

6G-SANDBOX

Supporting Architectural and technological
Network evolutions through an intelligent, secured
and twinning enabled Open eXperimentation
facility

Deliverable D5.5

FINAL REPORT ON KPI AND KVI VALIDATION

Date: --/--/----

Version: V1.0

Grant Agreement	101096328
Document number	D5.5
Document title	Final Report On KPI and KVI Validation
Lead Beneficiary	KEYD
Editor(s)	Filip Ivanovic (KEYD)
Author(s)	Michael Dieudonné (KEYB), Pedro Merino, Pablo Herrera, María del Mar Moreno, Javier Jiménez, María del Mar Gallardo, Rafael López-Gómez, Laura Panizo, Delia Rico (UMA), Filip Ivanovic, Matteo Pagin, The Dao (KEYD), Harilaos Koumaras, George Makropoulos, Vasilis Pitsilis (NCSRD), Fofy Setaki, Elina Theodoropoulou, George Lyberopoulos (COS), Antti Pauanne (OULU), Vaios Koumaras, Eugenia Vergi, Georgios Koumaras (INF), Bjoern Riemer (FOKUS), Bruno Rodriguez (ON)
Dissemination level	Public / Sensitive
Contractual date of delivery	30 June 2026
Status	Draft / Final
Version	V1.0
File name	6G-SANDBOX_Dx.x_Vy.y.pdf

Document History

Version	
V1.0	First Version, ready for peer review
Vy.y	Description of changes
Vy.y	Description of changes

TABLE OF CONTENTS

Keywords	1
1 Introduction	1
1.1 Notes on General Methodology	1
2 Reports of KPI Testing Efforts.....	1
2.1 KPIs measured using UESIM	1
2.1.1 NCP, ARR, SPE	2
2.1.2 NEE	12
2.2 KPIs relating to Signal Quality	5
2.2.1 Measurements at NCSRD	5
2.2.2 Measurements at OTE.....	9
2.3 KPIs relating to NFV Energy Efficiency.....	11
2.3.1 Results	12
2.4 KPIs relating to Device Energy Efficiency.....	13
2.4.1 Testbed Setup	13
2.4.2 Results and Analysis	14
2.5 KPIs relating to VM Compute Measurements	15
2.5.1 CUU	24
2.5.2 RAU	24
2.5.3 ECU	25
2.5.4 SOL	26
2.6 KPIs relating to Orchestration	20
2.7 KPIs relating to Localization	22
2.7.1 Scope	22
2.7.2 Functionality Methodology and Result Description	22
2.8 KPIs relating to RIS.....	24
2.8.1 Capacity Enhancement.....	24
2.8.2 Update Rate	27
2.9 KPIs relating to EMF Exposure	28
2.9.1 EMF limits for general audience and occupational persons.....	29
2.9.2 UOulu EMF measurement of 28 GHz basestation	30
2.9.3 EMF results of 28 GHz base station	32
2.10 KPIs relating to Sensing	33
2.10.1 6G-Sandbox sensing architecture/Methodology	33
2.10.2 Quest Headset-based sensing.....	34
2.10.3 External sensing with Physiotrack.....	36
2.10.4 Combining external solutions and headsets – The 6G-SANDBOX Sensing Layer	37
2.11 KPIs relating to Anomaly Detection.....	39
2.11.1 Anomaly Detection on N2 Interface	40
2.11.2 Anomaly Detection on N6 Interface and Near to Real-Time Reaction	43

2.12	KPIs relating to Security and Conformance Testing	49
2.12.1	Test Setup.....	50
2.12.2	Results	52
2.13	KPIs relating to AI/ML	53
3	<i>Report on Continuation of Benchmarking Efforts.....</i>	55
3.1	Phase 2	56
3.2	Phase 3	58
3.3	Phase 4	60
3.3.1	Consistency	61
3.3.2	Interference.....	62
3.3.3	CPU Allocation.....	66
3.3.4	Memory Allocation.....	68
3.4	Benchmarking: Final Analysis and Conclusions.....	71
4	<i>Report on Digital Twin Efforts.....</i>	74
4.1	EXata Digital Twin in Athens	74
4.2	Formal Digital Twins in UMA.....	79
4.2.1	The Case Study: PTP Protocol	80
4.2.2	Automated Modeling of PTP as a Formal Digital Twin	82
4.2.3	Experimental Results.....	84
4.2.4	Evaluation of the Approach	86
5	<i>Final KPI Overview.....</i>	86
6	<i>Conclusion.....</i>	89
7	<i>References.....</i>	89

List of FIGURES

Figure 1: Downlink data bursts in the example scenario. PHY throughput maxes out at 653.7 Mbps, indicating the DL NCP for this cell.....	3
Figure 2: Uplink data bursts in the example scenario. PHY throughput maxes out at 77.9 Mbps, indicating the UL NCP for this cell.	3
Figure 3: Power consumption of the server running the O-CU/O-DU while performing data bursts. A single UE is emulated in this case.	4
Figure 4: Power measurements in different conditions during downlink bursts.	4
Figure 5: The two gNBs (Amarisoft and Ericsson) located at the NCSR campus, marked in yellow and red respectively	5
Figure 6: DL signal quality metrics acquired at the Athens platform (Ericsson RAN) using Keysight's Nemo Handy Field Measurement Tool	6
Figure 7: DL RSRP trace acquired in the proximity of the Ericsson gNB at the NCSR Campus.....	6
Figure 8: DL SINR trace acquired in the proximity of the Ericsson gNB at the NCSR Campus	7
Figure 9: Amarisoft RAN equipment for signal quality metrics acquisition at NCSR Campus.	7

Figure 10: UL signal quality metrics acquired at the Athens platform (Amarisoft RAN) using Keysight's Nemo Handy Field Measurement Tool.....	8
Figure 11: UL SINR trace acquired in the proximity of the Amarisoft gNB at the NCSR Campus	8
Figure 12: UL RSRP trace acquired in the proximity of the Amarisoft gNB at the NCSR Campus	9
Figure 13 - Ericsson Radio 4408 n78R equipment at OTE Academy Campus.....	9
Figure 14: Ericsson signal quality metrics acquired at OTE Academy Campus using Keysight's Nemo Handy Field Measurement Tool.....	10
Figure 15: RSRP trace acquired in the proximity of the Ericsson gNB at the OTE Campus	10
Figure 16: SINR trace acquired in the proximity of the Ericsson gNB at the OTE Campus	11
Figure 17 - Energy usage in mJ of the 3 main components of an example TN deployed on the Berlin Platform ..	12
Figure 18 - Energy usage in mJ of the Core (Open5GS) of an example TN deployed on the Berlin Platform. The UPF consumes such a large amount that all other processes have been grouped for better visibility	12
Figure 19 – Non-UPF Energy usage in mJ of the Core (Open5GS) of an example TN deployed on the Berlin Platform. This is the same proportion marked “other NFs” in Figure 18	13
Figure 20 - Energy Efficiency for the Core (Open5GS) of an example TN deployed on the Berlin Platform. The average value in this case is 32804,81 Bytes per Joule.	13
Figure 21 – Diagram of the testbed used for Device Energy Efficiency measurements	14
Figure 22: Graph showing Device Energy Efficiency Measurements over time for both the Idle and Fully Loaded states	15
Figure 23 - CPU Usage results for an example TN running multiple different tests on the Berlin Platform. The CPU usage does not appear to change significantly when the amount of UEs present in the test is changed, or when the direction is changed from Uplink (UL) to Downlink (DL)	16
Figure 24 - RAM Usage results for an example TN run on the Berlin Platform. The TN dynamically uses more memory during the execution of a test campaign.....	17
Figure 25 - CPU Resource Usage for a RAN deployed within an example TN deployed within the Berlin Platform. Different usages can be seen for different amounts of UEs and traffic directions	17
Figure 26 - Memory Resource Usage for a RAN, Core and DN deployed within an example TN deployed within the Berlin Platform. The usage of the core does not seem to change much with load, while the change is significant for the DN and RAN	18
Figure 27 - Timing Diagram of Scale Out Latency Measurements conducted at Athens and Berlin. Each point an arrow touches one of the vertical lines indicates a collected timestamp	19
Figure 28 - Graph showing the total Scale Out Latency measurements for all 200 iterations in Athens	19
Figure 29 - Graph showing total Scale Out Latency measurements for all 200 iterations in Berlin.....	20
Figure 30 - Logs of the Automated Gathering Script for Orchestration KPIs.....	21
Figure 31 - Results for Orchestration Measurements on the Malaga Platform as seen in the execution script ..	22
Figure 32 - Architecture of the LMF testing setup at the Athens Platform	23
Figure 33 - Lab setup for LMF testing at the Athens Platform.....	23
Figure 34 - LPP Message generated at the LMF during the estimation procedure for a Google Pixel 6 (Successful Test).....	24
Figure 35 - LPP Message generated at the LMF during the estimation procedure for a Huawei/OnePlus device (Unsuccessful Test)	24
Figure 36 - Overview of the outdoor measurements using a cellular mmWave system with a Reflector to enhance connectivity in a narrow, NLOS urban environment	25
Figure 37 - RSRP, SINR and Throughput Performance for the Malaga RIS Field Trial with and without reflectors A,B,C,D,E,F.....	26
Figure 38 - UE Transmission power performance for the Malaga RIS Field Trial	26
Figure 39 - Time series plot for SINR and Base Station Beam Index for Malaga RIS Field Trial with point of reflector B removal marked	27
Figure 40 - Time series plot for RSRP and Transmit power for Malaga RIS Field Trial with point of reflector B removal marked	27

Figure 41 - Coherence time and RIS Update Frequency plotted for a UE walking at 1,4m/s.....	28
Figure 42: Safety limits for different radiated power levels with varied EIRP values.....	30
Figure 43: Block diagram of EMF measurement setup.....	30
Figure 44: Measurement setups (a) Two meter OTA link and (b) 0.5-meter OTA link.....	31
Figure 45: Measured spectrum figures (a) Wide span spectrum and (b) 5G signal spectrum	31
Figure 46 - Diagram showing the structure of the 6G-SANDBOX Sensing Architecture present at UOULU	33
Figure 47 - The Meta Quest 2 and 3 headset devices, the primary internal sensing devices used in 6G-SANDBOX's Sensing lab at UOULU. They provide 6-DoF pose and velocity, video, audio and AR/VR Rendering. The Meta Quest 3 also offers environment depth sensing	34
Figure 48 - Physiotrack full inference pipeline.....	36
Figure 49 - The 6G-SANDBOX Sensing layer interface shown at two distinct points in time. The external camera view is shown in the main frame, with other inputs shown as inset panels	39
Figure 50: Setup to collect dataset and run experiments.....	41
Figure 51: Results of the N2 Layer Anomaly Detection testing, showing the differing performances from the 3 implemented models.	43
Figure 52: Experimentation setup for Anomaly Detection on the N6 Interface	45
Figure 53: Full pipeline for Anomaly Detection in the N6 Interface.....	47
Figure 54: Overall results of the Anomaly Detection testing at the N6 Interface	49
Figure 55: Graphical representations of the results of the Anomaly Detection testing at the N6 Interface	49
Figure 60 - Graph showing a comparison between Actual Energy Consumption vs Energy Consumption predicted by the AI/ML model in Athens AI/ML Experimentation	55
Figure 61 - Graph showing progression of Average Percentage Error as the amount of training data available to the AI/ML model was increased in Athens AI/ML Experimentation	55
Figure 62 - Heatmap of Phase 4 Interference Berlin Deltas	64
Figure 63 - Heatmap of Phase 4 Interference Athens Deltas	65
Figure 64 - Heatmap of Phase 4 CPU Allocation Berlin Deltas.....	67
Figure 65 - Heatmap of Phase 4 CPU Allocation Athens Deltas.....	68
Figure 66 - Heatmap of Phase 4 Memory Allocation Berlin Deltas	70
Figure 67 - Heatmap of Phase 4 Memory Allocation Athens Deltas.....	71
Figure 68 - Heatmap of all Phase 4 Deltas for both the Athens and Berlin Platforms, where Berlin is represented as "BI" and Athens as "A". The interference test cases are referred to as "I", the CPU allocation as "CPU" and the Memory allocation as "mem"	73
Figure 69 - Heatmap of Phase 4 Deltas for both Athens and Berlin summed together. Dark blue indicates an overall positive effect on both platforms, dark red indicates an overall negative, and light red/blue/white indicate either no or conflicting changes.....	73
Figure 70 - The new antenna shown in situ on the NCSR Campus	75
Figure 71 - Configuration parameters for the new antenna on the NCSR Campus	76
Figure 72 - UE Provided by KEYS for use with NEMO application testing	76
Figure 73 - Nemo Application export of the generated route based on the GPX data	77
Figure 74 - Location of the new antenna and coordinates of the different waypoints imported into the EXata scenario	77
Figure 75 - Final setup of the configuration of 5G SA deployment of the EXata Digital Twin and the path taken on the NCSR Campus	78
Figure 76 - Real-world measurements collected through the use of Nemo compared to the corresponding measurements generated by EXata	79
Figure 77 - Real-world measurements collected through the use of Nemo compared to the corresponding measurements generated by EXata. In this case, the antenna parameters are included in the simulation, as well as the radiation pattern files. The final graph shows the gap between the measured SINR values and the simulated ones.....	79
Figure 78 - PTP state machine of a slave device connected to one master.	81

