    11k    7    0   0%      0 |  25.3 -39.8   28    24k    5    0   0%      0
   -11n    4

          |--------------------DL--------------------|-----------------------UL-------------------
----------
 pci rnti | cqi  ri  mcs  brate   ok  nok  (%)  dl_bs | pusch  rsrp  mcs  brate   ok  nok  (%)    bsr
    ta  phr
   1 4602 |  15 1.0   26   1.3k    2    0   0%  14.1k |  29.3 -36.8   28   4.4k    1    0   0%      0
   -17n    3
   1 4602 |  15 1.0   26   914k   34    0   0%      0 |  27.8 -38.1   28    10k    3    0   0%      0
    -5n    6
```

```
Available radio types: uhd and zmq.
[INFO] [UHD] linux; GNU C++ version 11.4.0; Boost_107400; UHD_4.6.0.0-76-g7bc7e518
Making USRP object with args 'type=b200,num_recv_frames=64,num_send_frames=64'
F1-C: Connection to CU-CP on 10.1.14.54:38472 completed
E2AP: Connection to Near-RT-RIC on 10.1.14.54:36421 completed
==== DU started ===
Type <h> to view help
q
Stopping ..
oran@paradigms-p11:~/albert/srsRAN_Project/build/apps/du$ sudo ./srsdu -c /home/oran/albert/handove
r/du_rf_b200.yml

--== srsRAN DU (commit e5d5b44b9) ==--

srsLog error - Unable to create log file "/tmp/du.log": Permission denied
Lower PHY in quad executor mode.
Cell pci=2, bw=20 MHz, 1T1R, dl_arfcn=632628 (n78), dl_freq=3489.42 MHz, dl_ssb_arfcn=632256, ul_fr
eq=3489.42 MHz

Available radio types: uhd and zmq.
[INFO] [UHD] linux; GNU C++ version 11.4.0; Boost_107400; UHD_4.6.0.0-76-g7bc7e518
[INFO] [LOGGING] Fastpath logging disabled at runtime.
Making USRP object with args 'type=b200,num_recv_frames=64,num_send_frames=64'
F1-C: Connection to CU-CP on 10.1.14.54:38472 completed
E2AP: Connection to Near-RT-RIC on 10.1.14.54:36421 completed
```

FIGURE 4. IMAGE OF MONITORING XAPP AND TRAFFIC AT BOTH DUS

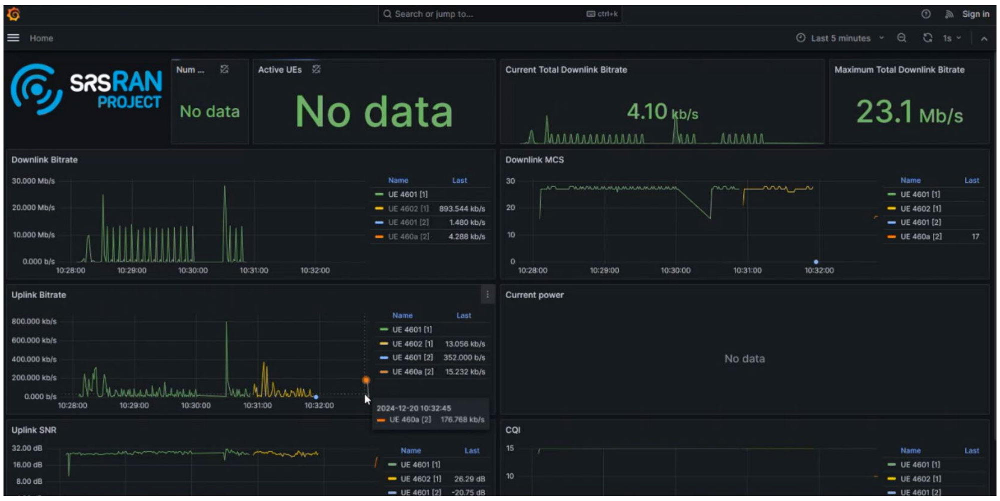
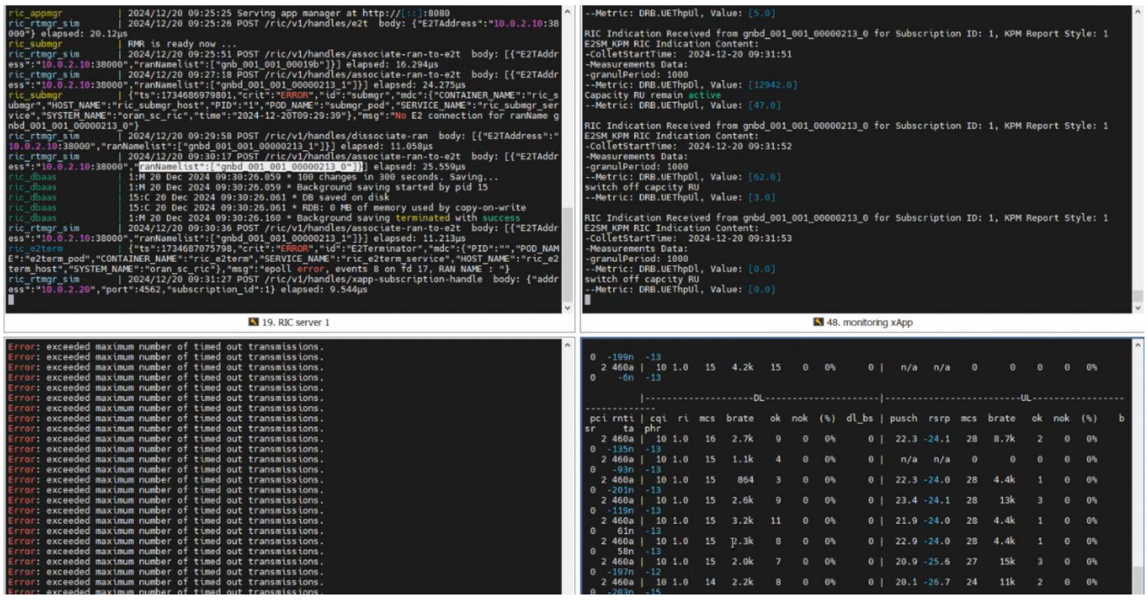


FIGURE 5. GRAFANA SHOWING KPMS COLLECTED BY THE MONITORING XAPP

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

A second Control xApp processes the collected KPIs and implements decision-making logic for dynamic energy optimization. This xApp evaluates RAN conditions—specifically PRB utilization and traffic distribution—to determine whether a handover (HO) should be triggered from the *capacity RU* to the *coverage RU*. This decision will be done if the prediction of the traffic decreases and the coverage cell is able to allocate resources to the UEs to be handoff. On the one hand, traffic predicted. If both conditions are accomplished, the xApp migrates active UEs to the coverage RU. Following the completion of the HO, the system powers down the capacity RU, enabling an approximate 50% reduction in RU-related energy consumption, as one of the two RUs becomes inactive.



```
ric_appmgr          | 2024/12/20 09:25:25 Serving app manager at http://[::]:8080
ric_rtmgr_sim       | 2024/12/20 09:25:26 POST /ric/v1/handles/e2t  body: {"E2TAddress":"10.0.2.10:38
000"} elapsed: 20.12µs
ric_submgr          | RMR is ready now ...
ric_rtmgr_sim       | 2024/12/20 09:25:51 POST /ric/v1/handles/associate-ran-to-e2t  body: [{"E2TAddr
ess":"10.0.2.10:38000","ranNamelist":["gnb_001_001_00019b"]}] elapsed: 16.294µs
ric_rtmgr_sim       | 2024/12/20 09:27:18 POST /ric/v1/handles/associate-ran-to-e2t  body: [{"E2TAddr
ess":"10.0.2.10:38000","ranNamelist":["gnbd_001_001_00000213_1"]}] elapsed: 24.275µs
ric_submgr          | {"ts":1734686979801,"crit":"ERROR","id":"submgr","mdc":{"CONTAINER_NAME":"ric_s
ubmgr","HOST_NAME":"ric_submgr_host","PID":"1","POD_NAME":"submgr_pod","SERVICE_NAME":"ric_submgr_ser
vice","SYSTEM_NAME":"oran_sc_ric","time":"2024-12-20T09:29:39"},"msg":"No E2 connection for ranName g
nbd_001_001_00000213_0"}
ric_rtmgr_sim       | 2024/12/20 09:29:58 POST /ric/v1/handles/dissociate-ran  body: [{"E2TAddress":"
10.0.2.10:38000","ranNamelist":["gnbd_001_001_00000213_1"]}] elapsed: 11.058µs
ric_rtmgr_sim       | 2024/12/20 09:30:17 POST /ric/v1/handles/associate-ran-to-e2t  body: [{"E2TAddr
ess":"10.0.2.10:38000","ranNamelist":["gnbd_001_001_00000213_0"]}] elapsed: 25.559µs
ric_dbaas           | 1:M 20 Dec 2024 09:30:26.059 * 100 changes in 300 seconds. Saving...
ric_dbaas           | 1:M 20 Dec 2024 09:30:26.059 * Background saving started by pid 15
ric_dbaas           | 15:C 20 Dec 2024 09:30:26.061 * DB saved on disk
ric_dbaas           | 15:C 20 Dec 2024 09:30:26.061 * RDB: 0 MB of memory used by copy-on-write
ric_dbaas           | 1:M 20 Dec 2024 09:30:26.160 * Background saving terminated with success
ric_rtmgr_sim       | 2024/12/20 09:30:36 POST /ric/v1/handles/associate-ran-to-e2t  body: [{"E2TAddr
ess":"10.0.2.10:38000","ranNamelist":["gnbd_001_001_00000213_1"]}] elapsed: 11.213µs
ric_e2term          | {"ts":1734687075798,"crit":"ERROR","id":"E2Terminator","mdc":{"PID":"","POD_NAM
E":"e2term_pod","CONTAINER_NAME":"ric_e2term","SERVICE_NAME":"ric_e2term_service","HOST_NAME":"ric_e2
term_host","SYSTEM_NAME":"oran_sc_ric"},"msg":"epoll error, events 8 on fd 17, RAN NAME : "}
ric_rtmgr_sim       | 2024/12/20 09:31:27 POST /ric/v1/handles/xapp-subscription-handle  body: {"addr
ess":"10.0.2.20","port":4562,"subscription_id":1} elapsed: 9.544µs
```

19. RIC server 1

```
--Metric: DRB.UEThpUl, Value: [5.0]

RIC Indication Received from gnbd_001_001_00000213_0 for Subscription ID: 1, KPM Report Style: 1
E2SM_KPM RIC Indication Content:
-ColletStartTime:  2024-12-20 09:31:51
-Measurements Data:
-granulPeriod: 1000
--Metric: DRB.UEThpDl, Value: [12942.0]
Capacity RU remain active
--Metric: DRB.UEThpUl, Value: [47.0]

RIC Indication Received from gnbd_001_001_00000213_0 for Subscription ID: 1, KPM Report Style: 1
E2SM_KPM RIC Indication Content:
-ColletStartTime:  2024-12-20 09:31:52
-Measurements Data:
-granulPeriod: 1000
--Metric: DRB.UEThpDl, Value: [62.0]
switch off capcity RU
--Metric: DRB.UEThpUl, Value: [3.0]

RIC Indication Received from gnbd_001_001_00000213_0 for Subscription ID: 1, KPM Report Style: 1
E2SM_KPM RIC Indication Content:
-ColletStartTime:  2024-12-20 09:31:53
-Measurements Data:
-granulPeriod: 1000
--Metric: DRB.UEThpDl, Value: [0.0]
switch off capcity RU
--Metric: DRB.UEThpUl, Value: [0.0]
```

48. monitoring xApp

```
Error: exceeded maximum number of timed out transmissions.
Error: exceeded maximum number of timed out transmissions.
Error: exceeded maximum number of timed out transmissions.
Error: exceeded maximum number of timed out transmissions.
Error: exceeded maximum number of timed out transmissions.
Error: exceeded maximum number of timed out transmissions.
Error: exceeded maximum number of timed out transmissions.
Error: exceeded maximum number of timed out transmissions.
Error: exceeded maximum number of timed out transmissions.
Error: exceeded maximum number of timed out transmissions.
Error: exceeded maximum number of timed out transmissions.
Error: exceeded maximum number of timed out transmissions.
Error: exceeded maximum number of timed out transmissions.
Error: exceeded maximum number of timed out transmissions.
Error: exceeded maximum number of timed out transmissions.
Error: exceeded maximum number of timed out transmissions.
Error: exceeded maximum number of timed out transmissions.
Error: exceeded maximum number of timed out transmissions.
Error: exceeded maximum number of timed out transmissions.
Error: exceeded maximum number of timed out transmissions.
Error: exceeded maximum number of timed out transmissions.
Error: exceeded maximum number of timed out transmissions.
Error: exceeded maximum number of timed out transmissions.
Error: exceeded maximum number of timed out transmissions.
Error: exceeded maximum number of timed out transmissions.
```

```
 0  -199n  -13
   2 460a |  10 1.0   15   4.2k   15    0   0%       0 |   n/a   n/a    0      0    0    0   0%
 0    -6n  -13