Figure 79 - Overview of the PTP Digital Twin Toolchain	83
Figure 80 - Visual representation of Passta learning the first four observations of the first trace	83
Figure 81 - Visual representation of the transfer of SRTA to UPPAAL	84

List of TABLES

Table 1: Fixed Common parameters during 6G-SANDBOX Device Energy Efficiency measurements	14
Table 2: Fixed Signaling parameters during 6G-SANDBOX Device Energy Efficiency measurements.....	14
Table 3: Statistical values derived from Device Energy Efficiency Measurements	15
Table 4 - Average Scale Out Latency per Function/Time Period for Athens and Berlin	20
Table 5 - Results for Orchestration Measurements on the Malaga Platform	21
Table 6 - EIRP signal analysis based on measured 5G signal spectral power density of 28GHz Basestation	32
Table 7 - Power density based on measured transmission of 28GHZ Basestation	32
Table 8 - EIRP signal analysis based on measured 5G signal spectral power density.....	32
Table 9 - Coverage of the 6G-SANDBOX Sensing KPIs by the Meta Quest 2 and 3 acting independently	35
Table 10 - KPI Coverage of external video tracking system	37
Table 11 - Combined Sensing KPI coverage of the 6G-SANDBOX Sensing Layer (Meta Quest 2 / Meta Quest 3 + Physiotrack).....	38
Table 12 - Consolidated traffic profile for the autonomous drive use case [42] [43] [44] [45] [46] [47] [48]	46
Table 13 - Security Assurance Specification (SCAS) Testing for the Access and Mobility Management Function (AMF). Reference: TS33.512.	51
Table 14 - Security Assurance Specification (SCAS) Testing for the Network Repository Function (NRF). Reference: TS33.518.....	52
Table 15 - Security Assurance Specification (SCAS) Testing for the Unified Data Management (UDM). Reference: TS33.514.....	52
Table 16 - Results of all carried out SCAS tests on the AMF, NRF and UDM of the Open5GS Instance	52
Table 17 - Benchmarking Phase 2 Results for PDR KPI	56
Table 18 - Benchmarking Phase 2 Results for UDR KPI.....	56
Table 19 - Benchmarking Phase 2 Results for PLR KPI.....	56
Table 20 - Benchmarking Phase 2 Results for UPL KPI.....	57
Table 21 - Benchmarking Phase 2 Results for DDJ KPI.....	57
Table 22 - Benchmarking Phase 2 Results for CPL KPI	57
Table 23 - Benchmarking Phase 3 Results for PDR KPI	58
Table 24 - Benchmarking Phase 3 Results for UDR KPI.....	58
Table 25 - Benchmarking Phase 3 Results for PLR KPI	59
Table 26 - Benchmarking Phase 3 Results for UPL KPI.....	59
Table 27 - Benchmarking Phase 3 Results for DDJ KPI.....	59
Table 28 - Benchmarking Phase 3 Results for CPL KPI	60
Table 29 - Phase 4 Consistency Results for the Berlin Platform	61
Table 30 - Phase 4 Consistency Results for the Athens Platform	62
Table 31 - Phase 4 Interference Results for the Berlin Platform	63
Table 32 - Phase 4 Interference Results for the Athens Platform	64
Table 33 - CPU Core Allocation Legend for CPU Core Allocation Testing	66
Table 34 - Phase 4 CPU Allocation Results for the Berlin Platform	66
Table 35 - Phase 4 CPU Allocation Results for the Athens Platform	67
Table 36 - RAM Allocation Legend for Memory Allocation Testing.....	69
Table 37 - Phase 4 Memory Allocation Results for the Berlin Platform	69
Table 38 - Phase 4 Memory Allocation Results for the Athens Platform	70

Table 39 - Validation trace acceptance rates across different network scenarios.....	84
Table 40 - Satisfaction probabilities for MITL properties across ptp4l versions.....	85
Table 41 - Final Overview of measured KPIs within 6G-SANDBOX	86

List of ABBREVIATIONS

AD – Anomaly Detection	LMF – Location Management Function
AMF – Access and Mobility Management Function	LOS – Line Of Sight
AoA – Angle of Arrival	LPP – LTE Positioning Protocols
AoD – Angle of Departure	MAPE – Mean Absolute Percentage Error
AR – Augmented Reality	MIMO – Multiple-Input Multiple-Output
AUSF – Authentication Server Function	MU – Measurement Uncertainty
BI – Beam Index	NAS – Non-Access Stratum
BS/BTS – Base Station	NEF – Network Exposure Function
COTS – Commercial Off the Shelf	NFV – Network Function Virtualisation
COV – Coefficient of Variance	NGAP – Next-generation Application Protocol
DDoS – Distributed Denial of Service	NLOS – Non Line Of Sight
DN – Data Network	NRF – Network Repository Function
DoF – Degree of Freedom	NRPPa – NR Positioning Protocol A
DoS – Denial of Service	NSSF – Network Slice Selection Function
DT – Digital Twin	OTA – Over The Air
DUT – Device Under Test	PA – Power Analyzer
EIRP – Effective Isotropic Radiated Power	PCF – Policy Control Function
EMF – Electromagnetic Field	PRB – Physical Resource Blocks
FDT – Formal Digital Twin	PSD – Power Spectral Density
FL – Federated Learning	PTP – Precision Time Protocol
FSPL – Free Space Path Loss	RAN – Radio Access Network
HPE – HP Enterprise	RAPL – Running Average Power Limit
HRLLC – Hyper Reliable and Low Latency Communications	RIS – Reconfigurable Intelligent Surface
KPI – Key Performance Indicator	RRC – Radio Resource Control
KVM – Kernel-based Virtual Machine	RRU – Remote Radio Unit
	RSRP – Reference Signal Received Power

RSRQ – Reference Signal Received Quality	SUL – System Under Learning
RTN – Reference Trial Network	SUT – System Under Test
RTT – Round Trip Time	TDD – Time Division Duplex
SAR – Specific Absorption Rate	TDOA – Time Difference of Arrival
SBA – Software Based Architecture	TN – Trial Network
SCAS – Security Assurance Specifications	TNLCM – Trial Network life Cycle Manager
SCS – Subcarrier Spacing	UDM – Unified Data Management
SINR – Signal to Interference plus Noise Ratio	UE – User Equipment
SLAM – Simultaneous Localization and Mapping	UPF – User Plane Function
SMAPE – Symmetric Mean Absolute Percentage Error	URLLC – Ultra Reliable Low Latency Communications
SMC – Statistical Model Checking	VM – Virtual Machine
SMF – Session Management Function	VR – Virtual Reality
SRTA – Stochastic Real Timed Automata	XR – Extended Reality

List of OFFICIAL 6G-SANDBOX KPI CODES

ACA – Audio Capture	LOF – Localization Functionality
ADP – Anomaly Detection Precision	NEE – NFV Energy Efficiency
AMA – AI/ML Accuracy	OBD – Object Detection
AME – AI/ML Training Effectiveness Improvement	PDR – Peak Data Rate
ARR – Augmented Reality Rendering	PLR – Packet Loss Rate
ATC – Area Traffic Capacity	POX – Pose and Orientation of XR Device
COD – Connection Density	RAU – RAM Usage
CPL – Control Plane Latency	RCE – RIS Capacity Enhancement
CUU – CPU Usage	RSI – Real Scenario Images
DCA – Depth Capture	RSQ – Received Signal Quality
DDJ – Delay Deviation/Jitter	RUR – RIS Update Rate
DEE – Device Energy Efficiency	SEC – Security Conformance
ECU – Edge Computational Resource Usage	SEE – Self EMF Exposure
EGS – Egocentric Segmentation	SLT – Slice Provisioning Time
IEE – Inter EMF Exposure	SOL – Scale Out Latency

SPE – Spectral Efficiency**UPL** – User Plane Latency**SPT** – Service Provisioning Time**VOX** – Velocity of Device**UDR** – User Experienced Data Rate**VRR** – Virtual Reality Rendering

ABSTRACT

6G-SANDBOX as a project consists of the development of test and measurement platforms, and as such one of the key ways that the success of the project can be evaluated is its ability to measure a wide variety of KPIs with a variety of different methodologies. This Deliverable D5.5 acts as a companion to Deliverable D5.3 and together they form the full report of the measured Deliverables defined in Deliverable D2.4. Within this Deliverable, KPIs are separated by methodology, and for each group, a description of the measurement methodology is given as well as some exemplary results. The Platform Benchmarking Investigation, first documented in Deliverable D5.3, is also continued here, with the continued development of the methodology and examination of further interesting results both documented within. The Deliverable also contains some documentation regarding Digital Twin measurements that have been conducted within the auspices of 6G-SANDBOX. Overall, the Deliverable acts as the final word in terms of validation of the capability of 6G-SANDBOX to act as a modern and innovative testing platform and can be used by future experimenters as a guide for both the possibilities within the project platforms and the methodologies needed to reach the measurements of their choosing.

KEYWORDS

6G, Measurement, Validation, KPIs, Digital Twin

1 INTRODUCTION

6G-SANDBOX as a project focuses on the development of testing infrastructure, and as such a large proof of its success is its ability to conduct tests and produce results that are consistent, trustworthy and useful for the purpose of being able to draw conclusions from. Earlier in the project period as covered in Deliverable D5.3, a subset of 6G-SANDBOX Key Performance Indicators (KPI) were measured and recorded as a pilot set, during the early days of the integration of all of the project's technologies. Once those KPIs were verified and all 6G-SANDBOX platforms entered full maturity, the next step was to develop and document the measurement results and methodologies for the remaining set of predefined KPIs. This process is documented in this Deliverable.

1.1 NOTES ON GENERAL METHODOLOGY

Previous efforts in implementing 6G-SANDBOX testing methodologies were centralized – the same KPIs were measured on every platform, and the coordination of the testing everywhere was the responsibility of a single working group within the consortium. The testing efforts covered in this Deliverable are far more varied, often requiring features and technologies only available at one or two platforms, or with the application of one of the Open Caller technologies. As such, the implementations described herein were tackled in a distributed manner, usually assigned a specific platform or partner as the coordinator of the specific measurement effort. As a result of this, there is no unified methodology or structure as was present in the testing recorded in Deliverable D5.3. Therefore, each measurement section/sub-section in this Deliverable is a self-contained report and can be referenced individually without the need to consult or cross reference any of the others.

2 REPORTS OF KPI TESTING EFFORTS

In this Section, the methodology of measurement and example results of KPIs that were not reported in Deliverable D5.3 are shown. These remaining KPIs were grouped into “blocks” based on shared testing methodologies and responsibilities within the consortium. Each subsection contains the report of a given block, with the related KPIs being listed at the start.

The reports contained within these sections are intended to serve as examples of the capability of 6G-SANDBOX to measure the relevant KPI. The methodologies described are taken from one or more experiments and test campaigns done over the course of the 6G-SANDBOX lifecycle. As such, each section contains both a description of the specific example measurements done to measure the necessary KPI, as well as more generalized methodologies that can be used for future test campaigns.