          |-------------------DL-------------------|-----------------------UL----------------
-------------
 pci rnti | cqi  ri  mcs  brate   ok  nok  (%)  dl_bs | pusch  rsrp  mcs  brate   ok  nok  (%)    b
sr      ta  phr
   2 460a |  10 1.0   16   2.7k    9    0   0%       0 |  22.3 -24.1   28   8.7k    2    0   0%
 0  -135n  -13
   2 460a |  10 1.0   15   1.1k    4    0   0%       0 |   n/a   n/a    0      0    0    0   0%
 0   -93n  -13
   2 460a |  10 1.0   15    864    3    0   0%       0 |  22.3 -24.0   28   4.4k    1    0   0%
 0  -201n  -13
   2 460a |  10 1.0   15   2.6k    9    0   0%       0 |  23.4 -24.1   28    13k    3    0   0%
 0  -119n  -13
   2 460a |  10 1.0   15   3.2k   11    0   0%       0 |  21.9 -24.0   28   4.4k    1    0   0%
 0    61n  -13
   2 460a |  10 1.0   15   2.3k    8    0   0%       0 |  22.9 -24.0   28   4.4k    1    0   0%
 0    58n  -13
   2 460a |  10 1.0   15   2.0k    7    0   0%       0 |  20.9 -25.6   27    15k    3    0   0%
 0  -197n  -12
   2 460a |  10 1.0   14   2.2k    8    0   0%       0 |  20.1 -26.7   24    11k    2    0   0%
 0  -203n  -15
```

FIGURE 6. HO DONE BY THE CONTROL XAPP TO THE COVERAGE DU

To enhance decision accuracy and anticipate future network conditions, the PoC incorporates state-of-the-art (SotA) AI-driven multivariate probabilistic forecasting methods. Specifically, techniques such as Temporal Fusion Transformers (TFTs) and Deep Vector AutoRegressive (Deep VAR) models are evaluated.

These models provide several advantages over classical AI-based forecasting techniques (e.g., LSTM, GRU), which typically generate *single-point predictions*. Such point forecasts often give a misleading sense of confidence and lack the uncertainty quantification needed for safe, autonomous decision-making in real RAN environments.

In contrast, probabilistic forecasting models offer:

- A full predictive distribution rather than a single-value estimate
- Uncertainty quantification, including confidence intervals around predicted load
- Robustness to outliers and anomalies, improving stability in volatile traffic conditions
- Better support for strategic and risk-aware decision-making in Distributed Elements (DEs)

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

By leveraging these probabilistic outputs, the system can predict future PRB load and determine whether the capacity RU will likely remain underutilized. If low load is forecasted *and* the coverage RU is predicted to have sufficient resources to serve all connected UEs, the control logic proactively schedules:

1. Handover from the capacity RU to the coverage RU
2. Powering off of the capacity RU to reduce energy consumption

This proactive energy-saving mechanism is substantially more reliable because decisions are informed not only by observed KPIs but also by predicted traffic patterns with quantified uncertainty.

Once the handover is completed and the traffic is fully offloaded, the system proceeds to shut down the capacity RU. This action eliminates its power draw, enabling a quantifiable energy reduction. In this PoC, disabling one of the two RUs results in an approximate 50% energy savings, directly corresponding to the removal of one RU's consumption from the total network energy footprint.

Overall, this PoC demonstrates how O-RAN's open interfaces and xApp-driven RAN Intelligence Controller (RIC) enable closed-loop automation. By combining fine-grained KPI monitoring with intelligent control policies, the system dynamically optimizes radio resource usage and significantly reduces energy consumption without compromising service continuity.

2.1.1.3 Achievements and lessons learned

This Proof of Concept demonstrates how AI-driven probabilistic forecasting, combined with xApp-based automation, can reduce energy consumption in an O-RAN environment by dynamically managing Radio Unit activity without impacting service quality. The setup includes two Radio Units—one acting as a capacity layer and the other as a coverage layer—each connected to its own Distributed Unit and aggregated at a Central Unit. A first xApp collects near–real-time KPIs such as PRB load, throughput, and power consumption, storing them in InfluxDB and visualizing them through Grafana. A second xApp implements control logic, taking decisions on handover execution and RU deactivation based on the monitored KPIs.

A key innovation in this PoC is the use of state-of-the-art multivariate probabilistic forecasting models such as Temporal Fusion Transformers (TFTs) and Gaussian-Process Vector Autoregressive (GPVAr). Unlike classical LSTM or GRU models that only produce single-point predictions, these models generate full predictive distributions and quantify uncertainty, offering a more realistic and reliable view of future network conditions. This uncertainty-aware forecasting improves robustness against anomalies and allows the system to make safer decisions. When the forecast indicates sustained low PRB load on the capacity RU and confirms that the coverage RU will have sufficient resources to serve all users, the system proactively triggers a handover and subsequently switches off the capacity RU.

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 results show that shutting down one RU leads to approximately a 50% reduction in RU-related energy consumption while maintaining service continuity. The PoC also highlights several lessons: the critical importance of high-quality KPI data; the value of probabilistic forecasting for improving the confidence of autonomous decisions; the need for tight synchronization between monitoring, forecasting, decision-making, and actuation; the complexity of integrating forecasting modules and xApps in the O-RAN architecture; and the essential role of strong observability tools such as Grafana and InfluxDB. These insights will guide next steps, including extending forecasting to additional KPIs, scaling the solution to multi-cell environments, enabling more adaptive RIC policies, and assessing long-term energy gains in operational networks.

2.1.2 ELASTIC UC1 PoC2 -Energy optimization system in NPNs

Energy efficiency in NPNs is crucial for environmental sustainability, cost efficiency, network reliability, and user experience. By implementing energy-efficient practices, we can reduce the carbon footprint, optimize resource utilization, and contribute to a greener and more sustainable 6G ecosystem. Our goal is to research mechanisms for evaluating, recommending, and enforcing alternative NPN 6G system configuration options and alternative network resource distribution options to optimize the energy efficiency of the NPN while ensuring the 6G connectivity service meets the expectations of enterprises and subscribers.

For this PoC the Network Digital Twin applies advanced ML methods within the 3GPP PNI-NPN framework, focusing on the selection of algorithms to optimize NPNs, balancing computational efficiency and interpretability of results based on problem context and data structure.

TABLE 7. KEY CONCEPTS MAPPING VERSUS POCS

Key Concepts
NPN Digital Twin System
Relation of vertical KPIs with the network configuration
Inter(a)-slice reconfiguration and massive slicing
NEF instance for KPI data and configuration capabilities exposures of NPNs
AI/ML methods for reducing energy consumption
Energy efficiency as service criteria
ML optimization with an embedded analytics engine in a Digital Twin

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

2.1.2.1 PoC Implementation details

2.1.2.1.1 Implemented Architecture

Implemented PoC follows the 3GPP PNI-NPN architecture, as depicted in following figure:

On the one hand we have the enterprise customer (NPN), with a series of connectivity needs depending on the use case or use cases. For this we deploy in the enterprise only what needs to be in their facilities; this is Radio and User Plane. With this we achieve cost savings in equipment and very low latency times.

On the other hand, we have the service provider, who will offer us two fundamental elements:

- 5G Core, which will give us the control plane capabilities as well as slicing capabilities.
- A Digital-Twin of the private network, which will give us advanced network design, testing, analytics, and optimization capabilities.

Finally, we have the consumer or user of these capabilities, in this case for example the enterprise customer's technology team, which is the one that knows and controls the 5G “users” of the network, in this case for example, AGVs, Robots or Drones.

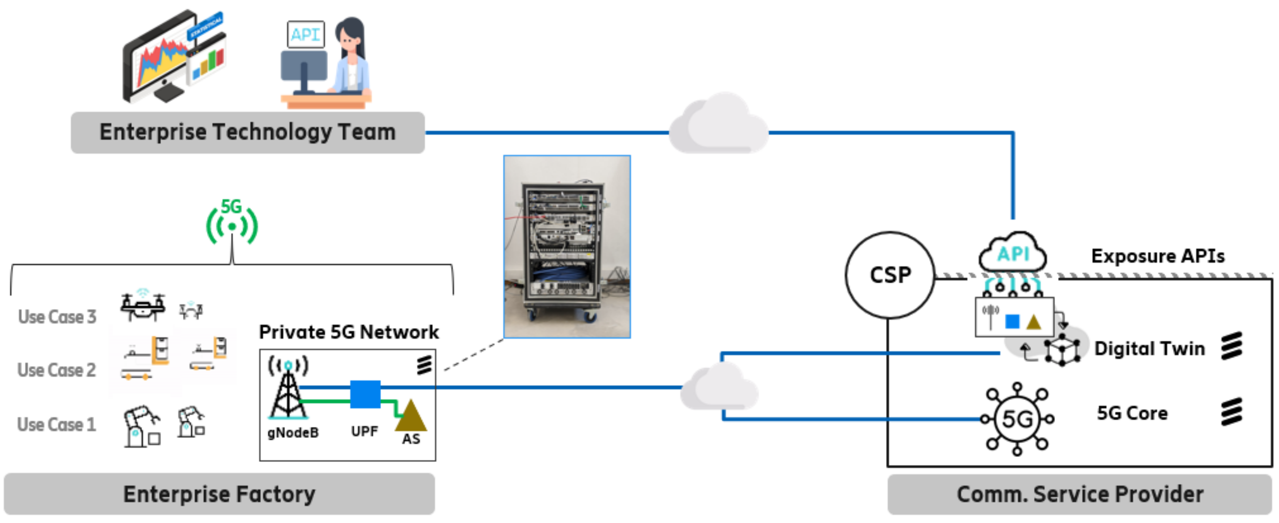


FIGURE 7. PNI-NPN ARCHITECTURE WITH ITS MAIN ACTORS

When it comes to the practical implementation of the PNI-NPN we can start with a geographical description of the setup deployed.

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

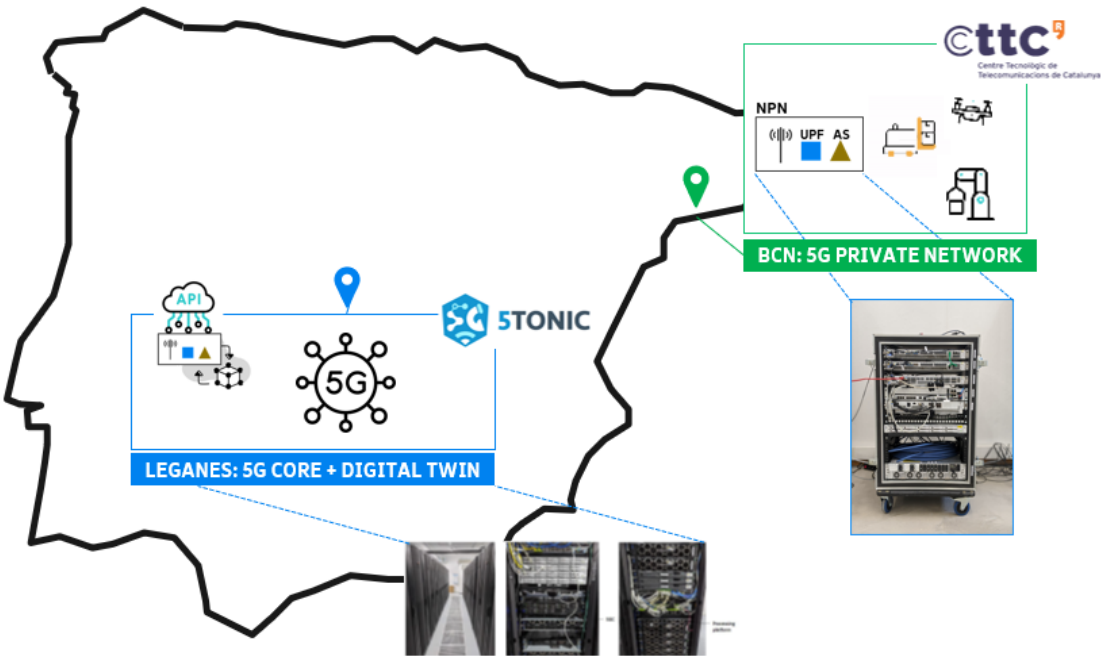


FIGURE 8. GEOGRAPHYCAL PNI-NPN SYSTEM COMPONENTS

- PNI part were deployed in 5TONIC lab at Leganes Madrid. Consisting of 5GC and Network Digital Twin Platform
- NPN part were delivered and deployed in CTTC Lab at Castelldefels, Barcelona. Consisting of a portable NPN system flight-rack cabinet hosting the following components:

1. 5G Radio DOT System (RDS): It provides 5G NR connectivity and is specially designed for indoor coverage.
2. 5G Indoor Radio Unit (IRU): The main purpose of the IRU is the transmission of signals providing an interface to the RDS through the Radio DOT Interface and supplying power to the RDS through this interface.
3. 5G Baseband: It provides switching, traffic management, timing, baseband processing, and radio interfacing.
4. IP Router: It is a high-capacity access router, designed to provide high-density 10G interfaces. It supports VPN services over IP/MPLS networks, service provider SDN, service exposure using NETCONF/YANG, extensive quality of service, and precise synchronization features.
5. GPS: a GPS receiver consists of a GNSS Active antenna and a GPS03. The GPS signal is needed to have a Time & Phase Synchronization required for 5G with TDD strategy.
6. Dedicated COTS Dell PowerEdge server for UPF: It is the physical server that hosts Kubernetes K3S distribution as the CaaS layer and UPF CNF will be deployed over this layer.

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

7. Dedicated COTS Dell PowerEdge server for hosting application function: It is the physical server that hosts application function software.
8. Power Supply Unit: It is used to provide power supply to baseband, IRU, and IP router.
9. Flight Rack: 19" rack flight case to host all physical NPN components explained above.

2.1.2.1.2 PoC actors and workflows

The optimization procedure focuses on the reduction of energy while guaranteeing the required KPIs tuning different parameters of the network and observing their impact on the energy and performance KPIs, exploring the tradeoffs between both.

A workflow representation of the PoC is depicted in following picture:

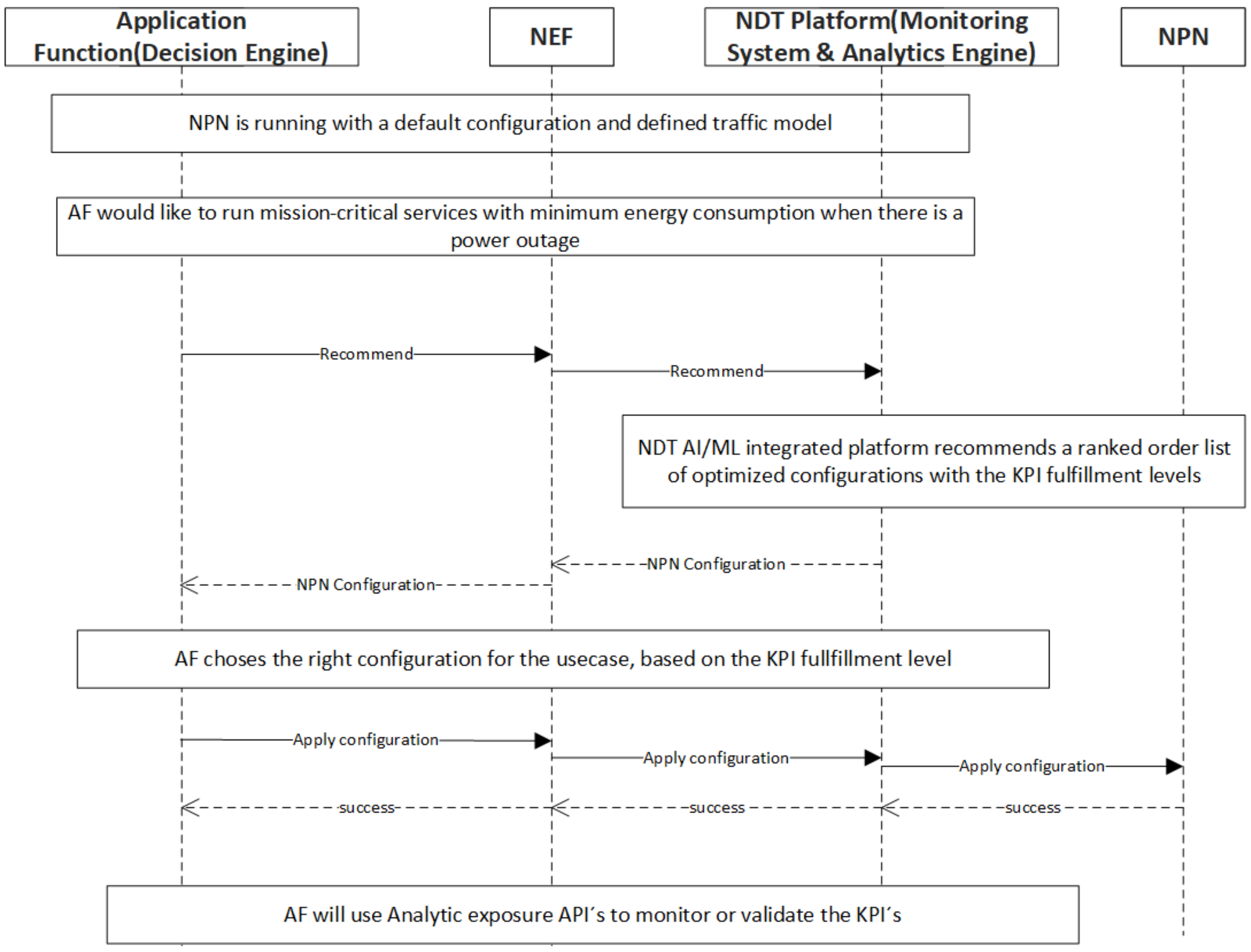

Application Function(Decision Engine)
NEF
NDT Platform(Monitoring System & Analytics Engine)
NPN
NPN is running with a default configuration and defined traffic model
AF would like to run mission-critical services with minimum energy consumption when there is a power outage
Recommend
Recommend
NDT AI/ML integrated platform recommends a ranked order list of optimized configurations with the KPI fulfillment levels
NPN Configuration
NPN Configuration
AF choses the right configuration for the usecase, based on the KPI fullfillment level
Apply configuration
Apply configuration
Apply configuration
success
success
success
AF will use Analytic exposure API´s to monitor or validate the KPI´s

FIGURE 9. OPTIMIZED NPN CONFIGURATION FOR ENERGY CONSUMPTION

Following actors intervene in the PoC workflow:

Analytical Engine:

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

The functionality of the Analytical Engine is primarily planned to be implemented in the NDT in this PoC. In the NDT implementation, the primary objective of the AE is to achieve energy consumption optimization of the PNI-NPN system.

NDT AE utilizes the data captured by MS and it employs an AI/ML model to forecast energy consumption patterns of the PNI-NPN system. This forecast is based on historical collected data, network configuration, traffic patterns, etc.

Details about AI/ML model to forecast energy consumption are detailed in chapter **2.1.2.2.1.2** Selected AI/ML model for Energy Consumption:

The output of the NDT AE is a ranked ordered list of PNI-NPN configurations to optimize the energy consumption of the system. This list of optimized configurations is conveyed to the Decision Engine of the AF with the support of NEF.

Decision Engine:

In this PoC, the AE component of NDT will generate a ranked ordered list of PNI-NPN configurations along with their corresponding achievable levels of energy optimization. **The responsibility of selecting the specific PNI-NPN configuration lies within the DE of the AF**. **AF's DE will make this decision, considering additional factors and requirements specific to the vertical use case**. The interaction and communication between the NDT and AF will be facilitated through the NEF.

Actuator:

In this PoC, once AF's DE selects one of the PNI-NPN configurations proposed by NDT via NEF, it triggers AF's Actuator to apply that configuration on the NPN system. Then AF ACT interacts with the specific NEF Service - Extension for RAN Configuration, to convey the request to the NDT. Once NDT receives the configuration change request from NEF, it applies the selected configuration on NPN System to implement energy optimization.

2.1.2.1.3 NPN Recommender API

Use case leverage on the so called "recommender" exposure API enabling users to acquire a set of recommended NPN configurations tailored to their input traffic model and specified KPIs.

These configurations are designed to meet the requested KPIs and traffic model requirements.

2.1.2.1.3.1 API inputs:

The endpoint is `POST /5tonic-exposure/v1/npn/<npnId>/recommend`. The request body is a JSON object containing the following input data:

1. **npnId:** Identifier of the NPN
2. **TrafficModel:**
   - protocol_stack: Protocol used in traffic model [UDP, TCP, ICMP]
   - bandwidth: target bandwidth required for TCP or UDP
   - traffic_direction: Direction of traffic (Downlink, uplink) required for TCP or UDP

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. **KPIs (**SLA parameters**)**
    - **MaxOWDMean**: The maximum allowable mean one-way delay in milliseconds for the NPN configuration with the specified traffic models.
    - **MaxPowerConsumptionMean**: The maximum allowable power consumption in Watts for the NPN configuration with the specified traffic models.
    - **MinThroughput**: The minimum throughput the configuration must achieve.
4. **optimizationType:** Recommender given results will be prioritized based on the selected criteria*
    - THROUGHPUT_OPTIMIZATION. The API output JSON will be ordered from highest max throughput to lowest max throughput
    - LATENCY_OPTIMIZATION. The API output JSON will be ordered from lowest owd prediction to highest owd prediction
    - **ENERGY_OPTIMIZATION**. The API output JSON will be ordered from lowest energy consumption to highest energy consumption *** (This criteria must be selected for UC1 PoC2)**
    - RESOURCE_OPTIMIZATION. The API output JSON will be ordered from lowest NPN cell bandwidth to highest.

2.1.2.1.3.2 API outputs:

NDT returns a list of NPN configurations fulfilling target KPIs (NpnConfigurationwithfulfillmentKPI);

- **NpnConfigurationwithfulfillmentKPI**:
    - **NpnConfiguration**: Suggested NPN Configuration
    - **KPIFulfillment**: predicted KPI values for the suggested NPN configuration based on the requested traffic model

The user gets a list of configurations that fulfill the traffic models and the request KPIs, ordered. Along with the suggested configurations there's a list of predicted KPIs for each of those NPN configurations.

The API output will be a JSON object containing a list of objects **NpnConfigurationwithfulfillmentKPI**. This object contains two objects:

- **NpnConfiguration**: it's the NPN system **suggested** configuration, following the API model.
- **KPIFulfillment**. Will contain the predicted KPIs for that configuration:
    - **MaxThroughput**: This is the maximum throughput the NPN configuration can achieve with the requested input traffic model in Mbps.
    - **PredictedPowerConsumption**: the predicted power consumption in Watts for the traffic model.
    - **PredictedOWDMean**: the predicted one-way delay in milliseconds for the traffic model.

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

TrafficModel: the trafficModel itself where the previous KPIs apply for the NpnConfiguration.The list of **NpnConfigurationwithfulfillmentKPI** is sorted according to the selected **optimizationType,** in our case (UC1 PoC2) **ENERGY_OPTIMIZATION**

2.1.2.1.4 NPN Configuration API

Once that list of NPN configurations fulfilling target KPIs is available, next step would be to apply a selected configuration, for that Configuration API allows to apply to the NPN system a given configuration while Configuration Retrieval API can be used to validate the applied configuration.

2.1.2.1.4.1 Input

The configuration to be applied takes into account the following parameters: Spectrum Assignment, Spectrum usage technique, Radio Unit type, gNB configurations, air conditions, KPIs and Traffic Models.
The endpoint for this API is `POST /5tonic-exposure/v1/npn/<npnId>/configure`. The request body is a JSON object containing the following input data:

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 6G BLUR 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 6G BLUR 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 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 cells 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 configurable values in the model are: 20, 40, 80 and 100 MHz.
- **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)).

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

- **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 1000mW to 10000mW in steps of 1000mW and for the AIR6488 we reduced the possibilities to 2400mW and 10000mW.
- **MIMO**: It 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. It is a configuration 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 into account the use of the maximum antennas also depends on the air conditions.
- **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 6G BLUR project single carrier is the one allowed.

2.1.2.1.4.2 Output

If configuration is successfully applied NDT returns a "`201 Created: A new resource was created successfully`".

2.1.2.1.5 NPN Configuration Retrieval API

In order to get information about the current configuration of the NPN system, NDT offers a NPN Configuration Retrieval API

2.1.2.1.5.1 Input

The endpoint for this API is `GET /5tonic-exposure/v1/npn/<npnId>/configuration`. The request body is a JSON object containing the following input data:

2.1.2.1.5.2 Output

The API output will be a JSON object containing an object NpnConfiguration similar to the one used as input in Configuration API

2.1.2.1.6 NPN Analytics API

The Analytics API is designed to gather diverse data from the Non-public Network (NPN) system. It provides users with the capability to obtain a wide range of aggregated Key Performance Indicators (KPIs), such as throughput, data volume, latency, energy consumption, active users, and link availability, among others. Users can define the period for data collection with high flexibility, allowing for aggregation over days, hours, or specific time frames within a given period.

2.1.2.1.6.1 Input

The endpoint for this API is `POST /analyticsexposure/{npnId}/fetch.`

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 request body is a JSON object containing the following three objects input data:

- **AnalyticsEvent** (enum):
  - UE_MOBILITY: the Exposure API will retrieve UE mobility analytics.
  - NPN_KPI: the Exposure API will retrieve analytics from the NPN system.
- **AnalyticsEventFilter**:
  - Start: start time to query data.
  - Stop: stop time to query data.
  - TimeSlotStart and TimeSlotStop in case these values are in the request, the analytics API will take the time slot defined TimeSlotStart and TimeSlotStop and apply it every day since Start to Stop time.
  - TemporalGranSize: the duration in seconds of the time window to query data.
  - Dnn.
  - Snssai.
  - CellId.
  - 5-tuple:
    - Protocol.
    - Ipv4AddrDest.
    - Ipv4AddrSrc
    - PortSrc
    - PortDest
- **TgtUe**. It's the id of the UE, it can be identified by:
  - GPSI
  - IMSI
  - IPV4 ADDRESS

In the AnalyticsEventFilter thetimeSlotStart and timeSlotStop are optional in case the user wants to obtain analytics from the same time range over different days.

2.1.2.1.6.2 Output

The output will be an object containing different data depending on the Analytics Event in the input data.

2.1.2.1.6.2.1 NPN KPI

For NPN_KPI, the response will be an object TrafficKpiData which contains:

- **TrafficKpiRanData**: This section provides detailed metrics related to the radio access network (RAN):
  - **RanDownlinkThroughput**: Represents the average and variability of data throughput in the downlink. It provides also **potencial anomalies**.
  - **RanUplinkThroughput**: Represents the average and variability of data throughput in the uplink. It provides also **potencial anomalies**.
  - **ActiveUsersDL**: Indicates the range of active users in the downlink.
  - **ActiveUsersUL**: Indicates the range of active users in the uplink.
  - **HarqDlAck**: Contains acknowledgment counts for different modulation schemes in the downlink, such as 16QAM, 64QAM, 256QAM, and QPSK.
  - **HarqUlAck**: Contains acknowledgment counts for different modulation schemes in the uplink.

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

    - **PdschTableMcsDistr**: Shows the distribution of modulation and coding schemes (MCS) for PDSCH across three tables, indicating how often each MCS is used.
    - **PuschTableMcsDistr**: Shows the distribution of MCS for PUSCH across three tables, indicating how often each MCS is used.
    - **UtilizationDistributionDL**: Provides the distribution of resource utilization in the downlink.
    - **UtilizationDistributionUL**: Provides the distribution of resource utilization in the uplink.
    - **RanEnergyConsumption**: Represents the total energy consumed by the RAN.
    - **RanDataVolume**: Indicates the total volume of data handled by the RAN.
    - **RanEnergyConsumptionPerByte**: Shows the energy consumption per byte of data transmitted.
- **TrafficKpiCoreData**: This section provides metrics related to the core network:
    - **UpfDataVolume**: Represents the total volume of data processed by the user plane function (UPF).
- **TrafficKpiNpnData**: This section includes metrics for non-public networks (NPN):
    - **Latency**: Represents the average and variability of network latency.
    - **LinkLocalAvailability**: Indicates the availability of local network links.
    - **LinkCentralAvailability**: Indicates the availability of central network links.

2.1.2.1.7 Vertical service – extension of the operational time of critical sites during power outages

The PoC additionally demonstrates how the implemented APIs in the NPN (Recommender, Configuration, Configuration Retrieval and Analytics) could be exploited by a vertical service, namely, how could these APIs be used by PPDR network operators to predict and extend the expected operational time of critical sites during grid failures. Natural disasters often cause grid failures, leaving critical sites used by PPDR organizations dependent on finite power sources such as batteries or generators. Ensuring the sustained operation of these sites, and accurately being able to predict their remaining operation time before the backup power sources exhaust, is essential for effective emergency response. Therefore, by using data collected from the different power sources together with the energy consumption predictions and configuration suggestions provided by the NDT, the implemented AF is capable of (i) predicting how much time will the power source take to run out of power and (ii) modify the network configuration, from the list of suggested ones, with the goal of consuming the least energy as possible and therefore extending the life of the site as much as possible without damaging performance. Additionally, a vertical-specific MS and AE, which collect and analyze metrics of interest for PPDR network operators has been implemented. This system, together with the generic KPIs provided by the NDT, is key in supervising the performance of the network, ensuring it is not impacted negatively by newly applied configuration changes.

Application KPI collection

The implemented system which, as already indicated, collects key QoE KPIs related to the PPDR communication services, tracking the performance of the network, consists of the following sub-components:

- **Collection of client-side radio information**: apart from the radio KPIs provided by the NDT based on metrics collected by the network infrastructure, this module retrieves and sends to the AE at the AF NR client-side information at the UE. This collection system has been implemented via Android's Telephony interface and via the function cellInfo() , collecting data such as the CSI CQI, SINR or TA.

- **Collection of audio data**: in order to analyze the quality of transmitted audio via PPTX or other communication forms, the collection of audio samples from real calls has been implemented. Sampling starts based on the monitoring of SIP messages within Android phones where the 6GDAWN MS agent is installed, which detects when a new call is being established. Then, audio is collected via Android's CallRecord API (for supported devices, since this interface is not supported in all Android models) and send to the centralized AF AE (i.e. the QOE estimator) for processing.
- **QoE estimator**: this AE module is capable of producing QoE estimators for PPDR voice communications via the analysis of the collected audio sample. These QoE estimations are computed in the form of estimated Mean Opinion Scores (MOS) using an ML model. Unlike referenced-based MOS estimators, where the original (reference) audio is required, the implemented model can compute estimations using only the degraded signal.

More concretely, for the QoE estimator, a non-intrusive Mean Opinion Score (MOS) estimator inspired by state-of-the-art (SoA) techniques was developed, designed to assess audio quality without the need for a reference signal. The approach begins by processing audio inputs into spectrograms using the Fast Fourier Transform (FFT), specifically generating Mel-spectrogram segments that capture the frequency distribution of the audio over time. Each Mel-spectrogram segment is then passed through a Convolutional Neural Network (CNN) to produce feature vectors that represent the underlying characteristics of the audio. To capture time-dependent relationships across these feature vectors, we employ a Self-Attention (SA) network, which is adept at identifying patterns and dependencies over sequential data. The final quality estimation is generated through a pooling mechanism that combines information from all feature vectors to produce a robust MOS prediction.

The estimator was trained using public datasets containing audio samples labeled with MOS values, primarily derived from crowdsourcing methods where human evaluators rated the audio quality. These datasets provide a rich ground truth for training and validating the model. The training process involved optimizing the model to minimize the difference between predicted MOS and ground-truth MOS values, ensuring the estimator could generalize well to unseen data. Once trained, the model was evaluated on a separate test set, where it successfully predicted MOS scores, demonstrating its capability to accurately estimate audio quality in a non-intrusive manner. This method combines efficiency with the ability to capture complex audio characteristics, making it suitable for real-world applications such as streaming, VoIP, or audio quality monitoring systems.

Distributed KPI collection

As already seen, the implemented QoE estimator leverages a deep learning pipeline that produces MOS predictions using a combination of Mel-spectrogram extraction, CNNs, and SA mechanisms.

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

While this approach ensures a high degree of accuracy in assessing audio quality, it also introduces significant computational complexity and high GPU/memory usage, especially when handling multiple audio streams concurrently. Furthermore, the SA mechanism scales quadratically with sequence length, making it a major computational bottleneck. The sequential execution of these stages requires substantial processing power, particularly for real-time or large-scale deployment scenarios where online processing of audio streams is not viable, reducing the capability of rapidly responding to QoE degradation.

Given these challenges, to enable real-time processing and scalability capable of fulfilling acceptable latency and resource constraints, a distributed architecture is required. By distributing the workload across multiple computing nodes—whether through data parallelism, model partitioning, or edge-cloud offloading—the system can process multiple audio streams in parallel, reduce latency, and improve overall efficiency.

As a first measure to reduce the amount of used bandwidth, distribute the computation load and improve the privacy of users, the devised distributed architecture supports the ability to run the less computationally intensive tasks (i.e. the Mel-spectrograms and CNN-based feature extraction) in the Edge devices, sending the extracted feature vectors to the AE for self-attention processing and MOS prediction. This way, bandwidth usage is reduced by sending compact feature vectors instead of raw audio, further improving in the privacy of users (where audio is only processed locally). Additionally, the designed architecture proposes the parallelization of the compute-intensive tasks via the distribution in multiple AE instances of (i) the self-attention tasks and (ii) the pooling mechanisms, reducing per-node memory usage and improving throughput by overlapping computation. Finally, a parallel fast feedback model could be implemented, where prediction accuracy could be traded in favor of better response latency. In order to achieve this, Long Short-Term Memory (LSTM) networks can be used as a replacement for self-attention in this auxiliary model, offering a more efficient alternative for capturing temporal dependencies while maintaining real-time performance.

Power source & outage simulator

Due to the absence of a secondary power source, such as a battery, in the 5G site testbed, along with the lack of a switching system capable of detecting grid power outages and transitioning to a secondary source, as well as the absence of a battery management system to monitor critical metrics like the state of charge (SoC), a simulator has been implemented to address these limitations. The simulator replicates the behavior of a secondary power source, the switching mechanism, and the battery management system, enabling the emulation of realistic power failure scenarios.

Implementing a simulator offers several advantages. First, it allows the proof of concept (PoC) to scale more effectively to complex scenarios involving multiple sites, as the simulated environment can easily accommodate diverse configurations and conditions without additional physical hardware. Second, it accelerates the development and testing process by enabling faster iterations. Developers can simulate different outage and recovery scenarios, refine algorithms, and validate system

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

functionality more efficiently than with physical systems, which may require extensive setup time or incur higher costs. Additionally, the simulator provides flexibility for testing edge cases and extreme conditions that might be challenging or unsafe to replicate with actual hardware, thereby improving the robustness and reliability of the final system. Overall, the simulator is a cost-effective and efficient tool for advancing the testbed's capabilities and ensuring the scalability of the solution.

Extension of operational time & control dashboard

Therefore, using this simulator, a NPN network operator admin dashboard was developed, enabling (i) the real-time visualization of all collected KPIs (both those produced by the NDT's Analytics API, those associated to the power sources and those collected from the vertical services), (ii) the alerting of power outages, (iii) the request and visualization of different suggested configurations to optimize energy consumption in selected sites based on the selected traffic model and (iv) the selection and enforcement of the most appropriate configuration.

2.1.2.2 E UC1 PoC2 KPIs Evaluation and Results

The Proof of Concept (PoC) for Energy optimization system in NPNs using a Digital Twin successfully demonstrated the capabilities of the NPN system by reproducing a vertical use case that exploits the data and analysis models provided by the Digital Twin System to optimize the energy efficiency at critical sites, achieved by trading off some of the performance characteristics delivered by the network.

For example, data rates and expected delay could be traded off in order to achieve lower energy consumption by tuning network parameters (such as forcing the usage of lower modulation and coding schemes, enforcing the pre-emption of non-critical users, dynamically modifying the assigned NR Resource Blocks or adjusting the antenna tilts and transmission powers). **This ability would be especially useful in (i) certain high-impact sites, where it could be desirable to intentionally reduce the delivered capacity in order to increase their autonomy in the event of power outages or (ii) to reduce Energy consumption in areas of low demand, a quite common case in NPN networks**

Data Exploration and Analysis:

The collected data originates from NPN Characterization, various experiments, and real-world use case experiments conducted on the NDT, each lasting a different duration. These experiments were facilitated by APIs (Refer to the Exposure APIs shared in gitlab repository) specifically developed for the automatic execution and collection of the necessary data through the configuration and execution of a battery of experiments. This automated approach ensured the creation of a comprehensive database, which serves as the foundation for all subsequent analysis and predictive solution development using machine learning.
Each experiment involved the execution of varied parameter configurations for the NDT settings and a traffic model, ensuring a diverse and robust dataset suitable for training and evaluating predictive models.

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 raw dataset comprises 40 unique parameter configurations, with multiple traffic models executed for each, resulting in diverse performance samples. Key performance metrics include energy consumption, one-way delay (OWD), and traffic received. The dataset, free of missing values, includes both categorical and numerical variables, making it well-suited for exploratory analysis and predictive modeling.
Data preprocessing involved cleaning, variable selection, and transformations like One-Hot Encoding for categorical variables and scaling for numerical features to ensure equitable model training. A pipeline integrates these preprocessing stages with model training, enhancing consistency and preventing data leakage.
Exploratory Data Analysis (EDA) revealed important correlations: traffic received is strongly correlated with traffic injected, while energy consumption shows moderate correlations with these traffic metrics. Delays, however, are influenced by categorical variables like direction and protocol, suggesting the need for complex modeling techniques. Descriptive statistics highlighted distribution patterns and the impact of categorical variables on target metrics, emphasizing their significance in model development.
Raw Data:

The table below contains the collected raw data samples for different types of use case executions at the NPN system.

TABLE 8. RAW DATA SAMPLE

id	start_time	stop_time	duration	configuration_cell_name	configuration_bandwidth	configuration_power	configuration_antenna	configuration_tdd_pattern	energy	owd	traffic_model	traffic_received	traffic_sent
006321cc-633b-4df9-b73d-8c7e5e84684f	2024-06-26T16:51:31.525361523Z	2024-06-26T16:52:32.757544522Z	60	PELICAN_5G_DOT-1	40	1004	RD4479B78L	DDDSU(10:2:2)	0.74,0.47,0.63,0.8,0.37,0.73,0.47,0.63,0.03,0.13	296181908, 296301779, 296539951, 295315519, 295729536, 294187097, 296468752, 295639050, 294707463, 296092280	uplink_UDP_120M	27.940169564986405,27.95598615909394,27.9300813768452,28.52385829713195,28.007316952472877,27.931272916411377,27.938538186285808,27.945326234199385,27.932300358656203,27.932300358656203	123.43762172647875,123.51269851864329,123.45158621082237,123.49115464246913,123.46752685825786,123.45280680109232,123.48809042153728,123.473767299753,123.49366738604661,123.47132656794848
00f9d464-071d-46f2-b98f-99919b637ca5	2024-06-26T15:36:54.487597718Z	2024-06-26T15:37:55.983240647Z	60	PELICAN_5G_DOT-1	20	1004	RD4479B78L	DDDSUDDSUU(10:2:2)	0.37,0.73,0.47,0.63,0.03,0.13,0.73,0.47,0.63,0.8	519666315, 539733649, 497439809, 519619573, 516947830, 509527009, 501709014, 486166412, 484560798, 482452863	uplink_UDP_150M	15.905741121522565,15.174072518715834,15.977648047088252,16.760465176369998,16.25012181328539,15.479174125405743,16.545358722659824,16.708693434993428,16.78153980334216,17.04265012803471	154.35542387062077,154.32637335651947,154.34422969788184,154.32247496330123,154.36649507484617,154.3501683153286,154.31300843524377,154.34426726158404,154.3618735068018,154.32894079485223

As we can see in the example of the attached table, there are 14 columns, the contents of which are detailed below:

- id: Unique identifier of the experiment for each record.
- start_time: Start timestamp of the measurement of the experiment.
- stop_time: Stop timestamp of the measurement of the experiment.
- duration: Duration of the measurement of the experiment (in seconds).
- configuration_cell_name: Name of the cell configuration.
- configuration_bandwidth: Bandwidth configuration.
- configuration_power: Power configuration.
- configuration_antenna: Antenna type.
- configuration_tdd_pattern: TDD pattern configuration.

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

- energy: Energy measurement values for the duration of the experiment, presented as a series of comma-separated numbers, with each number representing a measurement taken every 6 seconds.
- owd: One-way delay (OWD) values as a series of comma-separated numbers. Each number is the measurement gotten every 6 seconds.
- traffic_model: Type of traffic model.
- traffic_received: Received traffic values as a series of comma-separated numbers. Each number is the measurement gotten every 6 seconds.