The descriptions of all of the KPIs measured in this Section can be found in 6G-SANDBOX Deliverable D2.4

2.1 KPIs MEASURED USING UESIM

This Section contains KPIs that were to be tested on a trial network connected to a real Radio Access Network (RAN)/gNodeB with data being gathered through the use of the UESIM testing tool:

- Network Capacity (NCP)
- Area Traffic Capacity (**ARR**)
- Spectral Efficiency (**SPE**)
- Network Energy Efficiency (**NEE**)

UeSIM is a tool able to emulate hundreds of UEs, generally used for conformance testing and stress testing of base stations [1]. Within 6G-SANDBOX it can be shipped around to all platform locations if necessary, but by default (and for the testing described herein) it is located in the Malaga platform. For the testing displayed in this deliverable, it is connected to the base station under test (Device Under Test (DUT)) via RF cables, thus allowing for perfect radio conditions. The uplink signal can then be manipulated to emulate poor radio

conditions, and similarly the downlink can be manipulated to emulate packet drops and other dissatisfactory network conditions.

A related tool, RuSIM, emulates not just UEs on the radio interface, but it also includes the emulation of an ORAN Radio Unit (O-RU), connecting to an ORAN Distributed Unit (O-DU) on the ORAN Front-Haul (O-FH) [2].

Connecting these tools via conducted RF cables directly to the gNB, or even to the O-DU via O-FH, ensures perfect channel conditions and thus the possibility to measure the Network Capacity (NCP) in a controlled and repeatable way without fluctuating channel conditions. Furthermore, the possibility to scale up and down seamlessly the number of emulated UEs from 1 to several hundreds, allows for the measure of Connection Density (**COD**).

Example results reported here are measured with a system that is composed of a Keysight RuSIM, srsRAN Project v24.10 gNB, and a Keysight Coresim [3] [4]. Traffic bursts have duration of 60s with a 5s pause between bursts, with the following data rates set at application layer:

- Downlink (DN to UE): 1M, 2M, 5M, 10M, 20M, 50M, 100M, 200M, 500M, 1G
- Uplink (UE to DN): 1M, 2M, 5M, 10M, 20M, 50M, 100M

The application used is the User Data Generator (UDG) app from Keysight, natively integrated in both RuSIM and CoreSIM and designed for such tests. KPIs are collected from RuSIM every 1s. The cell is transmitting in band n78 (TDD) with 2x2 DL MIMO over a 100 MHz bandwidth with 30 KHz SCS. TDD frame structure was set to DDDDDDSUU, i.e., heavily DL-focused. A single slice with a single DNN with a single QoS flow (5QI=9) is set.

Additional information on this testbed and general testing setup can be found in other sources documenting its use for the generation of similar data [5] [6] [7].

2.1.1 NCP, ATC, SPE

Network capacity can be measured by emulating a single UE under ideal conditions. The test case can be performed as follows:

- Register a single UE
- Request a PDU session
- Perform DL traffic bursts with an increasing amount of data rate (e.g., 1M, 2M, 5M, 10M, 20M, 50M, 100M, 200M, etc.)
- Perform UL traffic bursts with an increasing amount of data rate
- Release the PDU session
- Deregister the UE

Traffic bursts can be tuned to the expected capacity of the cell, always transmitting at least one burst with higher data rate than the cell's capacity. The PHY throughput will hit an upper bound, i.e., the real NCP.

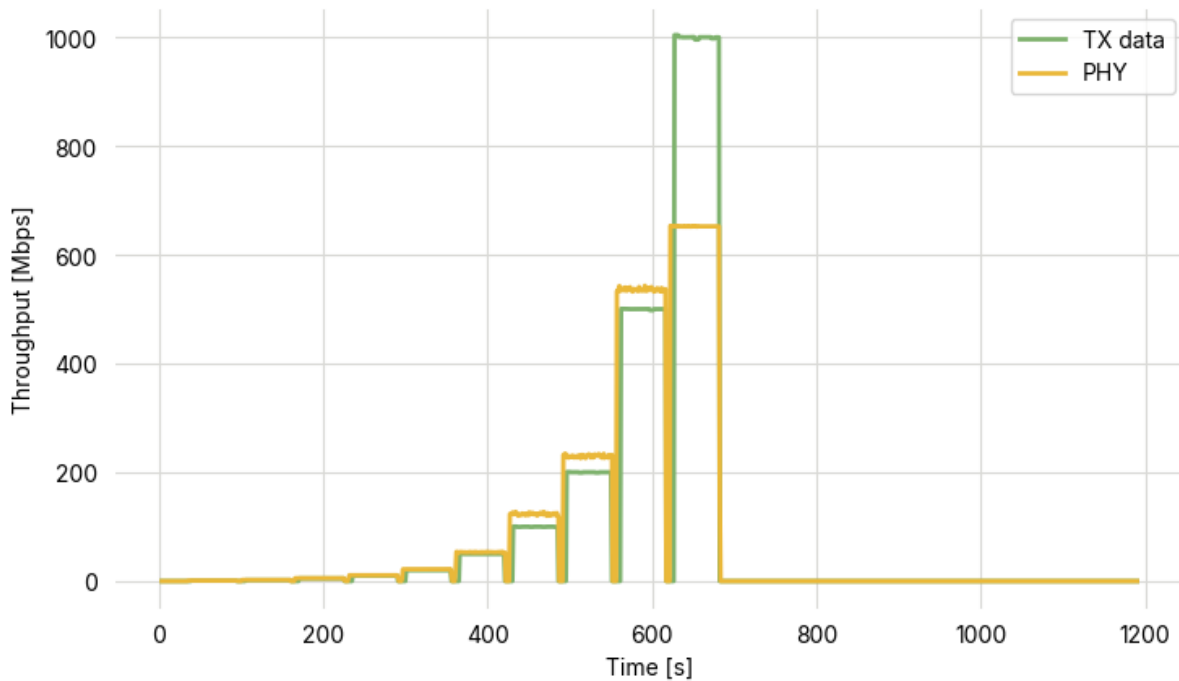


Figure 1: Downlink data bursts in the example scenario. PHY throughput maxes out at 653.7 Mbps, indicating the DL NCP for this cell.

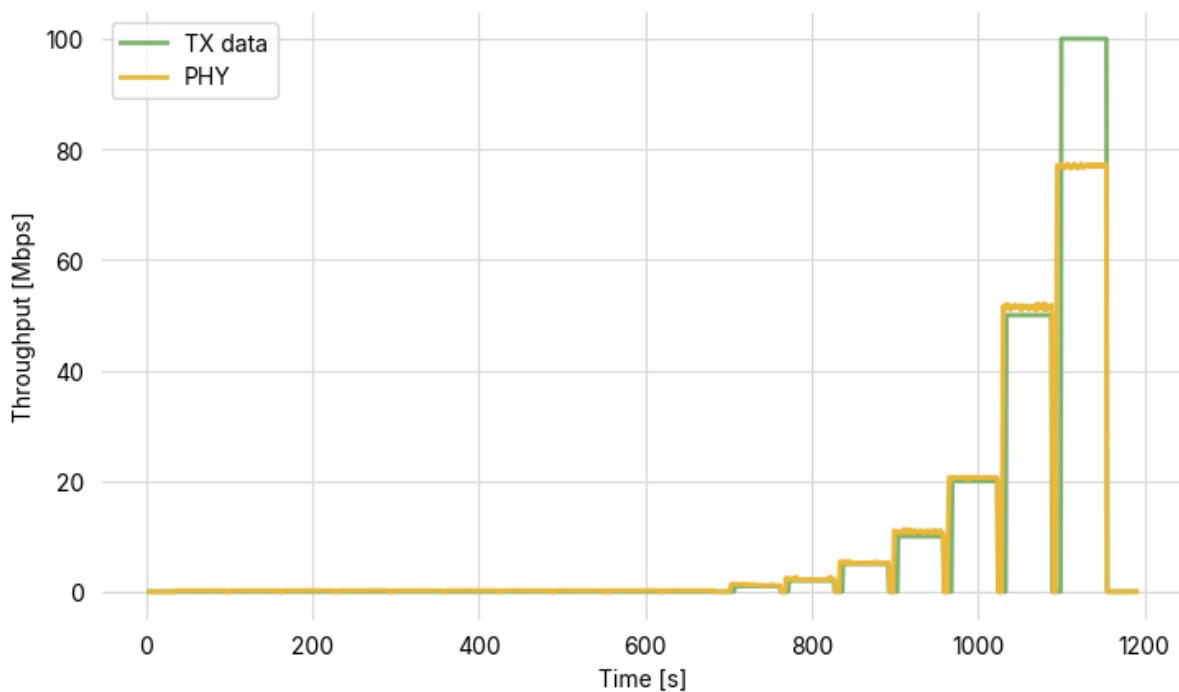


Figure 2: Uplink data bursts in the example scenario. PHY throughput maxes out at 77.9 Mbps, indicating the UL NCP for this cell.

In Figure 1 and Figure 2, both the App-level transmitted throughput and the PHY throughput effectively transmitted are shown respectively for DL and UL. It can be observed that in DL, the PHY throughput, which includes all layers' overhead and possible retransmissions, hits an upper bound at 653.7 Mbps for the 1G data burst. This is in line with the expected cell capacity (roughly 670M). Similarly, in UL the maximum PHY throughput reached is equal to 77.4 Mbps.

From NCP, **ATC** can be derived by dividing NCP by the target coverage area. Similarly, **SPE** can be computed by dividing the NCP by the cell bandwidth, in this case 6.537 bps/Hz in DL and 0.774 bps/Hz in UL. The big difference

between the two values comes from mostly two factors: the TDD frame structure is heavily DL-focused, and MIMO is rank 2 in DL while only rank 1 in UL.

2.1.2 NEE

By adding some means of direct power measurement (such as a Keysight AC Power Analyzer) to the testbed, power consumption and thus **NEE** can be measured at the same time as NCP [8].

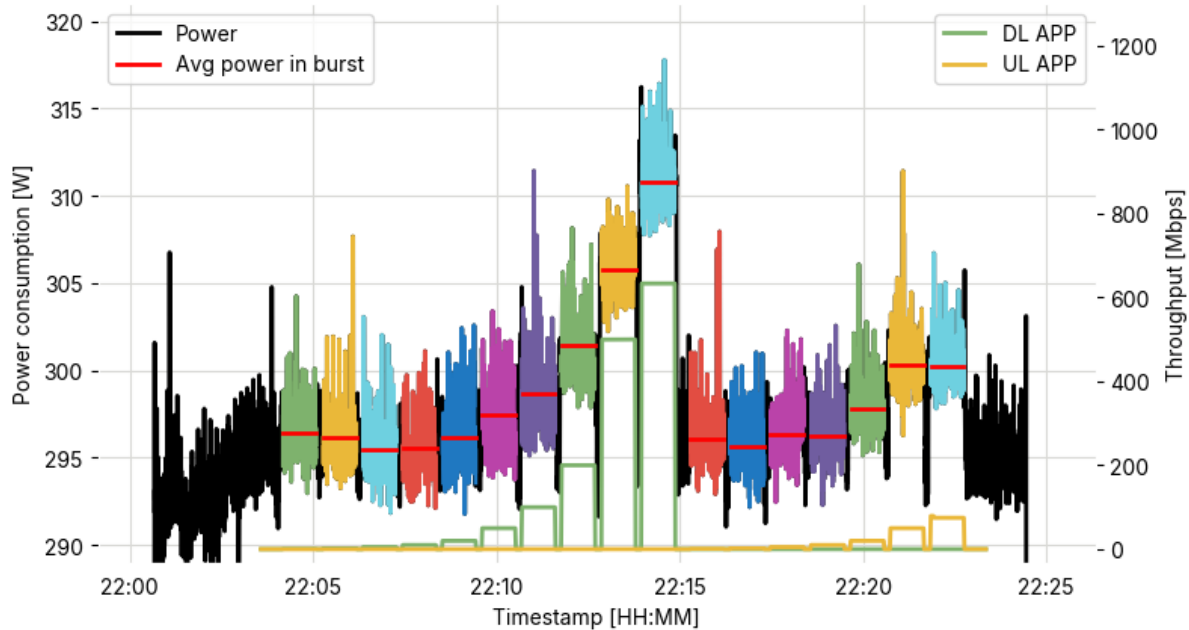


Figure 3: Power consumption of the server running the O-CU/O-DU while performing data bursts. A single UE is emulated in this case.

Figure 3 shows the power consumption over time of the server (Supermicro SYS-2029U-E1CR4T) running the O-CU/O-DU software. Other than the Rocky Linux OS, this is the only software running on this server. Although the power reading is quite noisy, a clear correlation with data rate can be observed. The same can be repeated in different conditions, e.g., with more UEs, with different cell bandwidth, with different MIMO ranks.