- energy_total: Total energy consumption measured for the duration of the experiment.

Predictive Model Development

Data Splitting

The objective for data splitting is to ensure robust model training and evaluation by dividing the dataset into training and testing subsets.

Steps Taken:

For each target variable (traffic_received_mean, energy_total, owd_mean):

1. **Dataset Division**:
    - The dataset was split into:
        - **80% Training Data**: Used for training the model and optimizing its parameters.
        - **20% Test Data**: Reserved for evaluating the model's performance on unseen data.
2. **Stratification of Target Variables**:
    - The splits ensured a good balance of the target variable values across both training and testing subsets.

Justifications for Each Action:

Why Split the Data?

- Splitting the dataset is a critical step in model development to prevent overfitting and ensure that the model generalizes well to new, unseen data.
- **Training Data**: By using 80% of the data, the model has sufficient information to learn patterns, relationships, and interactions within the dataset.
- **Test Data**: The remaining 20% provides an unbiased evaluation of the model's performance, simulating how it would perform on real-world data.

Why 80/20 Split?

- This ratio strikes a balance between providing enough data for training while reserving a substantial portion for reliable testing.
- It is a commonly accepted standard in machine learning practice, ensuring neither under-training nor insufficient testing.

Why Ensure Balance in Splits?

- **Target Variable Distribution**:

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 numerical targets like traffic_received_mean, energy_total, and owd_mean, ensuring a balanced distribution in training and testing sets prevents skewed predictions.
  - This balance avoids bias introduced by underrepresented scenarios in either split.

- **Dataset Integrity**:
  - The dataset includes diverse configurations (configuration_bandwidth, direction, protocol, configuration_tdd_pattern). Stratification ensures these features are proportionately represented in both subsets, maintaining the integrity of their impact on target variables.

Implementation Details:

- The dataset was split using a random seed for reproducibility, ensuring consistent results across multiple runs.
- Libraries such as sci-kit-learn were used for splitting, leveraging built-in functionality to maintain balance in the target variable distributions.

Model Creation

Models Tested

The following predictive models were tested to evaluate their effectiveness in predicting the target variables (traffic_received_mean, energy_total, owd_mean). Each model was chosen for its unique advantages and its potential to leverage the characteristics of the dataset.

Tree-Based Ensemble Models

1. **RandomForestRegressor:**
   - **Strengths**: Captures non-linear relationships, robust to overfitting when properly tuned, and provides feature importance.
   - **Why Suitable**: Handles both categorical (e.g., direction, protocol) and numerical features (e.g., configuration_bandwidth, traffic_injected) effectively. Its ensemble nature makes it resilient to noise in the dataset.
   - **Use Case**: Ideal for scenarios requiring interpretability of feature importance and moderate computational efficiency.
2. **GradientBoostingRegressor:**
   - **Strengths**: Performs well on smaller datasets, focuses on reducing errors iteratively, and captures complex patterns.
   - **Why Suitable**: Effective for modeling non-linear interactions, such as the relationship between configuration_bandwidth and traffic_received_mean.
   - **Use Case**: Suitable for datasets with moderately high dimensions and when high predictive accuracy is needed.
3. **XGBoost:**
   - **Strengths**: Highly optimized for speed and accuracy, includes regularization to prevent overfitting.

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

- **Why Suitable**: Can handle missing data and complex interdependencies between features like configuration_tdd_pattern and target variables.
- **Use Case**: Best for achieving high accuracy in datasets with mixed data types and higher computational budgets.

4. **CatBoostRegressor:**
   - **Strengths**: Handles categorical features directly, reducing preprocessing effort, and is robust to overfitting.
   - **Why Suitable**: The dataset includes significant categorical variables (direction, protocol), making CatBoost highly efficient and effective.
   - **Use Case**: Ideal when categorical variables significantly impact target predictions.
5. **LightGBM:**
   - **Strengths**: Highly efficient with large datasets and features, handles categorical data natively.
   - **Why Suitable**: Handles the mix of categorical and numerical data effectively, scales well for high-dimensional datasets.
   - **Use Case**: Recommended for large datasets with multiple categorical variables.
6. **ExtraTreesRegressor:**
   - **Strengths**: Robust to outliers, captures non-linear relationships, and provides feature importance insights.
   - **Why Suitable**: Handles high-dimensional data efficiently and performs well on datasets with imbalanced features.
   - **Use Case**: Selected for training as it combines robustness with computational efficiency.
7. **AdaBoostRegressor:**
   - **Strengths**: Combines weak learners iteratively to reduce bias and variance.
   - **Why Suitable**: Captures patterns overlooked by simpler models, suitable for moderate dataset sizes.
   - **Use Case**: Effective when dealing with challenging target variable distributions.

Linear Models

1. **LinearRegression:**
   - **Strengths**: Simple and interpretable, ideal for linear relationships.
   - **Why Suitable**: Serves as a baseline model to evaluate other complex models.
   - **Use Case**: Effective for identifying direct correlations in features like configuration_bandwidth and traffic_received_mean.
2. **Ridge, Lasso, and ElasticNet:**
   - **Strengths**: Introduce regularization to address multicollinearity and enhance generalization.
   - **Why Suitable**: Useful for feature selection and handling collinear features like traffic_injected and traffic_received_mean.

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

- **Use Case**: Ridge for simple regularization, Lasso for feature selection, ElasticNet for combining the two.

Support Vector Machines

1. **Support Vector Regressor (SVR)**:
   - **Strengths**: Captures non-linear relationships using kernel functions.
   - **Why Suitable**: Useful for predicting owd_mean, which may depend on complex interactions.
   - **Use Case**: Best for smaller datasets with non-linear patterns.

Instance-Based Models

1. **KNeighborsRegressor**:
   - **Strengths**: Makes predictions based on proximity to neighbors, requires minimal assumptions.
   - **Why Suitable**: Effective for local patterns, such as the effect of similar configuration_tdd_pattern values.
   - **Use Case**: Useful for small datasets or when interpretability of local decisions is needed.

Other Robust Models

1. **HuberRegressor:**
   - **Strengths**: Robust to outliers, combines linear regression with resistance to extreme deviations.
   - **Why Suitable**: Ideal for owd_mean, which includes significant outliers.
   - **Use Case**: Effective for targets with heavy-tailed distributions.
2. **BaggingRegressor (with Decision Trees):**
   - **Strengths**: Reduces variance by aggregating predictions of multiple decision trees.
   - **Why Suitable**: Provides stability and robustness in datasets with mixed feature types.
   - **Use Case**: Useful for datasets with high variance in target variables.

Model Training

For each target variable (traffic_received_mean, energy_total, owd_mean), **15 models** were trained, leveraging their unique strengths to evaluate their predictive capabilities. The training process was carefully designed to ensure fair comparison and optimal performance for each model. The models trained included a mix of ensemble methods, linear regressors, and robust algorithms, as previously detailed.

1. **Data Preparation**

The training process began with preprocessing steps outlined earlier:

- **Feature Encoding**: Categorical features (direction, protocol, configuration_tdd_pattern) were encoded using One-Hot Encoding.

- **Scaling**: Numerical features (configuration_bandwidth, traffic_injected) were standardized using *"StandardScaler"* to ensure uniform input across all models.
- **Data Splitting**: The dataset was split into an 80/20 ratio for training and testing, ensuring stratified splits for target variable distributions.

2. **Model Training Process**

Each of the **15 models** was trained for every target variable using the following procedure:

1. **Baseline Training**:
   - All models were initially trained using default hyperparameters to establish a baseline performance.
   - Models evaluated included tree-based methods (e.g., RandomForest, GradientBoosting, CatBoost), linear regressors (e.g., Ridge, Lasso), and others (e.g., SVR, Huber).

2. **Hyperparameter Optimization**:
   - To enhance performance, hyperparameters were tuned using *GridSearchCV* or equivalent tuning methods, optimizing parameters such as:
     - **Tree-Based Models**: Number of estimators, maximum depth, learning rate.
     - **Linear Models**: Regularization strength (alpha).
     - **SVR**: Kernel type and regularization parameters.
   - This process minimized overfitting while ensuring robust performance on test data.

3. **Cross-Validation**:
   - 5-fold cross-validation was used during training to evaluate model generalizability and prevent overfitting.

4. **Iterative Training**:
   - Performance metrics were monitored during training to identify underperforming models, refine hyperparameters, and iteratively improve.

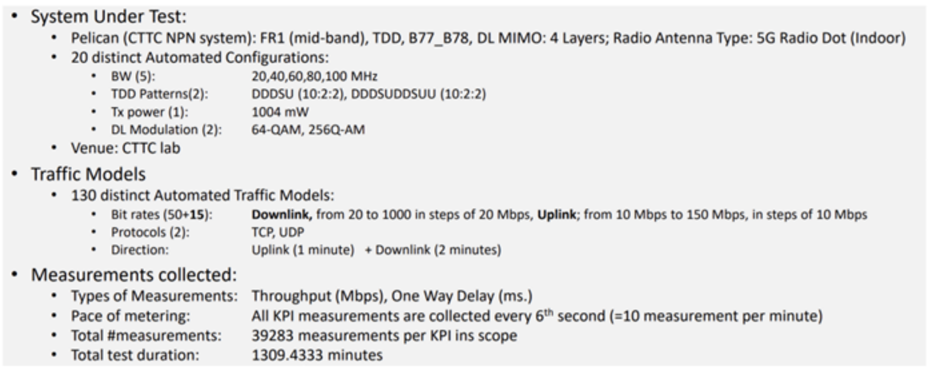
5. **Evaluation of Training Data**:

Training accuracy and metrics were computed to identify potential underfitting or overfitting before moving to test data evaluation.

2.1.2.2.1 Model Evaluation and recommender API performance

To generate the required data to perform the full characterization of the NPN system, an exhaustive measurement campaign is required. Indeed, in this case, multiple measurement campaigns have been carried out when the NPN was in preparation phase at 5TONIC lab (Q2-2024) and then, a second battery when deployed at CTTC (Q3-2024) premises to ensure the NPN system is operating with high efficiency. The generated experiment datasets are automatically collected and fed to ML models to update the training and prediction datasets to improve the AI/ML platform intelligence and accuracy of the prediction and recommendation capabilities.

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



- System Under Test:
  - Pelican (CTTC NPN system): FR1 (mid-band), TDD, B77_B78, DL MIMO: 4 Layers; Radio Antenna Type: 5G Radio Dot (Indoor)
  - 20 distinct Automated Configurations:
    - BW (5): 20,40,60,80,100 MHz
    - TDD Patterns(2): DDDSU (10:2:2), DDDSUDDSUU (10:2:2)
    - Tx power (1): 1004 mW
    - DL Modulation (2): 64-QAM, 256Q-AM
  - Venue: CTTC lab
- Traffic Models
  - 130 distinct Automated Traffic Models:
    - Bit rates (50+**15**): **Downlink,** from 20 to 1000 in steps of 20 Mbps, **Uplink**; from 10 Mbps to 150 Mbps, in steps of 10 Mbps
    - Protocols (2): TCP, UDP
    - Direction: Uplink (1 minute) + Downlink (2 minutes)
- Measurements collected:
  - Types of Measurements: Throughput (Mbps), One Way Delay (ms.)
  - Pace of metering: All KPI measurements are collected every 6th second (=10 measurement per minute)
  - Total #measurements: 39283 measurements per KPI ins scope
  - Total test duration: 1309.4333 minutes

FIGURE 10. DESCRIPTION OF A MEASUREMENT CAMPAIGN DRIVEN IN CUSTOMER PREMISES TO CHARACTERIZE THE NPN SYSTEM AND FEED AI/ML MODELS

The information obtained is crucial to train models that allow very reliable RAN configuration recommendation services for some target KPIs, or, complementarily, to estimate/predict the network performance for a given RAN configuration. This is a very powerful tool to focus on the specification of environments and pilot tests (reliability), as well as to save multiple unnecessary tests in the validation phase (efficiency and reduction of pilot lead times).

2.1.2.2.1.1 Model Evaluation

Models were evaluated on the test set using the following metrics:

- **Mean Absolute Error (MAE)**: Reflects average prediction error magnitude, interpretable in real-world units.
- **Mean Squared Error (MSE)**: Penalizes large errors, offering a comprehensive performance measure.
- **Root Mean Squared Error (RMSE)**: A direct interpretation of MSE in the same unit as the target variable.
- **R² Score**: Indicates the proportion of variance explained by the model, with higher values reflecting better fit.

Performance Results for Energy consumption

Following is a summary of the performance of all evaluated models for the variable *energy_total*.

TABLE 9. CONSUMED ENERGY [WATTS PER HOUR]

Model	**MSE**	**RMSE**	**MAE**	**R² Score**
RandomForest	2917.96	54.02	38.86	0.5253
GradientBoosting	2279.86	47.75	36.67	0.6291
XGBoost	2813.53	53.04	38.43	0.5423
CatBoost	2171.94	46.60	35.20	0.6467
LightGBM	2219.00	47.11	35.57	0.6390
AdaBoost	2596.94	50.96	42.72	0.5775

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

Model	MSE	RMSE	MAE	R^2 Score
ExtraTrees	3330.27	57.71	40.88	0.4583
KNeighbors	2665.66	51.63	37.75	0.5664
LinearRegression	3326.79	57.68	45.28	0.4588
Ridge	3314.11	57.57	45.12	0.4609
Lasso	3317.70	57.60	45.06	0.4603
ElasticNet	3818.01	61.79	51.04	0.3789
SVR	3671.49	60.59	45.01	0.4027
Huber	3561.76	59.68	43.76	0.4206
Bagging	3113.51	55.80	40.28	0.4935

Key Observations

After evaluating all the models across the target variable *energy_total* the following key observations were made:

General Observations

1. **Ensemble Models Outperformed Others**:
   - Tree-based ensemble models (e.g., RandomForest, ExtraTrees, GradientBoosting, CatBoost, and LightGBM) consistently outperformed simpler linear models and distance-based methods like KNeighbors and SVR.
   - These models were better at capturing non-linear relationships and interactions between features, especially for energy_total.
2. **Impact of Outliers**:
   - Robust models like HuberRegressor demonstrated resilience to outliers.
3. **Feature Dependencies**:
   - The categorical features (direction, protocol, configuration_tdd_pattern) played a significant role in predicting target variables, as shown in ensemble-based feature importance scores.
   - Numerical features such as configuration_bandwidth was significant for energy_total.
4. **Performance Trade-offs**:
   - While simpler models like LinearRegression and Ridge were computationally efficient, their inability to capture non-linear relationships limited their accuracy.
   - Models like SVR and AdaBoost struggled with large datasets, leading to poor performance and computational inefficiency.

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 behaviors for Energy Consumption:

- **ExtraTrees** delivered outstanding performance, capturing the gradual energy consumption increase in the pre-saturation region and stabilization near the threshold. Its predictions closely aligned with the real data across all regions.
- **RandomForest** performed very similarly to ExtraTrees but showed slight discrepancies at points of high energy consumption.
- **XGBoost**, while generally accurate, slightly overestimated energy consumption in certain areas, making it less reliable compared to ExtraTrees.
- **Bagging** showed significant dispersion near the saturation threshold, affecting its ability to predict stabilization accurately.
- **KNeighbors** had the worst performance, with scattered and inconsistent predictions.

2.1.2.2.1.2 Selected AI/ML model for Energy Consumption

After evaluating model behavior for Energy consumption, the **ExtraTrees** model was selected as the best option for the following reasons:

1. **High Precision:** ExtraTrees showed exceptional alignment with real data across all variables and regions, both before and after the saturation threshold.
2. **Consistency:** Unlike other models, ExtraTrees provided uniform and reliable predictions for all observed behaviors in each metric.
3. **Versatility:** This model effectively captured both linear patterns (in the case of traffic received) and nonlinear or abrupt changes (in the case of OWD and total energy).
4. **Computational Efficiency:** ExtraTrees offers an efficient implementation that combines fast training (approximately less than 5 seconds on a laptop with an x86 processor, 32 GB RAM, and Intel Iris Xe Graphics GPU with 16 GB of memory) with high predictive power.

In summary, **ExtraTrees** is the ideal choice to independently model each of the key variables in this context, ensuring an optimal balance between accuracy, robustness, and simplicity.

2.1.2.2.2 Real-world Use Case application

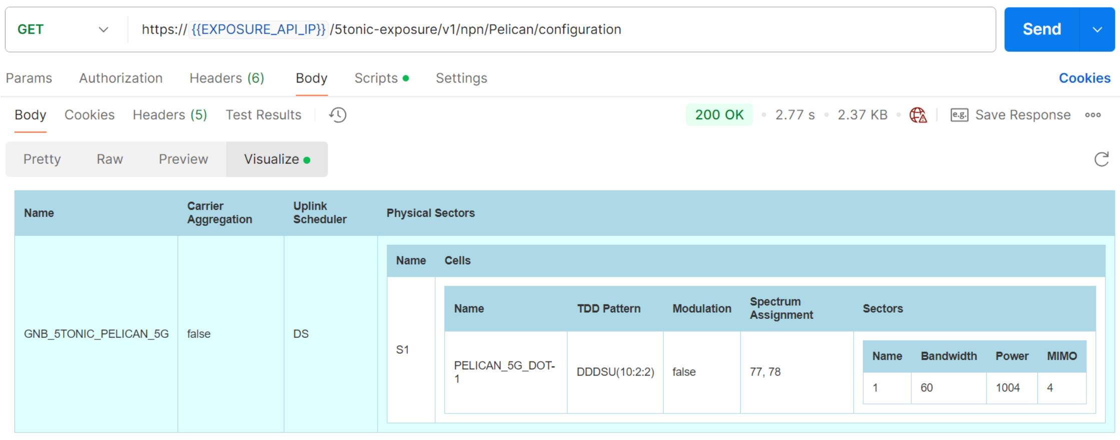
The real-world Use Case to be tested involves 3 main elements:

- NPN configuration
- Traffic models
- KPIs

2.1.2.2.2.1 NPN configuration

The starting point of the PoC is a configuration based on 60MHz bandwidth. To make sure what the current NPN configuration is we can make use of the Configuration Retrieval API of Pelican NDT

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



GET https:// {{EXPOSURE_API_IP}} /5tonic-exposure/v1/npn/Pelican/configuration

Name	Carrier Aggregation	Uplink Scheduler	Physical Sectors
GNB_5TONIC_PELICAN_5G	false	DS	

Name	Cells
S1	

Name	TDD Pattern	Modulation	Spectrum Assignment	Sectors
PELICAN_5G_DOT-1	DDDSU(10:2:2)	false	77, 78	

Name	Bandwidth	Power	MIMO
1	60	1004	4

FIGURE 11. CONFIGURATION RETRIEVAL EXAMPLE

2.1.2.2.2.2 Traffic models

We have considered 3 traffic models corresponding to 3 different types of Mission Critical traffic scenarios. All these traffic on the same coverage area, and the following characteristics:

Traffic Case	Purpose	KPIs	Duration
Traffic-A	Video conf	UL Tput UDP: 2 Mbps DL Tput UDP: 2 Mbps	Each 2 mins
Traffic-B	Video Broadcasting	UL Tput UDP: 5 Mbps DL Tput UDP: 5 Mbps	Each 15 mins for 3 mins
Traffic-C	Audio Calls	UL Tput UDP: 1 Mbps DL Tput UDP: 1 Mbps	Each 2 mins
		Common KPIs: Latency (OWD): 20ms Service Availability of 99.99%	

Here below a capture of these traffic models we have injected in Pelican NPN (UE probe capture)

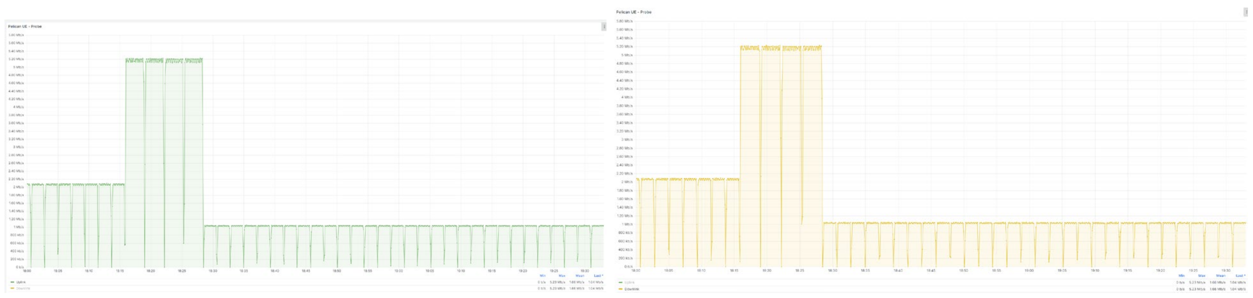


FIGURE 12. UPLINK AND DOWNLINK TRAFFIC OF THE REAL-WORLD USE-CASE TESTBED

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

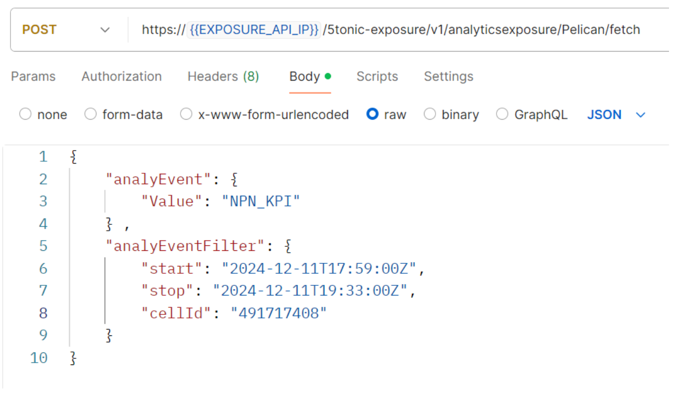
2.1.2.2.2.3 KPIs with current configuration

The most important KPIs on the 5G Network to successfully satisfy the Use Case demands are:

- Min Expected Tput of 2Mbps (video conf), 1Mbps (audio calls), 5Mbps (video broadcast)
- Latency (OWD): 20ms max
- Min Service Availability of 99.99%
- **Energy Consumption: Optimal configuration in terms of Energy consumption**

However, we should keep in mind that this PoC, which is Energy Consumption "centric" it is in fact the Energy Consumption KPI the one we should play with in situations where it could be desirable to intentionally reduce the delivered capacity in order to increase their autonomy in the event of f.e. power outages or to reduce Energy consumption in areas of low demand.

By using the Analytics API we can get access to real data, in this case we make use of the scheduler to get the data from the time slot when the traffic for the UC was injected in Pelican NPN (UE probe capture)

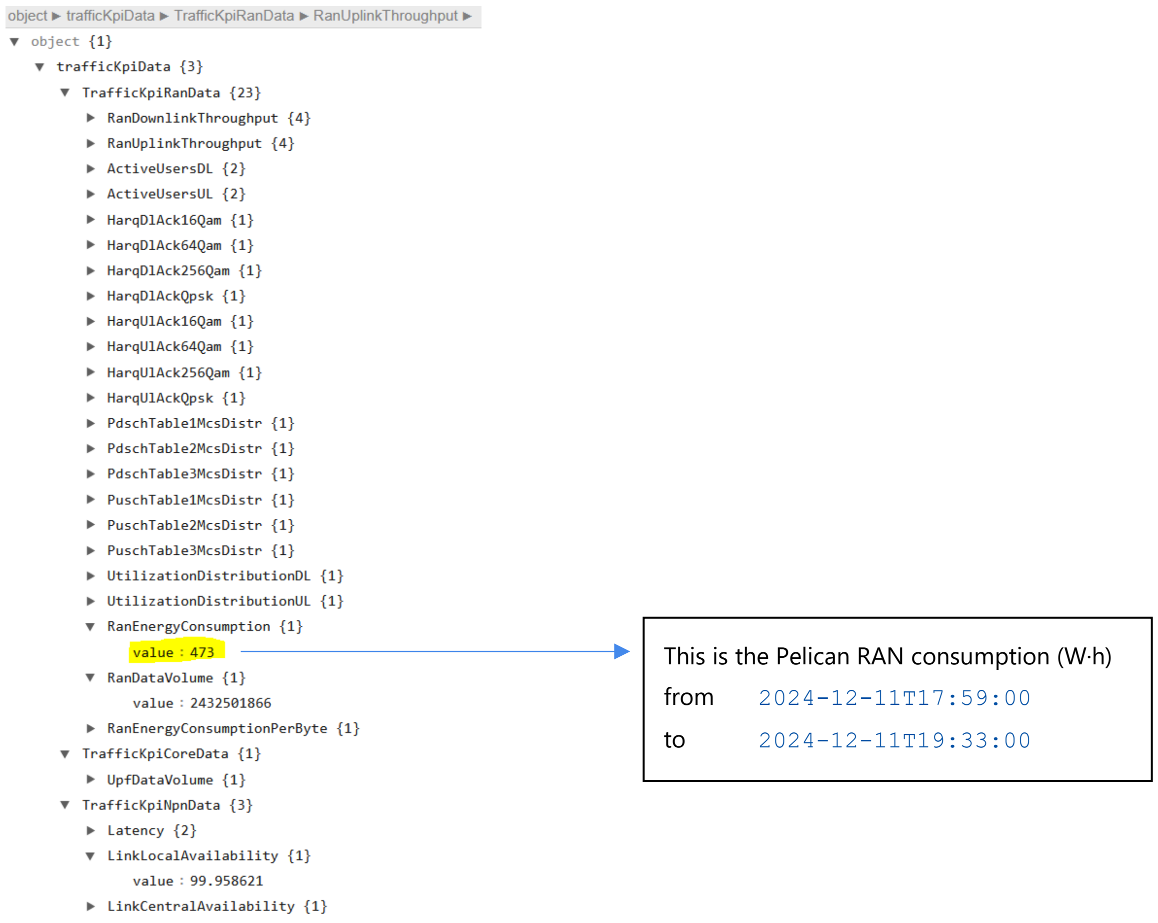


POST https://{{EXPOSURE_API_IP}}/5tonic-exposure/v1/analyticsexposure/Pelican/fetch