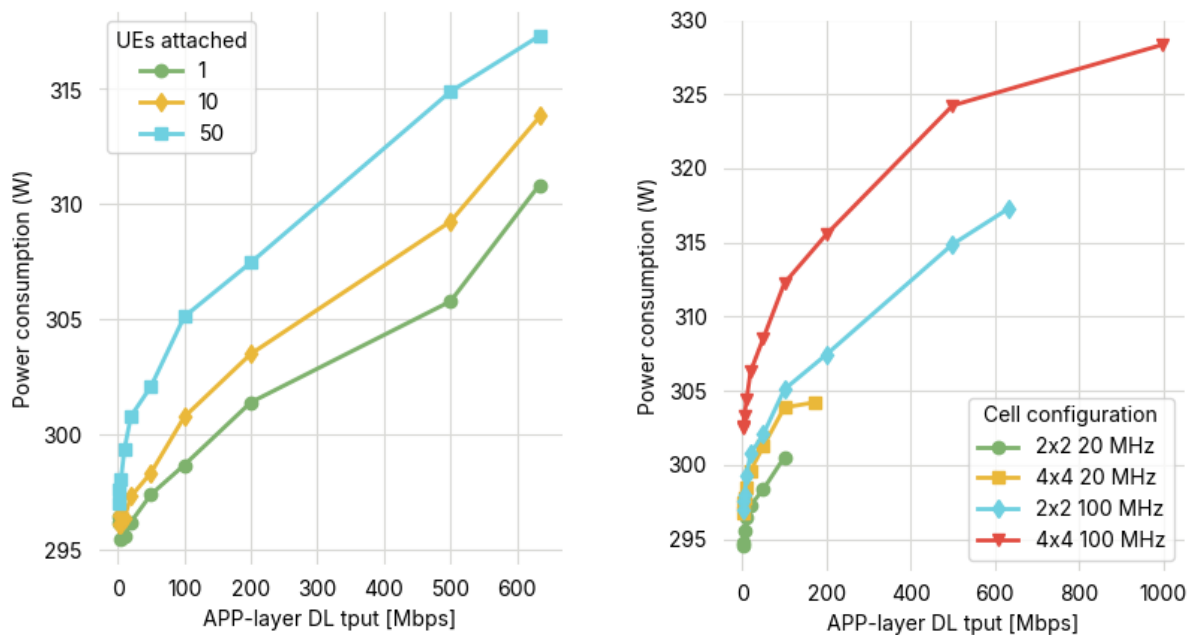


Figure 4: Power measurements in different conditions during downlink bursts.

Figure 4 shows average power consumption of the server in different conditions. The left figure shows the impact that multiple UEs have on power consumption. Even though the aggregated cell throughput is the same, having more users consumes more power due to the additional context, buffers, handling that each individual user needs from the cell. The right figure shows how different Multiple-Input and Multiple-Output (MIMO) ranks and cell bandwidths affect power consumption as well as the maximum achievable throughput.

2.2 KPIs RELATING TO SIGNAL QUALITY

This Section contains KPIs relating to Signal Quality:

- Received Signal Quality (**RSQ**)

As part of the Athens platform Digital Twin activities, an extensive measurement campaign to characterize the signal quality of the NCSR and OTE Academy campuses' antennas has been performed. Part of this measurement campaign consisted of the development of the methodology for measuring the **RSQ**, and as such it is documented here.

2.2.1 MEASUREMENTS AT NCSR

The first antenna assessed was the Ericsson Remote Radio Unit (RRU)/RAN, which is ultimately connected to the core of the HP Enterprise (HPE) system on the OTE side. This antenna is mounted on the roof of the IIT building at the NCSR campus and is illustrated with the red dot on Figure 5.



Figure 5: The two gNBs (Amarisoft and Ericsson) located at the NCSR campus, marked in yellow and red respectively

The measured metrics include Reference Signal Received Power (RSRP), Reference Signal Received Quality (RSRQ), and Signal to Interference plus Noise Ratio (SINR). All of these were measured with the use of the Keysight Nemo Handy field measurement tool, and when considered together they constitute the **RSQ** KPI [9]. As shown in Figure 6 below, the measured path indicates that the signal maintains an acceptable power, ranging from -80 to -110 dBm throughout the measurement campaign. The RSRQ values are mostly between -10 and -15 dB, while SINR values predominantly range from 10 to 20 dB, reflecting generally good signal quality along the measured path. Only at the farthest points from the antenna do SINR values drop below 10 dB as the distance increases. Figure 7 and Figure 8 illustrate the trace acquired in the proximity of the Ericsson gNB for RSRP and SINR respectively.

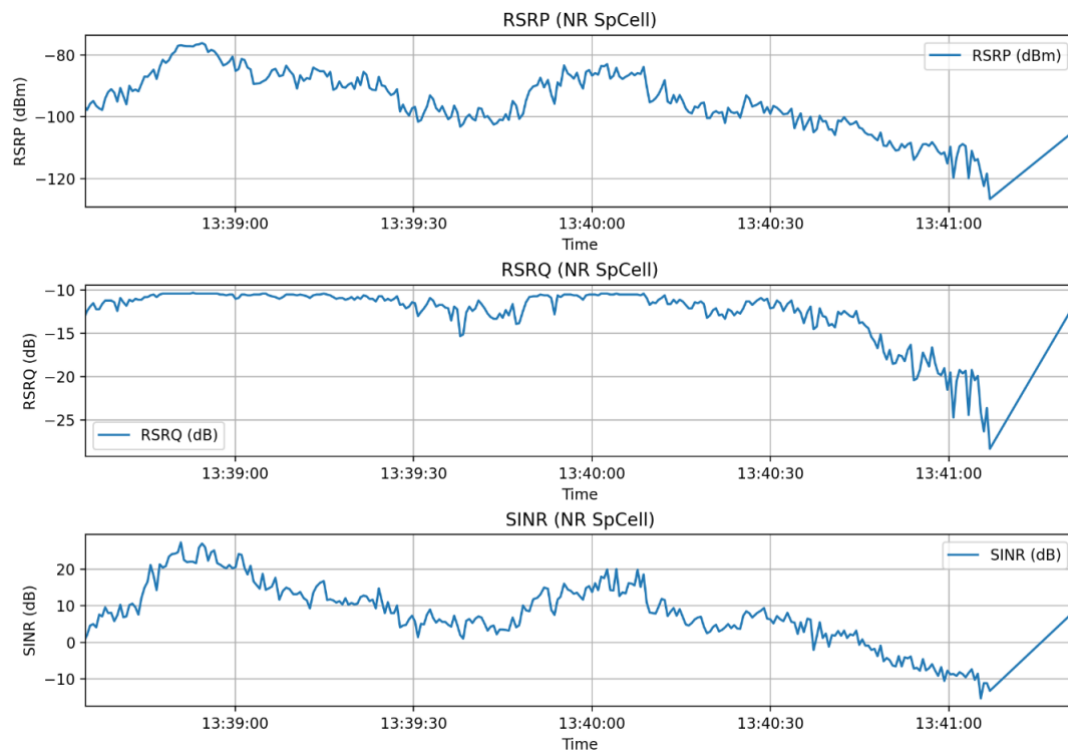


Figure 6: DL signal quality metrics acquired at the Athens platform (Ericsson RAN) using Keysight's Nemo Handy Field Measurement Tool

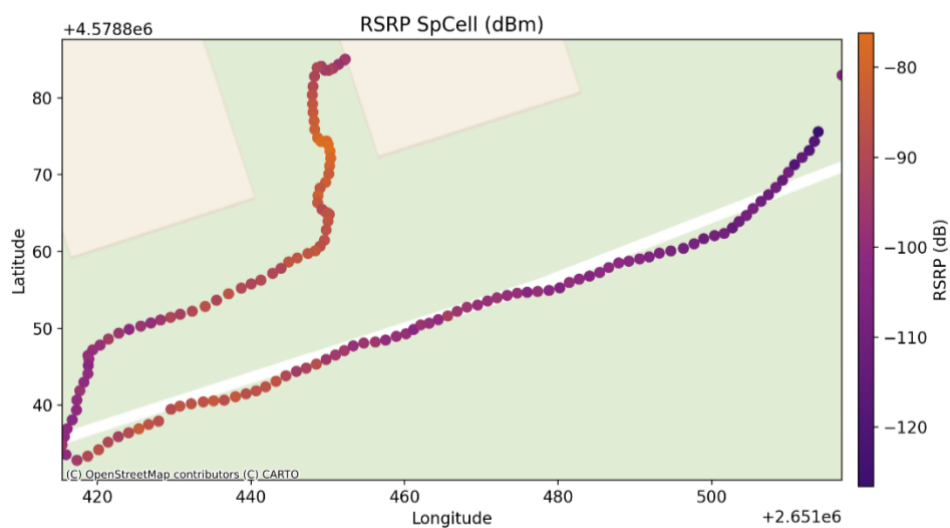


Figure 7: DL RSRP trace acquired in the proximity of the Ericsson gNB at the NCSR Campus

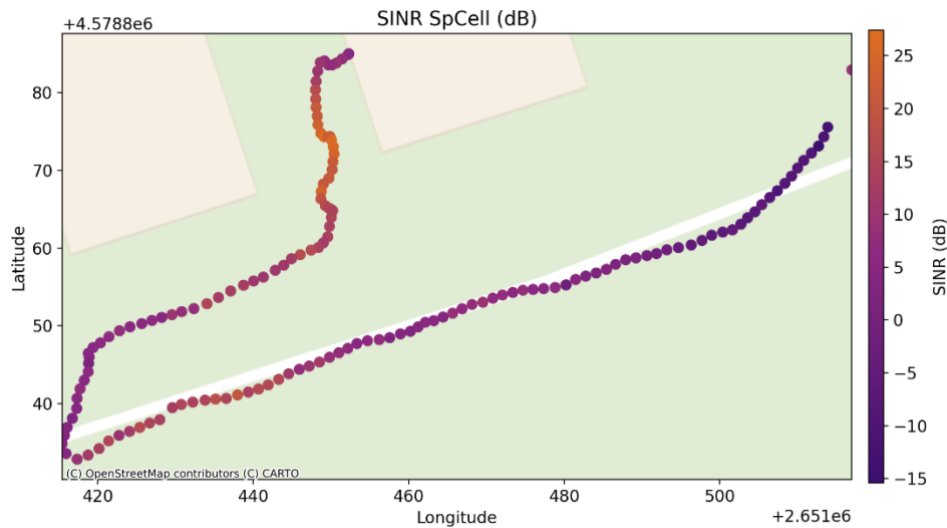


Figure 8: DL SINR trace acquired in the proximity of the Ericsson gNB at the NCSR Campus

The second antenna that was characterized is the Octo port multi-band antenna, illustrated in Figure 5 with the yellow dot, which is 1.4m long with 65° azimuth beamwidth and supports the 1427-2690 MHz and 3100-4200 MHz frequencies (4 wide band ports) [10]. In line with the measurements performed for the Ericsson antenna, the collected metrics include RSRP and SINR along an outdoor path surrounding the building, to determine the boundaries where the signal remains strong. The RSRP and SINR measurements of the specific antenna, reported in Figure 10, show moderate quality overall. While exhibiting fluctuations along the measurement path, the RSRP never experiences deep fading (RSRP below -110 dBm) for extended period of times, and averages around acceptable levels (-90 to -100 dBm). Similarly, the SINR occasionally drops to negative values, but rebounds to positive values quite quickly and is generally in the range [0-13] dB.



Figure 9: Amarisoft RAN equipment for signal quality metrics acquisition at NCSR Campus.