```
{
    "analyEvent": {
        "Value": "NPN_KPI"
    } ,
    "analyEventFilter": {
        "start": "2024-12-11T17:59:00Z",
        "stop": "2024-12-11T19:33:00Z",
        "cellId": "491717408"
    }
}
```

Among all the available data let's have a look to the energy consumption with the actual configuration:

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



```
object ▶ trafficKpiData ▶ TrafficKpiRanData ▶ RanUplinkThroughput ▶
▼ object {1}
  ▼ trafficKpiData {3}
    ▼ TrafficKpiRanData {23}
      ▶ RanDownlinkThroughput {4}
      ▶ RanUplinkThroughput {4}
      ▶ ActiveUsersDL {2}
      ▶ ActiveUsersUL {2}
      ▶ HarqDlAck16Qam {1}
      ▶ HarqDlAck64Qam {1}
      ▶ HarqDlAck256Qam {1}
      ▶ HarqDlAckQpsk {1}
      ▶ HarqUlAck16Qam {1}
      ▶ HarqUlAck64Qam {1}
      ▶ HarqUlAck256Qam {1}
      ▶ HarqUlAckQpsk {1}
      ▶ PdschTable1McsDistr {1}
      ▶ PdschTable2McsDistr {1}
      ▶ PdschTable3McsDistr {1}
      ▶ PuschTable1McsDistr {1}
      ▶ PuschTable2McsDistr {1}
      ▶ PuschTable3McsDistr {1}
      ▶ UtilizationDistributionDL {1}
      ▶ UtilizationDistributionUL {1}
      ▼ RanEnergyConsumption {1}
          value : 473
      ▼ RanDataVolume {1}
          value : 2432501866
      ▶ RanEnergyConsumptionPerByte {1}
    ▼ TrafficKpiCoreData {1}
      ▶ UpfDataVolume {1}
    ▼ TrafficKpiNpnData {3}
      ▶ Latency {2}
      ▼ LinkLocalAvailability {1}
          value : 99.958621
      ▶ LinkCentralAvailability {1}
```

This is the Pelican RAN consumption (W·h)

from 2024-12-11T17:59:00

to 2024-12-11T19:33:00

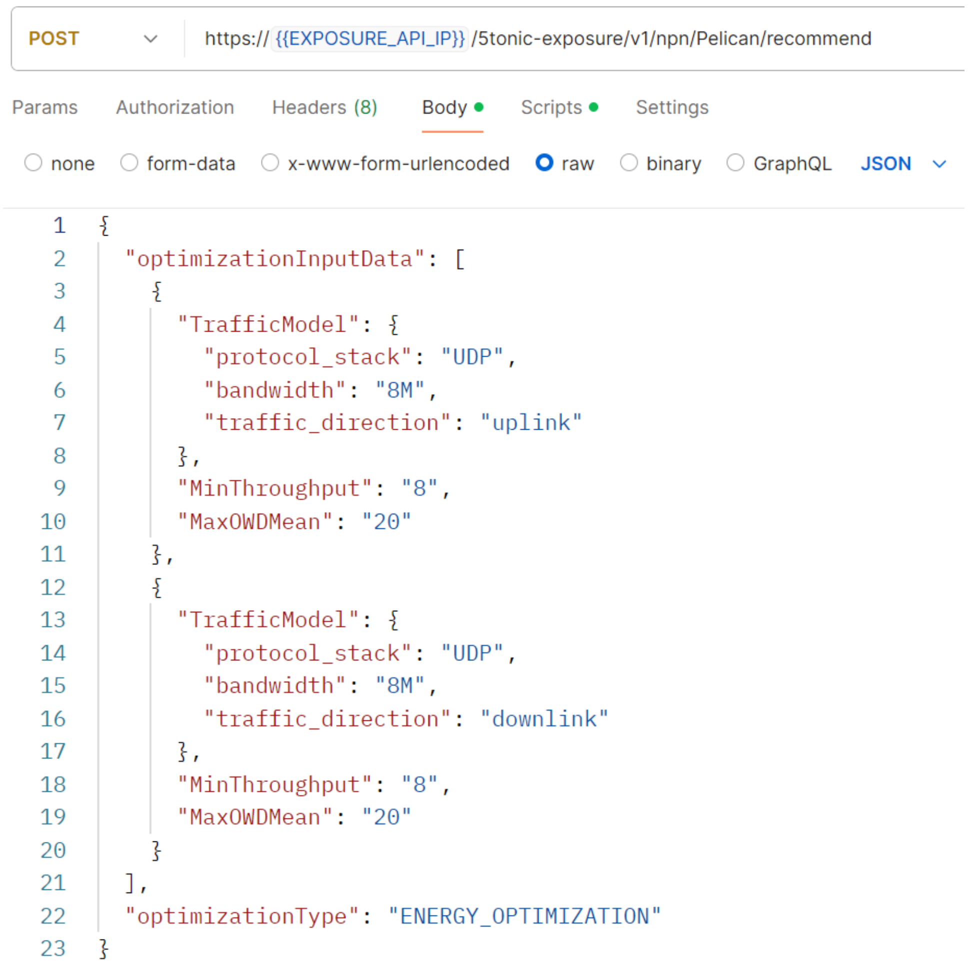
Pelican Energy consumption is 473W in 93′, giving a result of 301W·h every h

2.1.2.2.2.4 Optimized configuration for Energy consumption

From this point, we can leverage on the "recommender", which is a ML-based Exposure API, to try to find those configurations designed to meet the requested KPIs and traffic model requirements while getting most efficient one in terms of Energy consumption.

In the example below, as explained in section [2.1.2.1.3.1 API inputs] we provide as input the aggregated use case traffic models together with the minimum throughput the configuration must achieve (**MinThroughput**) and the maximum allowable mean one-way delay in milliseconds (**MaxOWDMean**) for the NPN configuration with the specified traffic models. Finally we indicate that the resulting feasible configurations should be ordered from lowest energy consumption to highest energy consumption **optimizationType: ENERGY_OPTIMIZATION**

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