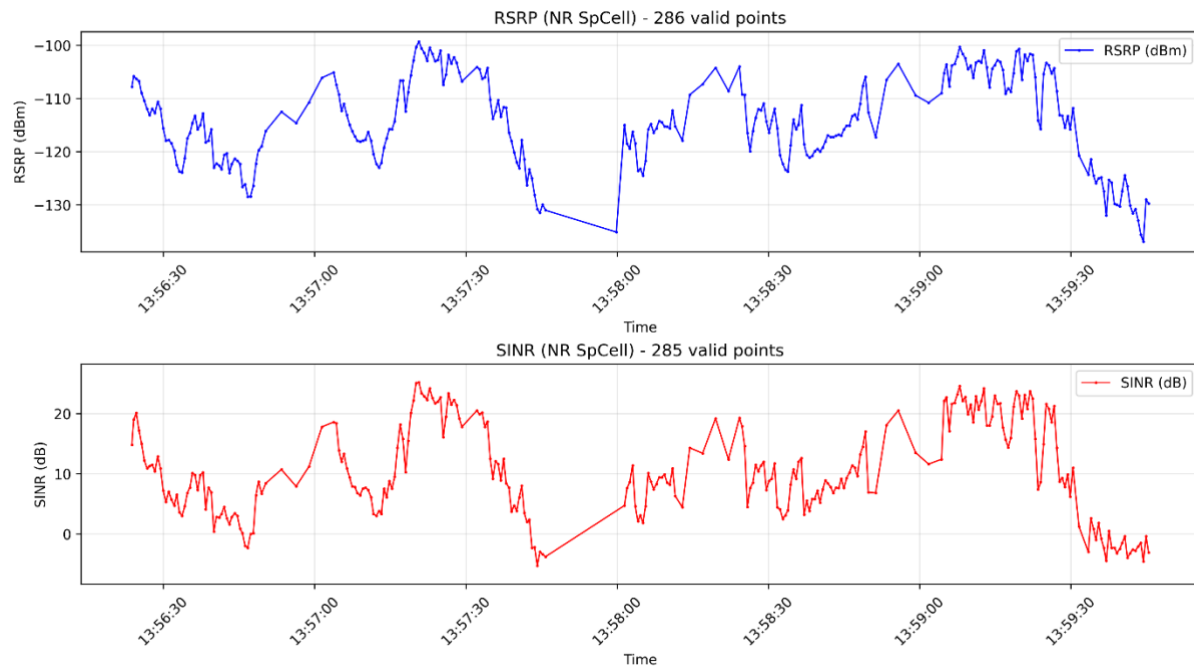


Figure 10: UL signal quality metrics acquired at the Athens platform (Amarisoft RAN) using Keysight's Nemo Handy Field Measurement Tool

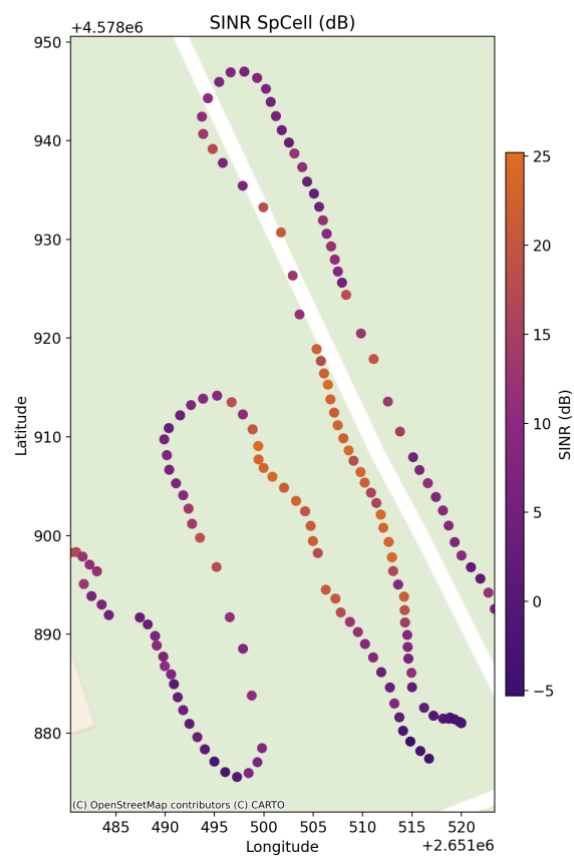


Figure 11: UL SINR trace acquired in the proximity of the Amarisoft gNB at the NCSR Campus

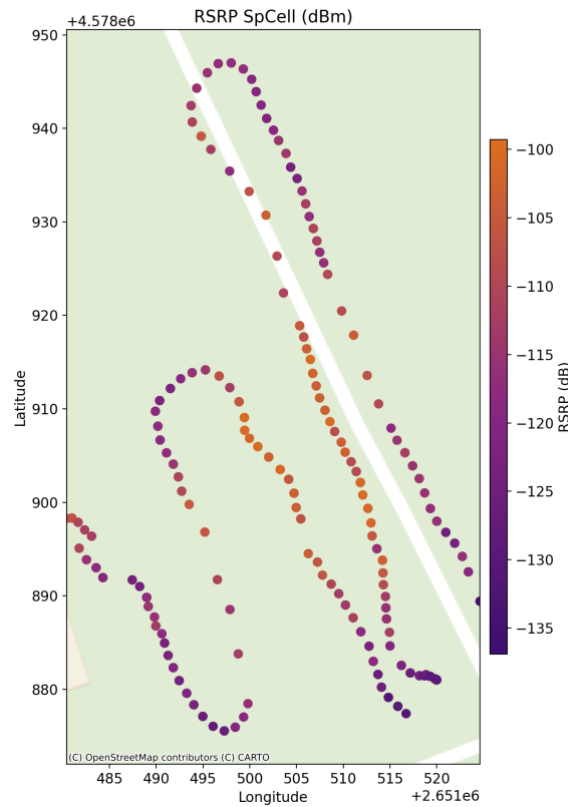


Figure 12: UL RSRP trace acquired in the proximity of the Amarisoft gNB at the NCSR Campus

2.2.2 MEASUREMENTS AT OTE

As part of the Athens platform Digital Twin activities, the signal quality of the Ericsson antenna (illustrated in **Error! Reference source not found.**) supporting the outdoor experiments at the OTE Academy campus were also evaluated [11].



Figure 13 - Ericsson Radio 4408 n78R equipment at OTE Academy Campus

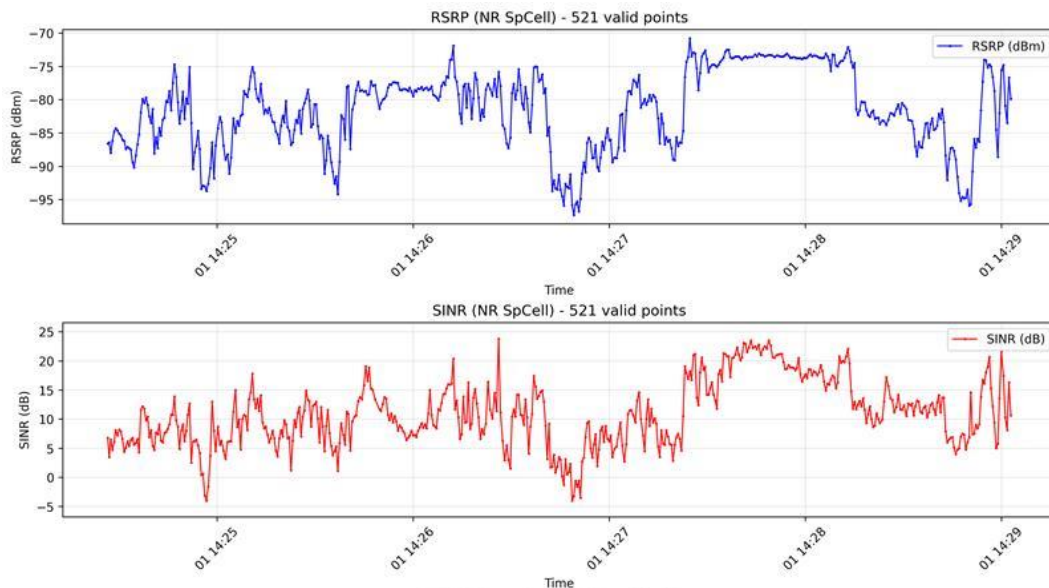


Figure 14: Ericsson signal quality metrics acquired at OTE Academy Campus using Keysight's Nemo Handy Field Measurement Tool

As depicted in Figure 14, the RSRP values typically range between approximately -95 dBm and -72 dBm, with most samples concentrated around -85 to -75 dBm, indicating very good signal strength for the outdoor Ericsson deployment. The RSRP measurement show brief drops to -95 dBm, suggesting temporary shadowing and/or partial blockages, but overall, the signal remains stable. The SINR values fluctuate between roughly -2 dB and 23 dB, with most measurements falling in the 5 – 20 dB range. This corresponds to moderate to very good radio signal quality. Positive outliers exhibit an SINR exceeding 15 – 20 dB, indicating excellent link conditions, while negative outliers show occasional dips around 0 dB. The very same trends can be seen in the signal quality traces projected on the map for in Figure 15 and Figure 16 for RSRP and SINR, respectively.

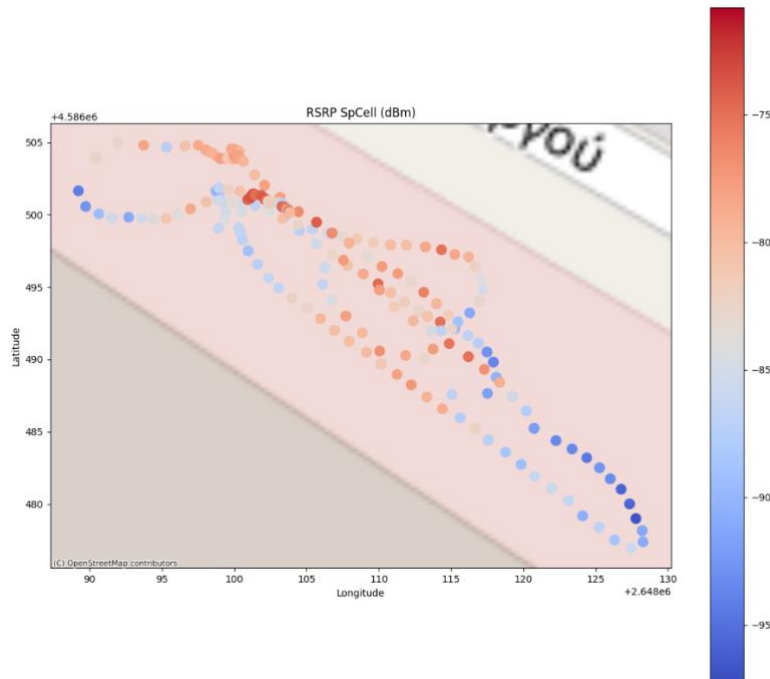


Figure 15: RSRP trace acquired in the proximity of the Ericsson gNB at the OTE Campus

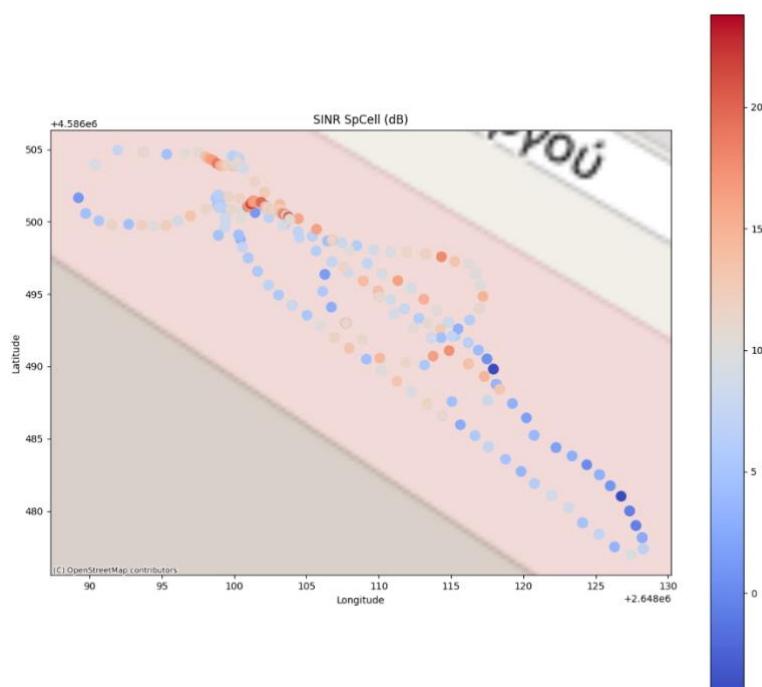


Figure 16: SINR trace acquired in the proximity of the Ericsson gNB at the OTE Campus

2.3 KPIs RELATING TO NFV ENERGY EFFICIENCY

This Section contains KPIs relating to the Energy consumption and efficiency of virtualised applications:

- Network Function Virtualisation (NFV) Energy Efficiency (**NEE**)

6G-SANDBOX computes the energy usage of a VM or individual process through an estimation of the instant power consumption of the infrastructure. This instant power consumption is calculated using Running Average Power Limit (RAPL) CPU extensions, that allow a fair approximation for the instant power usage on the processor and memory controller and can be used to interpolate the resource usage of processes running on the CPU [9].