```
https:// {{EXPOSURE_API_IP}} /5tonic-exposure/v1/npn/Pelican/recommend
```

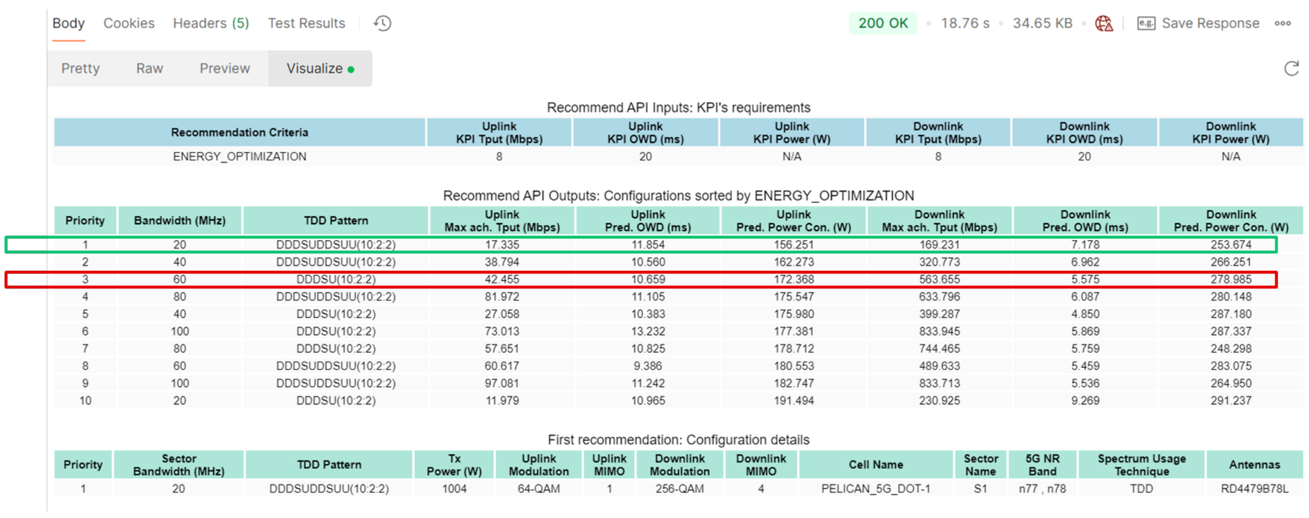
```
{
  "optimizationInputData": [
    {
      "TrafficModel": {
        "protocol_stack": "UDP",
        "bandwidth": "8M",
        "traffic_direction": "uplink"
      },
      "MinThroughput": "8",
      "MaxOWDMean": "20"
    },
    {
      "TrafficModel": {
        "protocol_stack": "UDP",
        "bandwidth": "8M",
        "traffic_direction": "downlink"
      },
      "MinThroughput": "8",
      "MaxOWDMean": "20"
    }
  ],
  "optimizationType": "ENERGY_OPTIMIZATION"
}
```

The ML algorithms supporting this API in NDT gave us the following results: For this particular real-world Use Case there are 2 additional configurations with less Energy consumption that the current configuration of

- **60MHz** DDDSU (10:2:2)

The optimal configuration in terms of Energy consumption that fulfills the requested KPIs (Tput, OWD and Energy Consumption) would be:

- **20MHz** DDDSUDDSUU (10:2:2)



Recommend API Inputs: KPI's requirements

Recommendation Criteria	Uplink KPI Tput (Mbps)	Uplink KPI OWD (ms)	Uplink KPI Power (W)	Downlink KPI Tput (Mbps)	Downlink KPI OWD (ms)	Downlink KPI Power (W)
ENERGY_OPTIMIZATION	8	20	N/A	8	20	N/A

Recommend API Outputs: Configurations sorted by ENERGY_OPTIMIZATION

Priority	Bandwidth (MHz)	TDD Pattern	Uplink Max ach. Tput (Mbps)	Uplink Pred. OWD (ms)	Uplink Pred. Power Con. (W)	Downlink Max ach. Tput (Mbps)	Downlink Pred. OWD (ms)	Downlink Pred. Power Con. (W)
1	20	DDDSUDDSUU(10:2:2)	17.335	11.854	156.251	169.231	7.178	253.674
2	40	DDDSUDDSUU(10:2:2)	38.794	10.560	162.273	320.773	6.962	266.251
3	60	DDDSU(10:2:2)	42.455	10.659	172.368	563.655	5.575	278.985
4	80	DDDSUDDSUU(10:2:2)	81.972	11.105	175.547	633.796	6.087	280.148
5	40	DDDSU(10:2:2)	27.058	10.383	175.980	399.287	4.850	287.180
6	100	DDDSU(10:2:2)	73.013	13.232	177.381	833.945	5.869	287.337
7	80	DDDSU(10:2:2)	57.651	10.825	178.712	744.465	5.759	248.298
8	60	DDDSUDDSUU(10:2:2)	60.617	9.386	180.553	489.633	5.459	283.075
9	100	DDDSUDDSUU(10:2:2)	97.081	11.242	182.747	833.713	5.536	264.950
10	20	DDDSU(10:2:2)	11.979	10.965	191.494	230.925	9.269	291.237

First recommendation: Configuration details

Priority	Sector Bandwidth (MHz)	TDD Pattern	Tx Power (W)	Uplink Modulation	Uplink MIMO	Downlink Modulation	Downlink MIMO	Cell Name	Sector Name	5G NR Band	Spectrum Usage Technique	Antennas
1	20	DDDSUDDSUU(10:2:2)	1004	64-QAM	1	256-QAM	4	PELICAN_5G_DOT-1	S1	n77 , n78	TDD	RD4479B78L

FIGURE 13. RECOMMENDER RESULTS FOR OPTIMAL CONFIGURATION IN TERMS OF POWER CONSUMPTION

In summary

- 20MHz Uplink-156W·h vs 60Mhz Uplink-172W·h give us a 9% energy saving
- 20MHz Downlink-253W·h vs 60Mhz Downlink 278W·h give us a 9% energy saving

As we can see, the recommended configuration not only allows us to comply with the UL Tput, DL Tput, OWD KPIs and a lower Energy consumption prediction, but also suggests working in a bandwidth of 20MHz (instead of the current 60MHz), which means a considerable saving of resources in the radio spectrum.

It is extremely important to note that this optimization exercise is applicable and can be extend to any of the traffic models and configurations with which the NPN has been characterized, and which have allowed training the ML prediction model of the NDT, so it can be applied to a wide variety of real use cases.

2.1.2.3 Achievements and lessons learned

Achievements:

- **Unlocked Synergies between NDT & NPN for enabling AI/ML supported APIs**

The implementation of the Network Digital Twin (NDT) platform marks a transformative advancement in the management and optimization of 5G Non-Public Networks (NPNs), enhancing the operational efficiency of PNI-NPNs.

One of the most significant achievements of the NDT platform is its comprehensive digital twin functionality, which includes performance monitoring, analysis, recommendations, and actuation. This positions the platform as an innovative, all-in-one solution that is ahead of anticipated standardization efforts. The scenario-specific modeling allows for precise optimization and learning tailored to PNI-NPN scenarios, offering insights and performance enhancements that generic analytics tools cannot achieve.

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 inclusion of specialized APIs—ranging from ML-based exposure to configuration, experimentation, and optimization—further extends the platform's capabilities. These APIs empower CSPs and enterprise customers to design, deploy, and optimize 5G connectivity use cases with greater precision and insight by having access to the real traffic patterns of the UEs through the analytics APIs, **allowing the enterprise to fine-tune the network configuration to obtain an optimization of resources based on real data**. The Machine Learning (ML) models implemented by the digital twin offer very accurate network performance prediction capabilities and thus anticipate the need for possible reconfigurations in your NPN system.

In summary, the Network Digital Twin platform provides a pioneering service that stands at the forefront of telecom innovation. By delivering comprehensive, scenario-specific solutions, it empowers organizations to fully harness the potential of their 5G Non-Public Networks, setting a new standard for network management and optimization. This innovative approach not only enhances current operational efficiency but also establishes a foundation for future advancements and standardizations in digital twin technologies.

- **Energy optimized Architectural design and technology applied**

In the context of 6GDAWN project, both the architectural options selected for the 5G NPN system supplied by Ericsson (Pelican) and the technology options already contribute to having an environment of experimentation where the energy consumption of the 5G system is extremely low and indeed optimized by design. Three main aspects contribute to that, namely:

1. PNI-NPN architecture: All the Control Plane functions related to pelican are executing in a remote data center (5TONIC) where the energy consumption is optimized and also mutualized with other workloads related to supporting other 4 NPN systems connected with 5TONIC in a similar set-up. SO, from the “enterprise” point of view the bill of electricity related to the many servers making the 5G Control Plane functionality is zero.
2. UPF location at CTTC: Since Pelican integrates an Ericsson LPG implementing the local UPF function, all User Plane traffic generated at CTTC is restricted to CTTC premises and the consumption of the necessary elements in the transmission path that would be needed for the case of relying on a Central Office UPF is also saved.
3. 5G Radio dots: Having selected 5G Radio dots for the indoor coverage, with max power of 1W (compared to other micro solutions for indoor coverage) allows a superior level of by-design energy efficiency

Lessons Learned

- **Energy consumption measurements**

Along the various research activities carried out in the project over the 5G NPN system supplied by Ericsson to the project (Pelican) measurements on energy consumption have been collected for several RAN configurations, traffic models, and for periods of collection from just 2 minutes up to 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

hour. Ericsson Radio system allows for querying for the consumption of energy of BBU, RRU (IRU in Pelican case) and Antennas (5G Dot in Pelican case) every 15 minutes. For the project we have implemented a mechanism to increase this frequency to just 30 seconds. That change comes along with the risk of (given the very low energy consumption of the system, as explained above) that the measurement for such a short period is affected by a larger relative error, although it may allow, anyway, to detect relative peaks or downs of energy consumption more effectively (timely). That said, three major learnings can be extracted from the analysis of the measurements collected:

1. The precision of the energy consumption measurements is 1 W·h. This means that for a measurement of the order of 2.0 W·h (for a period of 30'') since the potential error is given by that precision of 1 W·h, the relative error is as high as ½ = 50%
2. Tracking individual measurements of energy consumption (measurements collected every 30'') shows, for a constant stimulus of traffic, an apparent regular with fluctuation around a potential average value that it is simply a consequence of the error of the individual measurements compared with the measurements themselves. Take this example: Consider an actual average consumption of 2.5 W·h for periods of 30'', what the counter will be providing every 30'' is values of either 2 W·h or 3 W·h, being very unlikely that they offset each other and cancel the overall relative error of the accumulated consumption unless this is tracked for a long period of time (order of 1 h).
3. The high relative error induced by the precision level of the individual measurements can totally hide a sudden relevant variation of energy consumption (up to 50%) for quite some time too. This implies that checking the actual influence of anomalously high traffic on energy consumption cannot yield results of high confidence until quite some time has elapsed. Same happens with checking the effect of mitigation of that anaomaly (for instance isolating those users).

- **Data-driven approaches for improving energy consumption levels**

Combining both above-explained learnings (by-design optimized levels of energy consumption plus impossibility to secure effective actions on energy consumption optimizations when acting on very short periods of time) we have concluded that considering longer periods of actuation there are two types of approaches that can have a relevant effect. Over time, in reducing/further improving energy consumption levels of the 5G NPN system:

1. Considering advanced data-driven approaches for evaluating how to change the configuration of the RAN in order to safeguard the required performance KPIs and also save on energy consumption. This can be done by using NDT exposed APIs leveraging the trained model for energy consumption with 1-h long measurements collected for Pelican. The recommender API can give a very good indication of alternative configurations that may support target KPIs and also help reduce energy consumption. Moreover, using the configuration, experimentation and Analytics APIs that potential energy reduction can be easily confirmed for the real use case traffic model with an experiment of at least 1 h.

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. Considering a simple data-driven approach for identifying potential periods for activating deep sleep mode. The Analytics API enhanced by Ericsson in this project can be used to survey actual traffic in hour-intervals along arbitrarily long periods of time (weeks, months) in order to help spot those opportunities (periods of no traffic) and then program deep sleep regular intervals (for instance at the night shift, exactly from one safe hour to another) accordingly.

2.2 Use Case ELASTIC E2E network resources optimization

2.2.1 ELASTIC UC2 PoC1- Proactive infrastructure status prediction module

This PoC is intended to some functionalities that are considered relevant regarding the integration of the extreme-edge domain as part of the network continuum, in connection with the E2E Network Resources Optimization Use Case E2 described in the previous Deliverable. Specifically, the objective of this PoC is the implementation of a prototype of the so-called "Infrastructure Status Prediction Module" (ISPM) component, considering also its integration into the overall 6G DAWN architecture.

Figure 14 shows the high-level architecture approach to the PoC in line with the architectural components introduced in previous deliverables with some updates. That architecture aims a proactive orchestration provided by DMO, with data collection from devices by MS, analysis of it along with prediction patterns by ISPM, and decision-making on the services deployed on the infrastructure layer done by DE and ACT. All of these components achieve a Zero Touch Closed Control Loop.

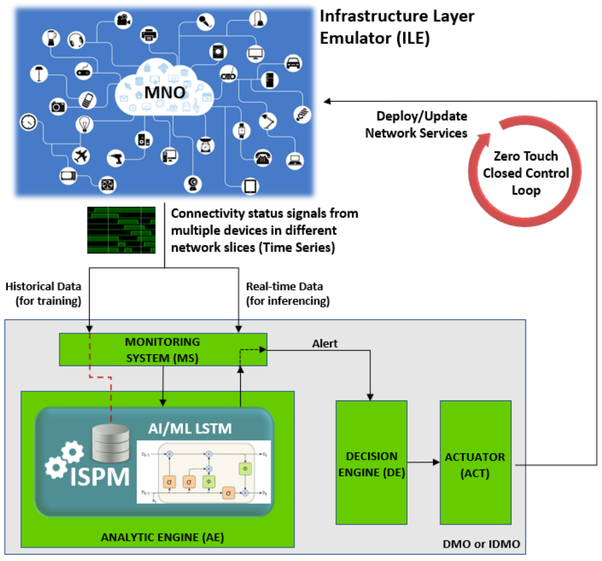

MNO
Infrastructure Layer Emulator (ILE)
Deploy/Update Network Services
Zero Touch Closed Control Loop
Connectivity status signals from multiple devices in different network slices (Time Series)
Historical Data (for training)
Real-time Data (for inferencing)
MONITORING SYSTEM (MS)
Alert
AI/ML LSTM
ISPM
ANALYTIC ENGINE (AE)
DECISION ENGINE (DE)
ACTUATOR (ACT)
DMO or IDMO

FIGURE 14. UC2 POC1 HIGH-LEVEL 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

2.2.1.1 PoC Implementation details

The following Figure shows how the architecture presented above has been implemented, from a lower point of view, showing the different interfaces presented in each component of the stack. Implementation details are reported deeply in the following sections.

The main interfaces are shown in the image below:

- ILE and MS: Through this interface each container sends its status to MS.
- MS and Serving Platform: MS preprocess the data from ILE and request the inference to Serving Platform which the model pretrained is running. Serving Platform (torch server) and model.mar act as the ISPM module.
- MS, RabbitMQ and DE: MS send the prediction made by ISPM to DE via rabbitMQ's asynchronous message queue.
- DE and ILE: DE modifies the deployment services through this interface.
- There are some auxiliary interfaces:
    - Front-End recollect data from ILE to show the behavior of infrastructure.
    - DataBase stores data to train new models with updated data.
    - Administrators send the model.mar through the REST or gRPC interface.

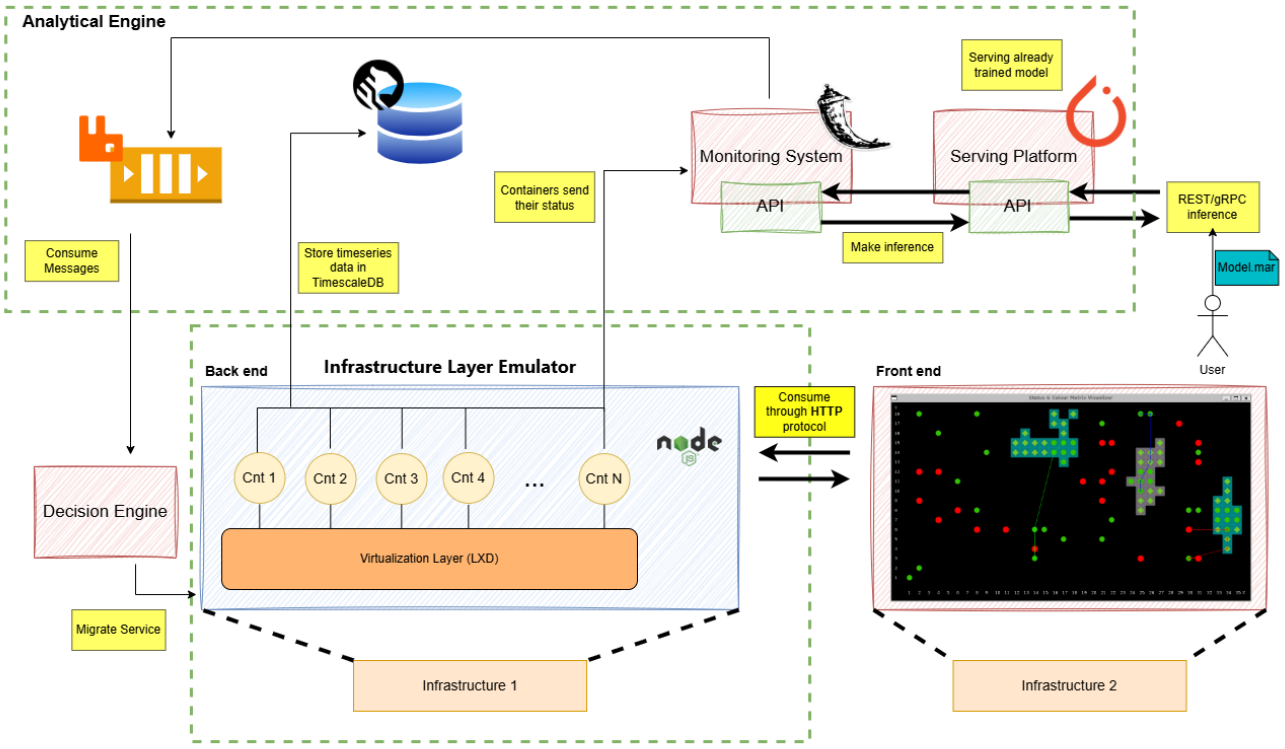


FIGURE 15. UC2 POC1 IMPLEMENTATION DIAGRAM

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

2.2.1.1.1 Analytical Engine

The Analytical Engine (AE) implements a time series prediction model using a Long Short-Term Memory (LSTM) network in Pytorch. Its primary purpose is to predict the future status of a set of containers based on historical binary status data and timestamp information. The model is particularly well-suited for sequential data and is designed to learn temporal patterns in how the containers behave throughout the day. It is designed to generalize across multiple containers and time intervals, and its output is a multivariate binary prediction indicating the expected state of each container at a future point in time, which is the input of the DE to take decision according to this predictions and others variables (explained in depth in the corresponding section 2.2.1.1.3)

The data is loaded from a PostgreSQL database with TimescaleDB extensions. The input dataset includes the binary status of each container over time, along with a timestamp. The continuous input is derived from the timestamp — specifically, the time of day is normalized and transformed using sine and cosine functions to represent its cyclic nature. This encoding helps the LSTM network understand patterns related to daily cycles. Ground truth labels represent the actual container state that the model should predict, that are also retrieved from the database.

Finally, the AE's model is served using TorchServe, which is a flexible and production-ready model serving library developed by PyTorch. In particular, TorchServe is deployed using official docker image aligned with cloud-native requirements.

Figure 16 shows the training and predictions workflow:

In the training phase, the user launches the training and data requests are made to the database, this database contains the state of each device at each moment, and the future state called "ground_truth". The preprocessing module normalizes the timestamps and sends these sequences to the trainer, which performs the training cyclically and saves the model.

In the service phase, the administrator serves the model (saved in the previous phase) to the AI server (Torch Server), which performs its processing, after that the MS can make inferences through the API described in Figure 16.

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

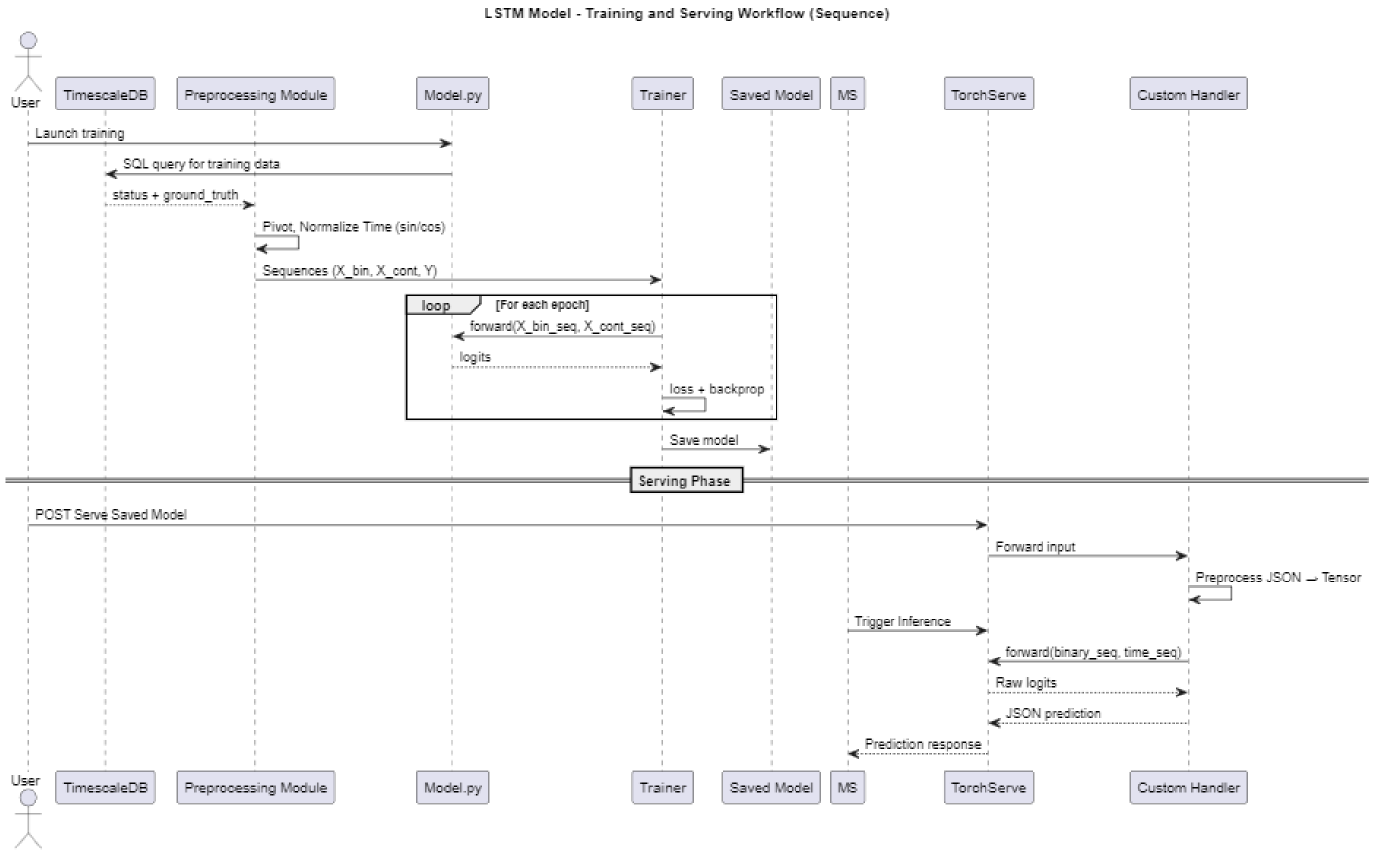


FIGURE 16. UC2 POC1 TRAINING AND SERVING WORKFLOW

2.2.1.1.2 Monitoring System

The Monitoring System (MS) is responsible for collecting and processing the data received from the network infrastructure. This component also requests the prediction to the AE and forward it to the Decision Engine (DE). All these interactions and interfaces are implemented using API Rest that offers high scalability, flexibility and is a widely used and well-established HTTP. The API Rest used by MS is implemented using flask.

The following Figure 17 shows the sequence of messages and interactions between different components of the PoC with the MS. This sequence shows how the MS collects data from each of the Network nodes and counts it, to request the inference to the AE, the result of this inference is processed and sent to RabbitMQ microservice.

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

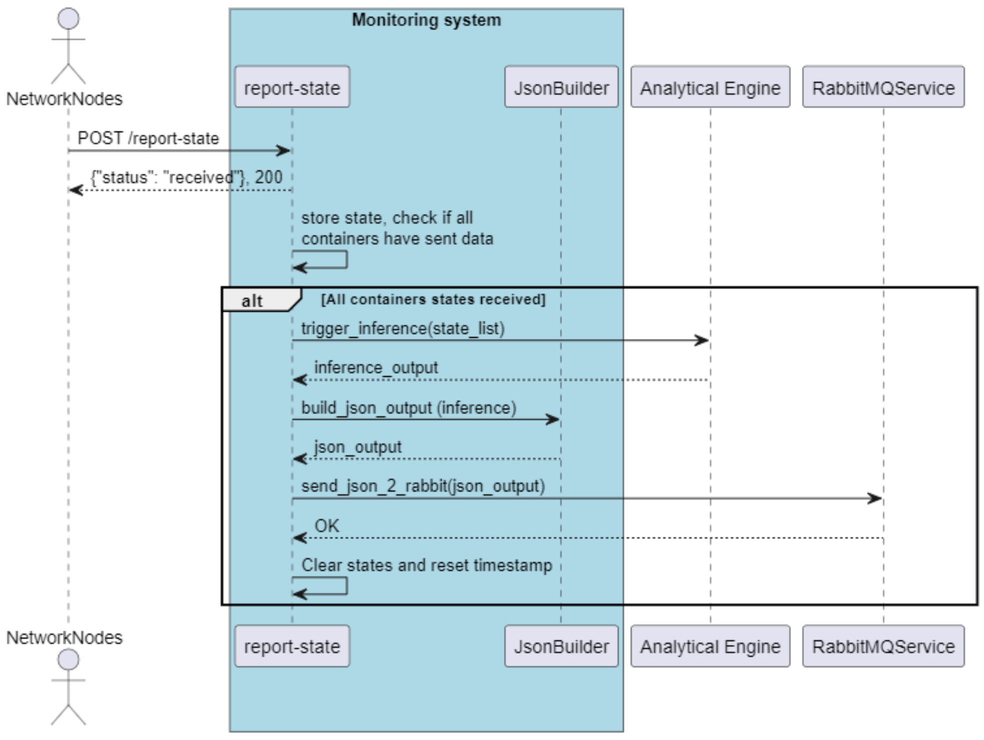

Monitoring system
NetworkNodes
report-state
JsonBuilder
Analytical Engine
RabbitMQService
POST /report-state
{"status": "received"}, 200
store state, check if all containers have sent data
alt
[All containers states received]
trigger_inference(state_list)
inference_output
build_json_output (inference)
json_output
send_json_2_rabbit(json_output)
OK
Clear states and reset timestamp

FIGURE 17. UC2 POC1 MONITORING SYSTEM WORKFLOW

2.2.1.1.2.1 Discovery Mechanism

The system updates the status of each node through an API Rest which collects data from all nodes and processes it to be served in the correct format to the Analytical Engine (AE). The endpoint is called "report-state" and is implemented through a POST action using Flask library. For this PoC, the nodes report the following information to the MS:

- TimeStamp, indicates the time when the container sends the update.
- State that, in this particular case, considers only two states (ON/OFF).
- Container ID, that identifies each container univocally and is assigned at the deployment phase.

From the point of view of the MS, that process is a reactive mechanism, since each node starts the process of discovery.

2.2.1.1.2.2 Auxiliary tools

There are some auxiliar tools that are represented in Figure 17:

- RabbitMQ: To Connect MS and DE.
- TimescaleDB: To store data for model train.
- Pytorch: To serve the AI/ML Models.

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

To implement communication between the MS and the DE, RabbitMQ has been used, which is an open-source message broker that delivers messages to a receiver (DE) in an asynchronous way. This characteristic makes it great for our purpose: decoupling services (easier to integrate components from different stakeholders) and scaling systems (one of the main goals of 6G-DAWN Project). It is also a cloud native option due to the fact that there are official docker images and is oriented to microservices communication.

Regarding the data of the system, TimescaleDB has been used under the same assumptions as RabbitMQ: Scalability and feasibility to integrate and cloud-native orientation. It is also a time-series database to manage large volumes of time-stamped data, which is an important feature to manage data from the containers and infrastructure monitoring.

There is also a Graphic User Interface that has been used to show the migrations applied by the proposed system during the simulation of the volatile infrastructure.

2.2.1.1.3 Decision Engine

The Decision Engine (DE) is a central component responsible for intelligent orchestration and automated migration of services within a Kubernetes-based cloud infrastructure. Its primary function is to ensure service continuity and optimal resource utilization by actively monitoring the health and status of compute nodes, and dynamically re-assigning services as the environment changes.

The DE in this PoC is providing key values as follows:

1. Resilient Service Placement and Migration

DE continuously listens to updates regarding the operational status of all nodes in the Kubernetes cluster, which are communicated via a message bus (RabbitMQ). Each node may transition between ON (available) and OFF (unavailable) states, for example due to failures, scheduled maintenance, or network disruptions. The DE ensures that key services are always running on available (ON) nodes and are moved promptly from nodes that become unavailable.

2. Resource-Aware Decision Making

Rather than statically assigning services to nodes, the DE considers the real-time resource capacity (CPU and memory) of each available node. For every service, it calculates a dynamic "score" for each ON node, based on their current allocatable CPU and memory resources, giving preference to nodes with higher available capacity. This prevents overloading of nodes and contributes to the overall performance and resilience of the cluster.

3. Autonomous and Granular Operations

- Initial Placement: Upon detecting at least three ON nodes, the DE automatically assigns each service to one of the available nodes, forming an initial, balanced placement.

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

- Continuous Adaptation: As the cluster state evolves (e.g., nodes go offline or recover), the DE autonomously decides if any service needs to be migrated to a better or currently available node.
- Seamless Migration: If a node hosting a service is predicted to go offline, or a better resource becomes available, the DE triggers migration. This involves temporarily marking the current node as unavailable for new workloads (via "tainting") and reallocating the service to a new node, ensuring minimal disruption.

4. Event-Driven and Synchronized Operation

DE operates in an event-driven manner, instantly reacting to every new update about node status received via RabbitMQ. This ensures the engine is always synchronized with the real-time cluster state and can promptly initiate mitigation or migration actions.

5. Transparency and Auditability

All migration decisions and actions taken by the DE are transparently logged. This provides operators with a clear historical record of when and why services were moved, supporting operational transparency and facilitating root-cause analysis in the case of disruptions.

In nutshell, the DE acts as a real-time "decision-maker," always aware of the health and resource state of the entire cluster, and ready to relocate services whenever necessary to maintain optimal operation and availability.

2.2.1.2 E UC2 PoC1 KPIs Evaluation and Results

Table 10 shows the KPIs, as defined in E3, for the Elastic UC2 PoC1, as well as the refence points of evaluation defined for the final pilot implementation.

Due to the project evolution, some of these KPIs have been merged in just one, while others have been removed; for each case, the reasons are also explained in the following table.

TABLE 10. UC2 POC1 KPI EVALUATION

KPI	KPI	Type	Definition	Evaluation Points
Infrastructure Size	Nodes	Capacity	Number of nodes in the network infrastructure layer, including both: physical and virtual nodes.	35 nodes, including 25 at the extreme edge, 6 at the edge and 4 at the core
Infrastructure heterogeneity	Bits	Entropy	Measure of the different kind of devices in the infrastructure, computed according to the Shannon Diversity Index.	Shannon Diversity Index equal to **H=0,68**. There are 20 alpine nodes and 15 ubuntu nodes.

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

Network Volatility Index (NVI)	Percentage	Volatility	Number of devices in the network that have changed their status during a given period of time over the total number of devices in the network.	All the extreme edge nodes change their status each 24 hours. Around 25% of extreme edge nodes change their status each 3 hours.
Device availability	Percentage	Availability	Percentage of uptime in a given year for each specific device in the network.	Since execute the simulation for a year, it has been taken a week to measure the device availably during a long period of time, each extreme edge device has an uptime around 30%. Other nodes, like cores nodes, are assumed 100% of uptime
MTBF	Seconds	Performance	Mean time between failures, applied to the services deployed on the infrastructure (the network volatility can impact on this KPI). Calculated using an arithmetic mean over a period.	This KPI is calculated using packets flow captured (fig 18 and 19) during simulations. - Reactive case: 75% Uptime and 0.06% of retransmission packet. - Proactive case: 99% uptime and 0.03% of retransmission packet due to migration process.
MTTR	Seconds	Performance	Mean time to recovery, applied to the services deployed on the infrastructure (the network volatility can impact on this KPI). Calculated using an arithmetic mean over a period.	This KPI is directly related to the service and how it is deployed on the infrastructure. In this particular case, the service is light, and its images are pre-

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

				pulled. Under this condition, MTTR stays between 10 and 20 seconds. That case only occurs when all replicas of the service are down.
Services QoS	Multiple	Quality	The infrastructure volatility may affect the QoS of the deployed services. The measurement of this KPI should be based on the selection of those specific service KPIs considered relevant for each service, such as latency, jitter, packet loss, bandwidth, etc.	In particular, latency and bandwidth metrics have been used. These metrics are shown in the figure 22.
Services QoE	Multiple	Quality	The infrastructure volatility may affect the QoE perceived by users for the deployed services. The measurement of this KPI should be based on the service end-uses feedback	During the PoC the video stream, the service did not stop at any time; this implies that the user did not detect any of the migrations done.
Services deployment time	Seconds	Performance	It may vary depending on the infrastructure resources available at any given time, including the re-deploy migration time.	Using a pre-pulling image of the service, less than a minute. This KPI is closer to MTTR.
Device detection time	Seconds	Performance	Regarding the devices discovery mechanism.	Around ~15 sec, since system updates device status each 30seg.
Devices discovery accuracy	Percentage	Accuracy	Proportion of successfully discovered devices compared to total actual devices in the network.	During PoC, with an emulation environment, there are no errors during the discovery period.
Devices Discovery Mechanism Traffic	Bits per second	Performance	Amount of network traffic generated by the device's discovery mechanism.	200 Bytes per node each 30 seg in a speed up simulation (mechanisms explained in 2.2.1.1).
Devices Discovery Registry Size	Bits		Amount of memory required to store the relevant information	In the PoC, 16977 KB for 32 nodes for one week.

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

			about the devices available in the infrastructure, regardless of how the storage mechanism is implemented.	
Devices Discovery System Scalability	Multiple	Scalability	Impact of the extension of the infrastructure pool of resources on the Devices Discovery System. The measurement of this KPI should be performed by increasing the Infrastructure Size to some extent, and by evaluating the related Performance, Quality and Accuracy type KPIs.	This KPI is covered by three previous ones.
Service adaptation time	Seconds	Performance	Time a network service needs to adapt to infrastructure changes.	This KPI is very similar to redeployment time and MTTR. It is achieved in around 25-30 seconds.
Device status update detection time	Seconds	Performance	Regarding the device´ monitoring mechanism.	The status of the devices is updated every 30 seconds.
Device status update prediction time	Seconds	Performance	Time it takes to anticipate the change of state of a device – inference time.	The inference real time is 3 hours, but during the PoC execution the inference is performed every 30 seconds due to the speed up modification.
True Positive Prediction Rate	Percentage	Accuracy	Number of events properly predicted, regarding the total amount of events.	At PoC, all the events are properly predicted, due to simulations constraints.
Predictions System Scalability	Multiple	Scalability	Ability of the predictions system to handle an increase in the number of devices of the infrastructure without degradation. The measurement of this KPI	This KPI is not directly checked during the PoC, but, according to other scalability KPI, the system can increase the number of devices since each

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

			would be performed by increasing the Infrastructure Size to some extent, and by evaluating the related predictions related KPIs	node sends low amount of information to the network.

2.2.1.2.1 Results

The analysis of the results obtained during the development of the PoC focused on the comparison between reactive or legacy behaviors and proactive behaviors through the use of artificial intelligence.

The following graph shows the numbers of packets received by users during the execution of volatile infrastructure. It was created using pcap files with different simulations.

The Figure 18 shows how the continuum of the service is interrupted due to the volatility of the simulated infrastructure by ILE. In that case, the different replicas of the services are managed in a legacy way; when the host of each replica switches its status to down, a new node is searched to host the replica. That behavior causes that, at certain times, no replica of the service is active, and the service is unavailable for the users. That situation is shown in the following figure, where there are some gaps in which there are no packets transmitted by the service.

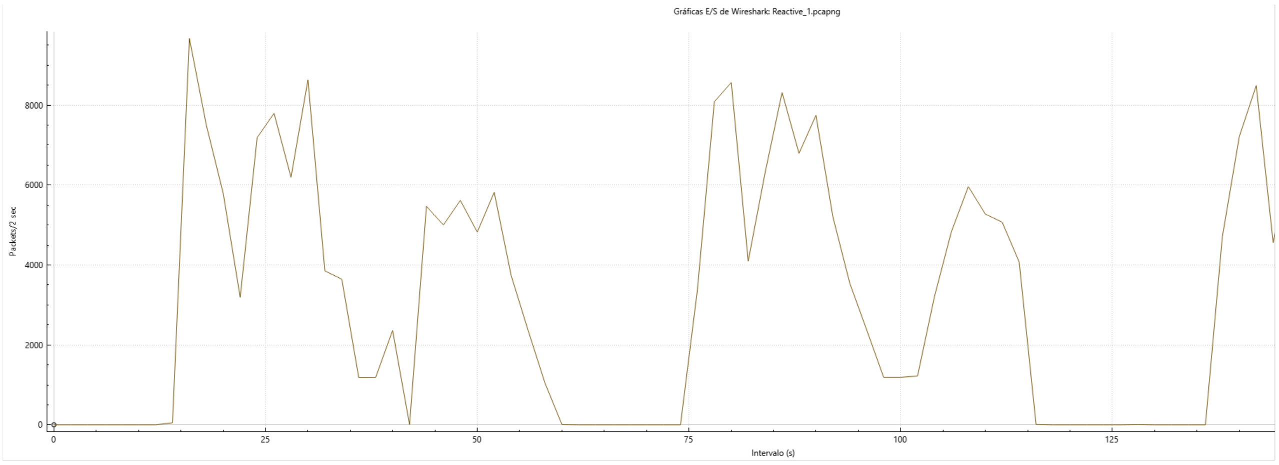


FIGURE 18. UC2 POC1 REACTIVE BEHAVIOUR

On the other hand, with a proactive behavior, the continuum of the service isn't interrupted due to the stack proposed for UC2 PoC1 with a prediction module implemented using AI/ML technologies (as explained in PoC Implementation details 2.2.1.1). In the following figure, the gap that appears in reactive simulation doesn't exist and the service is available for users during the whole simulation.

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

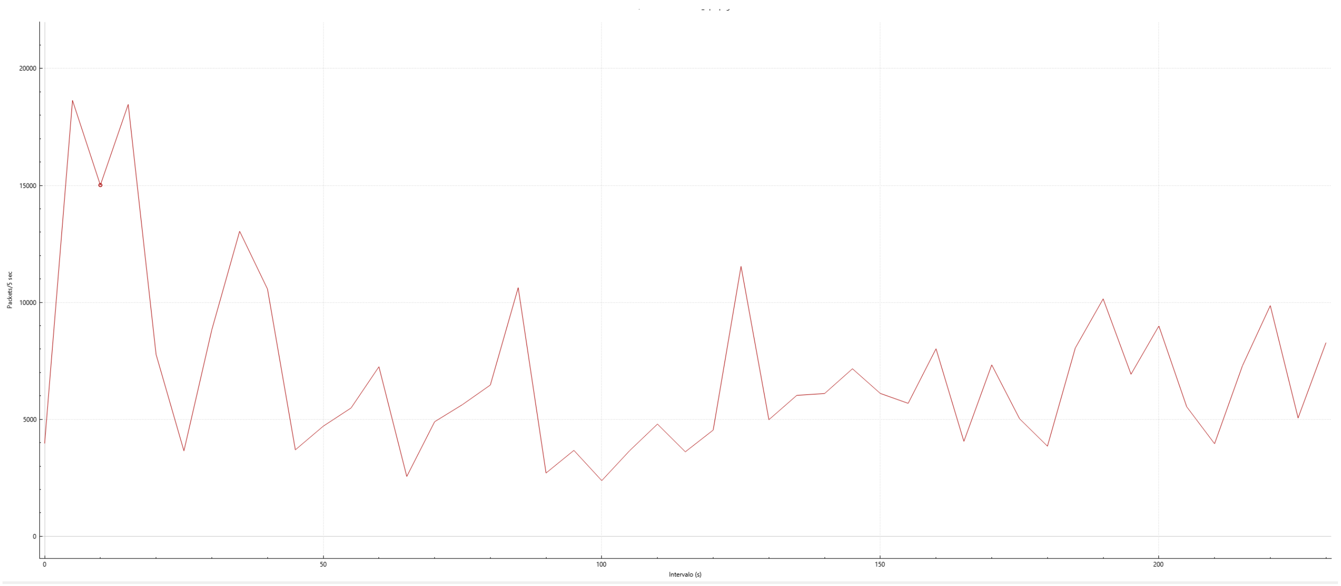


FIGURE 19. UC2 POC1 PROACTIVE BEHAVIOUR

The last graph shows the packet per second during a long simulation (around a week with a speed up modification).

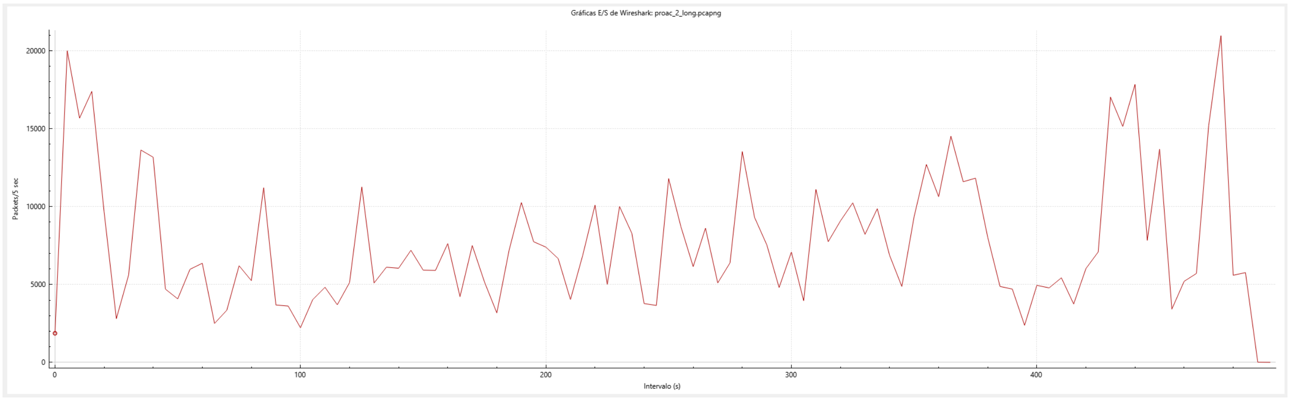


FIGURE 20. UC2 POC1 REACTIVE BEHAVIOUR (LONG SIMULATION)

Finally, some of KPIs explained in Table 10 are calculated using data offered by the pcap files:

- MTBF: Uptime is calculated adding the time without service and comparing with the total time of the simulation. The retransmission packets are calculated through wireshark tool.
  - Reactive: 271K of packet with 160 packets retransmitted.
  - Proactive: 326k packet with 120 packets retransmitted.
- MTTR: This period is calculated by checking and measuring the gap without packets during the simulations.

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

Others KPIs calculation:

- Infrastructure heterogeneity, Shannon Diversity Index calculated as:

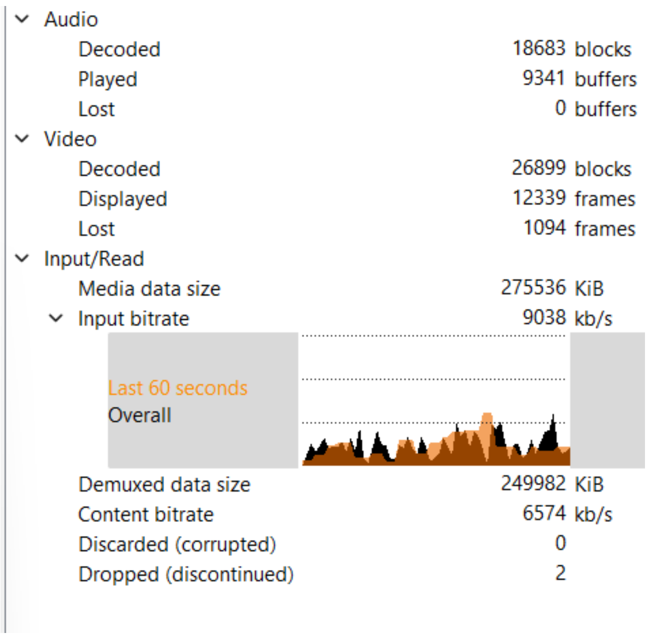
$$p_1 = \frac{20}{35} = 0.5714$$

$$p_2 = \frac{15}{35} = 0.4286$$

$$H = -(p_1 \, Ln \, p_1 + p_2 \, Ln \, p_2) = 0.68$$

QoS and QoE have been measured with the information provided by the VLC software, which is a media player that has been used to test the service deployed on the volatile infrastructure. VLC shows real-time statistics about the currently playing media, which are useful to measure the QoS and its degradation.

The following image (Figure 21) shows how Service works with a stable virtualized infrastructure simulated by ILE without migrations of the service.