The provided power measurements are estimations from specialized processor extensions, limited to the CPU, integrated memory controller, and integrated graphics (iGPU, when applicable) domains. They do not represent total system power, as they exclude significant components like the chipset (PCH), discrete GPUs (dGPUs), storage controllers, HDDs, and SSDs. For a true, comprehensive, and accurate total server power measurement, external calibrated equipment (e.g., a power meter) is necessary.

Nevertheless, despite not providing absolute total power, the **NEE** measurement methods within 6G-SANDBOX make for valuable dynamic indicators for monitoring power trends and relative efficiency within the monitored processor and its immediate subsystems. As every Kernel-based Virtual Machine (KVM) running on the 6G-SANDBOX OpenNebula hypervisor is treated as a process, its power consumption can be interpolated by making use of the RAPL extensions. Moreover, on Kubernetes every pod is seen as a process to the operating system, so a ponderation of its resource usage on the VM power makes a good approximation for its power usage. Three specific types of power consumption can therefore be measured in this way, listed here in order of descending scope:

- Hypervisor instant power consumption in microwatts (μ W)
- Instant power consumption per VM in microwatts (μ W)
- Synchronous VM process consumption in microwatts (μ W)

The software Scaphandre is utilized to calculate power consumption for both the hypervisor and individual Virtual Machines (VMs) which run as processes on the hypervisor. It operates by correlating the reported CPU energy consumption with the resource usage of a process over a specific period [10]. 6G-SANDBOX collects the metrics generated by Scaphandre through the use of Prometheus [11]. OpenNebula directly integrates a Prometheus exporter that allows for the outputting of these metrics within OpenNebula directly.

To estimate the power usage per process, the OpenNebula Prometheus exporter executes a small, statically linked binary executable on each VM. This binary synchronously reports the CPU cycles used by each process within the VM. The Prometheus exporter then uses this process-level resource usage to interpolate the overall VM power consumption. OpenNebula's Prometheus integration includes a Grafana dashboard to visualize the collected metrics, displaying power consumption per host and per Virtual Machine. The usage of pods and VM processes can also be displayed.

This energy usage result can then be extrapolated to the overall energy efficiency by dividing the number of processed bytes for a given TN/VM/Process by the energy used during the period covered.

2.3.1 RESULTS

Example results for **NEE** were collected on the Berlin platform, but this capability exists on every platform. Figure 17, Figure 18, and Figure 19 show the energy usage of an example TN in increasing granularity, while Figure 20 shows the **NEE** measurement over a period of 15 seconds for the core of said TN.

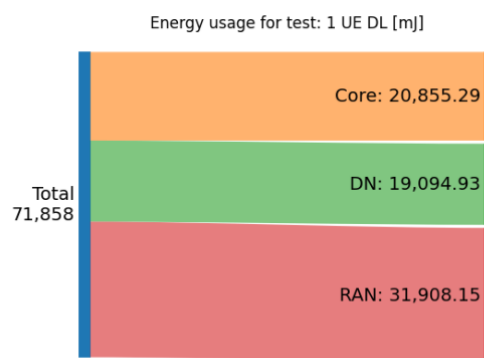


Figure 17 - Energy usage in mJ of the 3 main components of an example TN deployed on the Berlin Platform

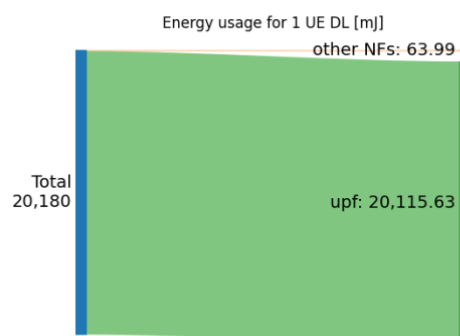


Figure 18 - Energy usage in mJ of the Core (Open5GS) of an example TN deployed on the Berlin Platform. The UPF consumes such a large amount that all other processes have been grouped for better visibility

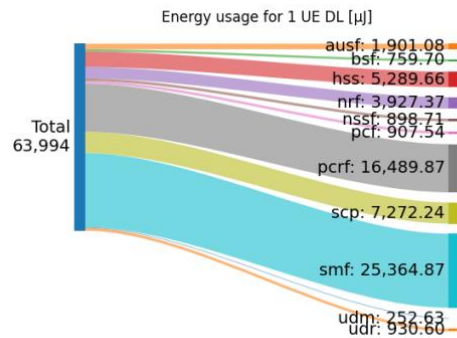


Figure 19 – Non-UPF Energy usage in mJ of the Core (Open5GS) of an example TN deployed on the Berlin Platform. This is the same proportion marked “other NFs” in Figure 18

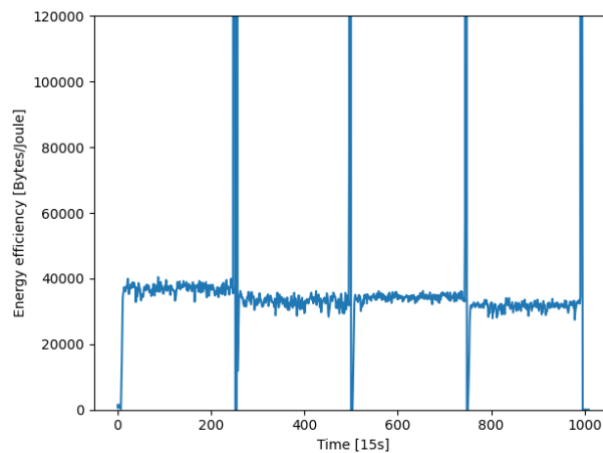


Figure 20 - Energy Efficiency for the Core (Open5GS) of an example TN deployed on the Berlin Platform. The average value in this case is 32804,81 Bytes per Joule.

2.4 KPIS RELATING TO DEVICE ENERGY EFFICIENCY

This Section contains KPIS relating to the energy consumption and efficiency of physical UEs:

- Device Energy Efficiency (**DEE**)

To accurately characterize the energy efficiency of a Commercial-Off-The-Shelf (COTS) 5G device, 6G-SANDBOX employs a controllable and flexible lab environment. This testbed, along with its associated measurement methodology, was originally developed within WP3 of the 6G-VERSUS project [12] and is physically hosted in the UMA-Keysight 6G-SANDBOX lab. The testbed enables fine-grained network configuration, as well as network-level and energy-level KPI collection. It is reused in 6G-SANDBOX to measure the device's energy efficiency.

2.4.1 TESTBED SETUP

The details of the testbed architecture and the measurement methodologies can be found in the 6G-VERSUS deliverables [12]. For completeness, a brief description of the testbed is described here. The physical setup consists of four main components: the UXM [13], the COTS phone (User Equipment (UE), Device under Test (DUT)), a USB charger, and a DC Power Analyzer (PA) [14]. The UXM emulates a 5G NR cell and connects to the UE via RF cables. The UE is configured in the bypass charging mode, so that the power supplied by the charger aligns with its real-time consumption. This ensures accurate power measurements by the PA. Figure 21 shows the diagram of the testbed setup.

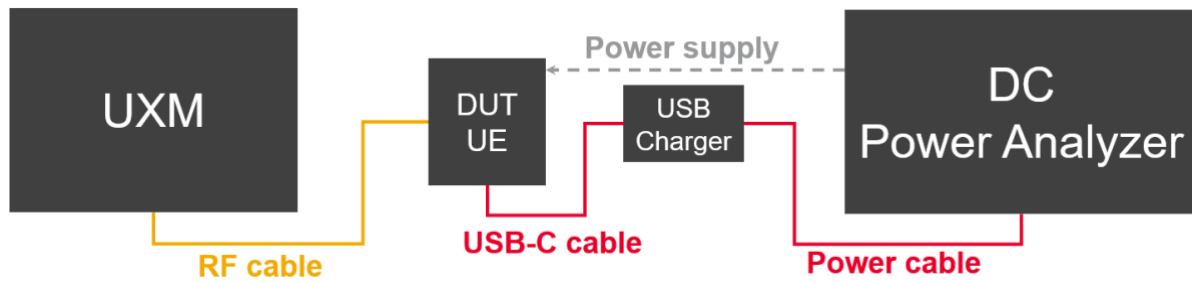


Figure 21 – Diagram of the testbed used for Device Energy Efficiency measurements

2.4.2 RESULTS AND ANALYSIS

Per the definition of the **DEE** KPI, there are two representative scenarios that need to be measured: idle and loaded. In the idle state, the phone is not transmitting or receiving any user data and remains in the RRC_IDLE state. This scenario corresponds to the low energy consumption aspect defined previously in the project. On the contrary, in the loaded state, the UE is fully loaded with user traffic in both Downlink (DL) and Uplink (UL). All available Physical Resource Blocks (PRB) are allocated to the phone under test in every slot, achieving maximum possible throughput in both directions. This configuration allows assessment of data transmission efficiency in high-volume data transfer cases.

Except for the Radio Resource Control (RRC) state and traffic load, all other configurations of the testbed remain constant for the duration of **DEE** testing. The details of these fixed settings are provided in Table 1 and Table 2.

Table 1: Fixed Common parameters during 6G-SANDBOX Device Energy Efficiency measurements

Parameter	Value
Radio access technology	NR
Duplex mode	TDD
TDD Configuration	DDDDDDDSUU
Nominal subcarrier spacing	30 kHz
Channel bandwidth	20 MHz
Number of PRBs	51
Central carrier frequency	3549.9 MHz (n78 mid)
Number of spatial/antenna streams	4x4 MIMO
Power control	Close Loop Mode: 0dBm
UE received power	-15 dBm / BW
Signal to Noise Ratio (SNR)	Static (no noise), by default
MCS index Table	Table 5.1.3.1-2-2 in TS 38.214 v18.5.0

Table 2: Fixed Signaling parameters during 6G-SANDBOX Device Energy Efficiency measurements

Parameter	Value
Paging Cycle	320 ms
SS/PBCH Periodicity	20 ms
CSI RS Period	10 ms

Periodic SR resource (SR period)	True (10ms)
Long DRX cycle	Off
Short DRX cycle	Off

The measurement results are presented in Figure 22, and the corresponding statistical values derived from this figure are summarized in Table 3. The figure clearly shows that the UE consumes significantly more energy in the loaded state (plotted in yellow), while the consumption in the idle state remains low (plotted in green). This behavior aligns with expectations, as a heavy traffic load implies much more workload in the phone and thus increased power draw. Also, in the fully loaded state, the measured power shows a clear 5 ms periodicity. This period corresponds to the configured Structured Cabling System (SCS) and Time Division Duplex (TDD) pattern, as each slot duration is 0.5 ms and a TDD cycle includes 10 slots, i.e., resulting in a total period of 5 ms and matching the observed pattern in power measurements. Moreover, within each cycle, the power peaks in the fully loaded measurement align with the UL transmission, which occupies only 2 of the 10 slots. These UL bursts therefore appear as narrow power peaks in the plot. The existence of these peaks indicates that, in the loaded state, UL transmission requires more power than DL, as the device needs to transmit RF signals in UL. On the contrary, in the idle state, power consumption of the device remains at a very low and stable level.

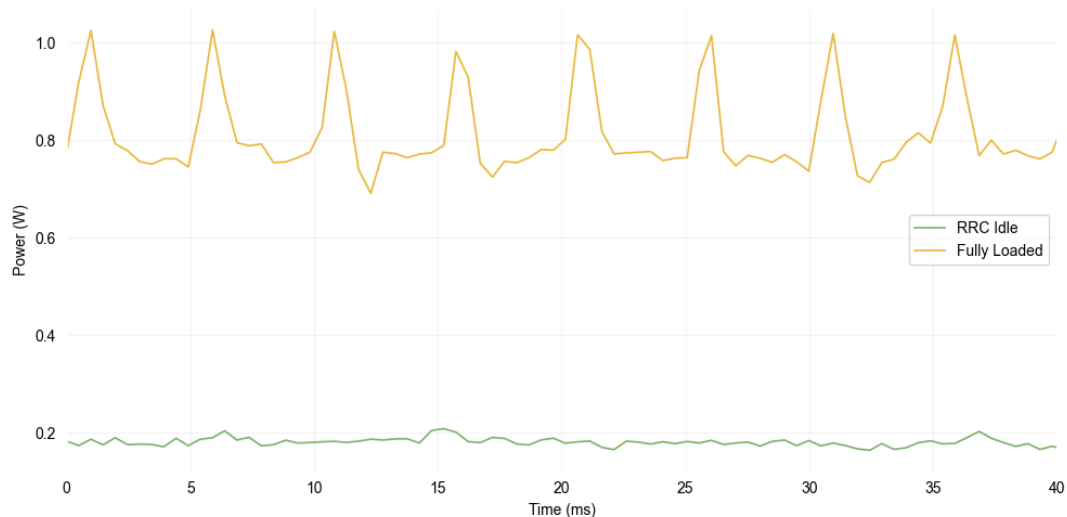


Figure 22: Graph showing Device Energy Efficiency Measurements over time for both the Idle and Fully Loaded states

Table 3: Statistical values derived from Device Energy Efficiency Measurements

Test scenario	Average power [W]	Maximum power [W]	Minimum power [W]
Fully Loaded	0.815	1.026	0.691
RRC Idle	0.181	0.208	0.164

These experiments demonstrate the capability of the 6G-SANDBOX Malaga platform at conducting accurate energy profiling on a device. These measurements can easily be replicated on other devices, and/or on devices connected to Trial Networks of interest, allowing for comprehensive insights on device energy usage within 6G-SANDBOX.