```
Audio
    Decoded                          18683 blocks
    Played                            9341 buffers
    Lost                                 0 buffers
Video
    Decoded                          26899 blocks
    Displayed                        12339 frames
    Lost                              1094 frames
Input/Read
    Media data size                 275536 KiB
  Input bitrate                       9038 kb/s
    Last 60 seconds
    Overall
    Demuxed data size               249982 KiB
    Content bitrate                   6574 kb/s
    Discarded (corrupted)                0
    Dropped (discontinued)               2
```

FIGURE 21. UC2 POC1 QOE ON STABLE INFRASTRUCTURE

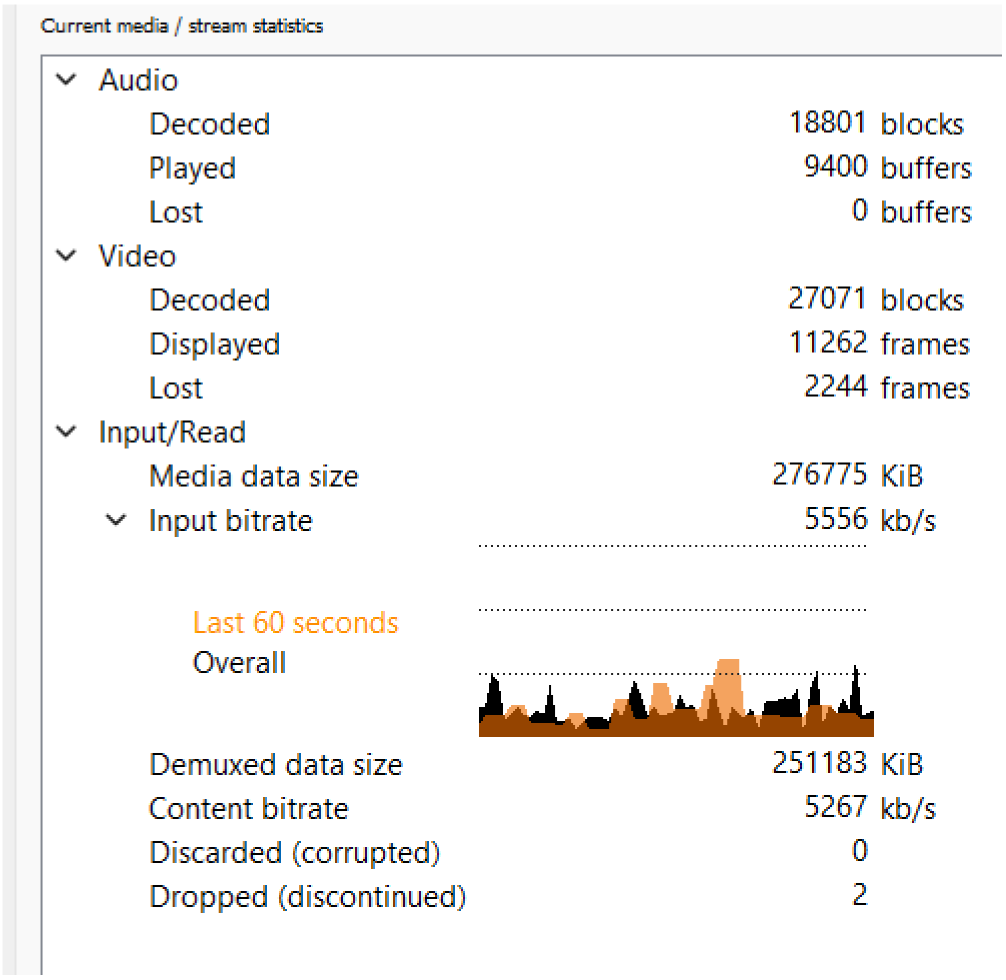
In that case, the Video Lost field shows Network issues due to that the option "--network-caching" is set to 0, configuration that maximizes sensitivity to network instability, in order to verify the stability of the service during the PoC. Instead of this configuration, VLC still use a minimal technical buffer at lower levels (decoder and OS), but network-caching equal to 0 disables the user-defined extra buffer meant for network smoothing, since there's no buffer to absorb network jitter, any delay, dropped packet, or fluctuation may cause service disruption. Also, the Hardware limitations (CPU decoder) have an impact on the loss frames.

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 “Input bitrate” section shows a real-time graph of input bitrate and the buffers:

- The **X-axis** represents time (in seconds).
- The **Y-axis** shows bitrate (in kbps).
- The curve shows how stable or variable the stream is.

Taking these statistics as a starting point, the following Figure 22 shows the comparison with a proactive behavior:



Current media / stream statistics

Audio	
Decoded	18801 blocks
Played	9400 buffers
Lost	0 buffers
Video	
Decoded	27071 blocks
Displayed	11262 frames
Lost	2244 frames
Input/Read	
Media data size	276775 KiB
Input bitrate	5556 kb/s
Demuxed data size	251183 KiB
Content bitrate	5267 kb/s
Discarded (corrupted)	0
Dropped (discontinued)	2

FIGURE 22. UC2 POC1 QOE ON PROACTIVE BEHAVIOUR

By comparing the statistics gathered from both the stable and volatile infrastructures, it can be observed that, although service continuity is maintained in both cases—thanks to the software designed and developed in the previously detailed proof of concept (PoC)—the number of lost frames is higher in the unstable execution. This increase is primarily due to service migrations occurring during runtime.

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

In particular, the QoE KPI, which is a subjective user perception of playback quality, is also demonstrated by the recorded video that is available at [5]. The media player never stops during the execution of the proactive prediction system, executing service migrations in a transparent way for the final user of the service.

2.2.1.3 Achievements and lessons learned

The implementation and integration of the Proactive Infrastructure Status Prediction Module, in line with the continuum orchestration concept, tries to integrate the volatile extreme-edge domains with stable cloud and edge infrastructures resources, which is the main goal of this PoC.

In particular, the set of components described in the present section met the specified requirements defined in Deliverable E3. The platform was designed using cloud-native, microservices-based architecture orchestrated by a DE supported by Kubernetes, enabling continuous deployment and robust lifecycle management. It features a comprehensive infrastructure data model that captures critical device attributes across core, edge, and extreme-edge nodes. A monitoring and diagnostics framework ensures continuous observation of these devices, while embedded AI/ML models enable predictive analytics to forecast device behavior and prevent service disruptions.

Service slicing is supported through the deployment of service replicas across the simulated infrastructure, and a responsive Decision Engine interprets analytics to trigger real-time orchestration actions such as service migration. Service deployment and assurance policies are aligned with Kubernetes orchestration principles, ensuring resilience even under volatile infrastructure conditions. Performance and service quality are validated through simulations and KPI exposed.

Beyond the technical achievements, the main lesson learned from this PoC is that AI/ML models, combined with orchestration software, enhance service continuity in volatile infrastructure, as the system is able to anticipate infrastructure volatility and proactively adapts.

[5] https://www.youtube.com/watch?v=YfTgk6bNM64

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 Conclusions

6G DAWN delivers a Decentralized AI closed loop (MS-AE-DE) that is implemented, validated with KPIs, by exploiting inter-building block interfaces across all ELASTIC PoCs, those defined in previous deliverable E3.

The implemented decentralized 6GDAWN AI (MS-AE-DE) achievements in ELASTIC PoCs are as follows:

- Decentralized AI-driven probabilistic forecasting which is combined with xApp-based automation reduce energy consumption in an O-RAN environment by dynamically managing Radio Unit activity without impacting service quality. The PoC results show that shutting down one RU leads to approximately a 50% reduction in RU-related energy consumption while maintaining service continuity.
- Energy optimization system in NPNs using a Digital Twin successfully demonstrated the capabilities of the NPN system by reproducing a vertical use case that exploits the data and analysis models provided by the Digital Twin System to optimize energy efficiency at critical sites, achieved by trading off some performance characteristics delivered by the network.
- The implementation and integration of the Proactive Infrastructure Status Prediction Module, in line with the continuum orchestration concept, tries to integrate the volatile extreme-edge domains with stable cloud and edge infrastructures resources. The PoC develops cloud-native, microservices architecture orchestrated by a Kubernetes-based Decision Engine for continuous deployment and lifecycle management. It unifies monitoring and diagnostics of core, edge, and extreme-edge devices through a rich infrastructure data model and embedded AI/ML for predictive analytics to prevent service disruptions.

Overall, this deliverable illustrates the successful implementation and validation of decentralized AI (MS-AE-DE) in all defined ELASTIC PoCs of 6GDAWN through their defined KPIs in E3.

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