2.5 KPIs RELATING TO VM COMPUTE MEASUREMENTS

This Section contains KPIs relating to the measurement of resource usage and allocation of underlying virtualisation architecture:

- RAM Usage (**RAU**)
- CPU Usage (**CUU**)
- Edge Computational Resource Usage (**ECU**)
- Scale Out Latency (**SOL**)

Within 6G-SANDBOX, these metrics are gathered from the monitorization data that OpenNebula gathers over the course of standard usage. Similarly to Section 2.3, these are then exported out through the use of a custom Prometheus collector [11].

2.5.1 CUU

This measurement quantifies the total number of physical, or "real," CPU cycles that have been actively consumed by each TN. This refers to the actual processing time utilized by the underlying physical CPU cores to execute the operations and instructions of Virtual Machines (VM) within a specific Trial Network (TN). It is a direct measure of the load the VMs place on the host hardware. This measurement goes beyond merely tracking allocated virtual CPUs (vCPUs) and instead focuses on the actual work done, offering an insight into the operational efficiency and resource demand of each virtualized workload.

CUU metrics can be viewed in real time via Grafana dashboards while a TN is running or collated at the end for further analysis. By default, the displayed unit is CPU seconds, with an example result shown in Figure 23.

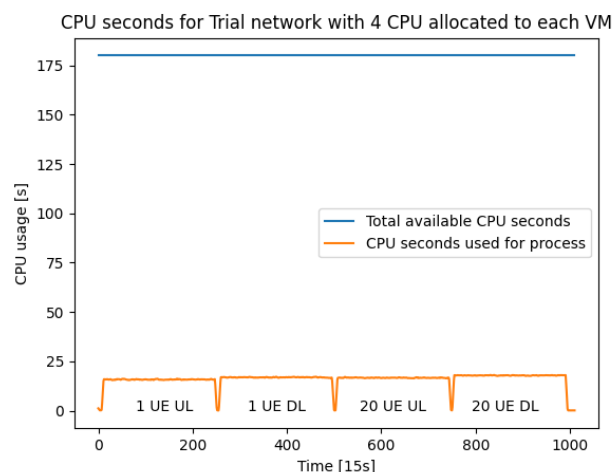


Figure 23 - CPU Usage results for an example TN running multiple different tests on the Berlin Platform. The CPU usage does not appear to change significantly when the amount of UEs present in the test is changed, or when the direction is changed from Uplink (UL) to Downlink (DL)

2.5.2 RAU

This metric provides the measurement of the cumulative RAM capacity consumed by each individual VM within the environment. The metric shows the total amount of physical host RAM that has been allocated to and is currently being actively used by the VM instance. This includes memory used by the operating system, applications, and system processes running within the guest environment.

RAU, much like **CUU**, can be viewed in real time via Grafana Dashboards while a TN is running, or collated at the end for further analysis. The minimum required amount is usually consistent, while the amount needed during a test execution differs depending on requirements. An example **RAU** result for a TN before and after a test execution can be seen in Figure 24

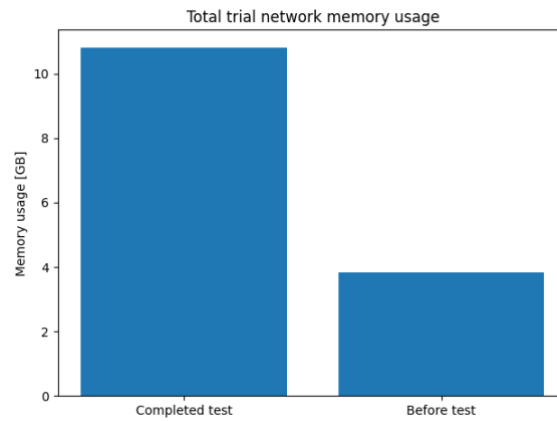


Figure 24 - RAM Usage results for an example TN run on the Berlin Platform. The TN dynamically uses more memory during the execution of a test campaign

2.5.3 ECU

This metric acts as a more granular version of **CUU** and **RAU**, tracking how individual VMs representing network edge devices are utilizing underlying physical CPU resources. For Edge CPU Usage, the metric reports real CPU time consumed by each VM, while for memory it reports both the minimum and the maximum used memory. A ponderation process is employed to present a clearer picture of internal resource competition. The results can be taken down even further in granularity, by examining individual processes within a given VM, if such a level of detail proves necessary.

Example results for both CPU **ECU** and Memory **ECU** can be seen in Figure 25 and Figure 26.

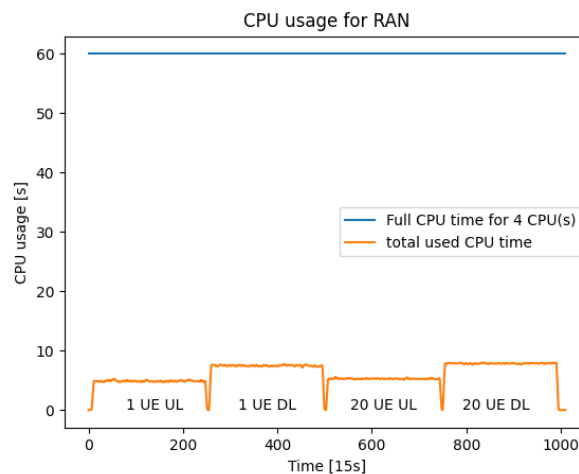


Figure 25 - CPU Resource Usage for a RAN deployed within an example TN deployed within the Berlin Platform. Different usages can be seen for different amounts of UEs and traffic directions

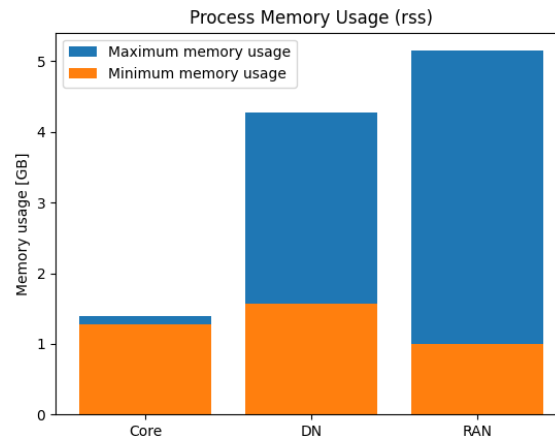


Figure 26 - Memory Resource Usage for a RAN, Core and DN deployed within an example TN deployed within the Berlin Platform. The usage of the core does not seem to change much with load, while the change is significant for the DN and RAN

2.5.4 SOL

This metric measures the system responsiveness to infrastructure necessities quantifying the total time required for a system to successfully complete an automatic or semi-automatic infrastructure rescaling operation in response to detected changes in workload or resource demand. This metric essentially measures the system's agility and responsiveness—the delay between identifying the necessity for additional resources (e.g., due to a spike in user traffic or a threshold breach in CPU utilization) and the point at which the newly provisioned resources are fully operational and contributing to the service capacity.

SOL was measured on both the Athens and Berlin platforms, by subjecting them to requests for CPU resizing from 6 to 8 cores and then back to 6 cores 100 times each, and measuring the time taken by doing log-based time stamp analysis. The Reference TN was used for this task, and it was hosted on a server isolated from other users. Testing and data collection was automated through the use of OpenTAP, and X shows a timing diagram of the entire process.

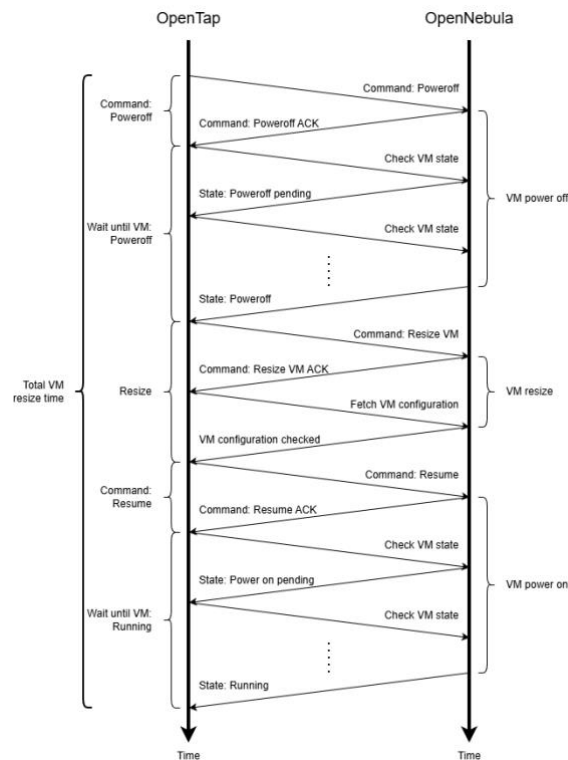


Figure 27 - Timing Diagram of Scale Out Latency Measurements conducted at Athens and Berlin. Each point an arrow touches one of the vertical lines indicates a collected timestamp

Figure 28 shows the results for the total resize time across all iterations for Athens, while Figure 29 shows the same for Berlin. The Berlin deployment appears to have a longer **SOL**, with an average of 50,93 seconds for the whole process, compared to 24,99 seconds for Athens. However, Berlin appears to also be more consistent, with Athens showing multiple iterations that take significantly longer than the average. Table 4 shows a full table of averages, including average times for each of the 5 functions defined in Figure 27.

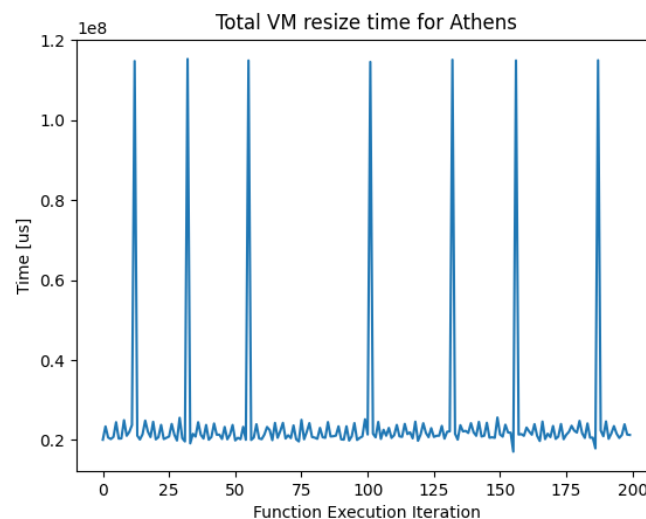


Figure 28 - Graph showing the total Scale Out Latency measurements for all 200 iterations in Athens

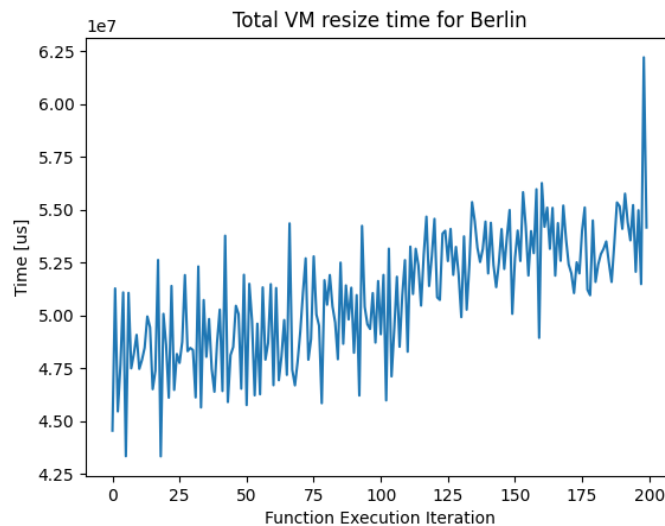


Figure 29 - Graph showing total Scale Out Latency measurements for all 200 iterations in Berlin

Table 4 - Average Scale Out Latency per Function/Time Period for Athens and Berlin

Function/Time Period	Athens Average (s)	Berlin Average (s)
Command: Power Off	0,49	2,74
Wait Until VM Power Off	16,45	32,42
Resize	0,47	2,47
Command: Resume	0,49	2,47
Wait Until VM Running	7,09	10,83
Total VM Resize Time	24,99	50,93

The function/period that appears to take the longest time across both platforms is waiting for the VM to shut down after the power off command has been received, with it taking 16 seconds and 32 seconds in Athens and Berlin respectively.

2.6 KPIs RELATING TO ORCHESTRATION

This Section contains KPIs relevant to measurement in the Orchestration layer:

- Service Provisioning Time (SRT)
- Slice Provisioning Time (**SLT**)
- Next Generation Application Part (NGAP) Restoration Time (NGT)

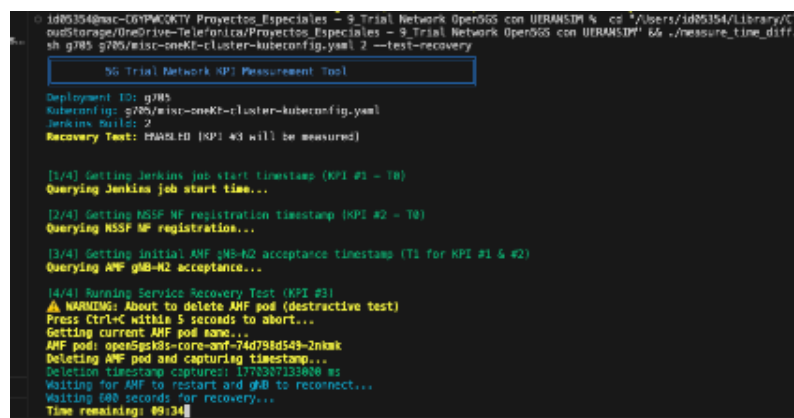
All three of these KPIs are related to the latency of orchestration and life cycle management of a Trial Network, measuring the time required to deploy, configure and recover network services within the 6G-SANDBOX infrastructure.

SRT relates to the time taken to provision the entirety of the Trial Network, that is to say all of the components listed in the Trial Network design. Once the time period is complete, the infrastructure can be said to be fully operational and ready for service usage. Said time period is defined as the duration from the timestamp when the Trial Network deployment pipeline is triggered to the timestamp when the Access and Mobility Management Function (AMF) successfully establishes the Next-Generation Application Protocol (NGAP) connection with the gNodeB.

SLT relates to the time taken from the provisioning of a slice within the 5G core to the establishment of a connection to the gNodeB. The time period is defined as the timestamp when the slice is created in the 5G core to the timestamp when the gNodeB successfully establishes the NGAP connection with the AMF.

NGT relates to the time taken to restore the connection between the gNodeB and the 5G core after a disruption, such as by an AMF restart or core network failure. The time period is defined as the timestamp when the AMF becomes unavailable to the timestamp when the NGAP connection is successfully re-established. This shows the recovery capabilities of the 5G Core/Radio Access Network (RAN)

For the measurements shown in this report, a Trial Network was deployed in the Malaga platform consisting of Open5GS acting as the network core connecting to a Nokia RAN which provided the gNodeB. Measurement of the KPIs is fully automated through the use of a Bash script framework that orchestrates a microcosm of the entire lifecycle of a Trial Network (TN) from instantiation to log extraction. The script is directly coupled with the setup in use and as such needs to be altered if the underlying TN is altered. The script first calls the TNLCM and initiates the pipeline for the provisioning of the TN. Once all necessary information is retrieved in order to have access to the TN cluster, the script can begin the process of situation triggering and timestamp collection. To ensure a high level of precision, timestamps are gathered from multiple different sources. The Jenkins API is used to extract timestamps from the pipeline itself for the purposes of the starting times for SRT and **SLT**. The starting time for NGT and the end time for all KPIs is extracted by interrogating the kubectl logs on the AMF and Network Slice Selection Function (NSSF) pods. The script filters for specific NGAP messages and uses the internal container timestamps for maximum accuracy. Once all timestamps are collected, the differences are calculated to generate the final KPIs. Figure 30 shows how the script interacts with the TNLCM, Jenkins, MinIO and the cluster in order to activate the Trial Network and gather timestamps for each event. The results for the example setup done on the Malaga platform can be seen in Table 5, with the results as presented by the measurement script seen in Figure 31.



```

id05154@mac:~/PROJECTS/Proyectos Especiales - 9.Trial Network Open5GS con UTRAN5G % cd ~/Users/id05154/Library/C/
outStorage/OneDrive-Telcel/Proyectos Especiales - 9.Trial Network Open5GS con UTRAN5G' && ./measure_time_diff.
sh g785 g785/misc-onekt-cluster-kubeconfig.yaml 2 --test-recovery

5G Trial Network KPI Measurement Tool

Deployment ID: g785
Kubeconfig: g785/misc-onekt-cluster-kubeconfig.yaml
Jenkins build: 2
Recovery Test: ENABLED (KPI #3 will be measured)

[1/4] Getting Jenkins job start timestamp (KPI #1 - T0)
Querying Jenkins job start time...

[2/4] Getting NSSF NF registration timestamp (KPI #2 - T0)
Querying NSSF NF registration...

[3/4] Getting initial AMF gNB-M2 acceptance timestamp (T1 for KPI #1 & #2)
Querying AMF gNB-M2 acceptance...

[4/4] Running Service Recovery Test (KPI #3)
WARNING: About to delete AMF pod (destructive test)
Press Ctrl+C within 5 seconds to abort...
Getting current AMF pod name...
AMF pod: open5gs-core-amf-74d79d549-2nbnk
Deleting AMF pod and capturing timestamp...
Deletion timestamp captured: 1770307133000 as
Waiting for AMF to restart and gNB to reconnect...
Waiting 600 seconds for recovery...
Time remaining: 09:34
  
```

Figure 30 - Logs of the Automated Gathering Script for Orchestration KPIs

Table 5 - Results for Orchestration Measurements on the Malaga Platform

KPI	Result
Service Provisioning Time (SRT)	19 min 11 s
Slice Provisioning Time (SLT)	7 min 9 s
NGAP Restoration Time (NGT)	4 min 5 s

These results indicate that:

- The full Trial Network deployment requires approximately **19 minutes**, including the orchestration of the infrastructure, the deployment of the 5GC and the configuration of connectivity with the RAN.

- The provisioning of a network slice requires approximately **7 minutes**, reflecting the time required for slice creation and network configuration.
- The recovery of the NGAP connection after a core disruption occurs within approximately **4 minutes**, demonstrating the resilience of the system under failure conditions.

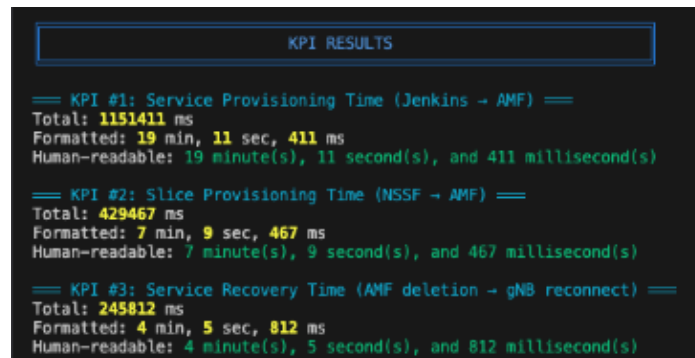


Figure 31 - Results for Orchestration Measurements on the Malaga Platform as seen in the execution script

2.7 KPIs RELATING TO LOCALIZATION

This Section contains KPIs relevant to measurement of Localization:

- Localization Functionality (**LOF**)

2.7.1 SCOPE

The capacity for Localization measurement within 6G-SANDBOX was provided by the integration of technology developed within the first round of Open Calls (OCs) at the Athens platform. This took the form of a custom version of the Location Management Function (LMF) implemented in free5GC [15]. The custom version included core modules such as communication, the LMF controller which supports selected LTE Positioning Protocols (LPP) and NR Positioning Protocol A (NRPPa) messages, and a basic localization algorithm based on Round Trip Time (RTT) and Time Difference of Arrival (TDOA) measurements. The LMF was implemented as a Python based, fully virtualized solution deployable on a server, virtual machine or a docker container, complying with 3GPP TS 29.572 [16].

Thus far, methodology has been developed for validating basic functionality of the Localization process within the free5Gcore environment. Full scale testing has not been executed at time of writing due to the requirements for very specialized hardware that is not available at this stage. Thus, the integration and methodology of the basic functionality will be documented here, but not the generation of the aforementioned KPIs.

2.7.2 FUNCTIONALITY METHODOLOGY AND RESULT DESCRIPTION

The previously mentioned LMF modules were refined to improve processing accuracy by incorporating a gradient descent-based localization algorithm that effectively utilized available RTT and TDOA data, compensating for the lack of other measurement types such as Angle of Arrival (AoA) and Angle of Departure (AoD). Concurrently, extensions to the free5GC Access and Mobility Management Function (AMF) were developed to facilitate communication with the LMF, enabling forwarding and processing of critical protocol messages, including LPP and NRPPa. These enhancements significantly improved data exchange between the AMF and LMF, ensuring precise message handling crucial for location estimation.

The Athens Testbed setup, used to validate the localization procedure, integrated the target User Equipment (UE), the Amarisoft gNB, and the LMF within the Free5GC core network. Figure 32 provides the functional architecture of the setup including the three functional modules of the LMF function: Controller,

Communication, and Localization Algorithms. The LMF is deployed as a Docker container within the other NFs in the core network, while Figure 33 illustrates the actual testbed components configuration deployed as part of the Athens Testbed.

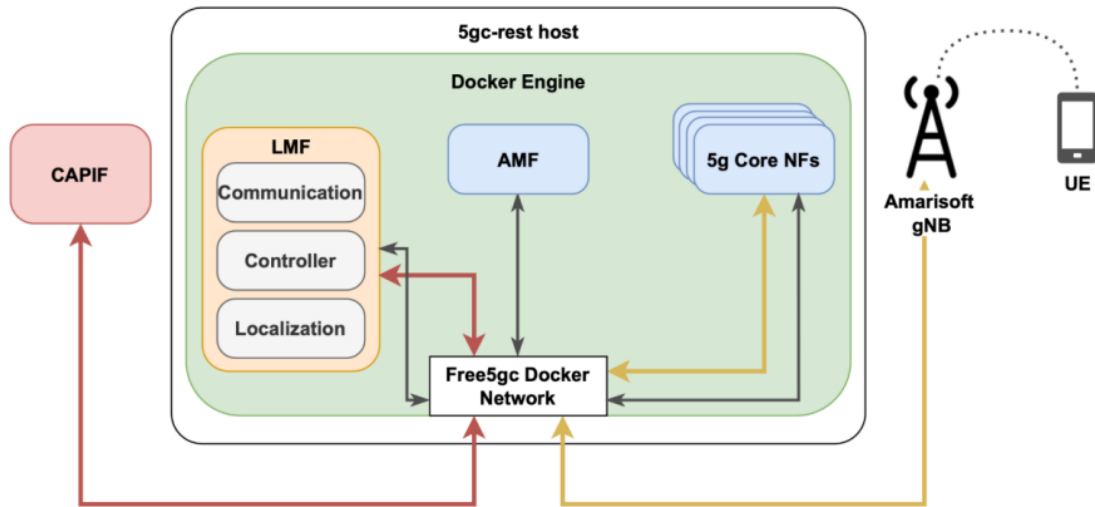


Figure 32 - Architecture of the LMF testing setup at the Athens Platform



Figure 33 - Lab setup for LMF testing at the Athens Platform

Three approaches/implementations were initially selected for evaluation. For both the first and second approach, the overall estimation procedure was not compliant with the 3GPP standard and was therefore rejected.

The third approach is a combination of the previous two and ended up being fully compliant. In the first phase of the procedure, positioning capabilities are exchanged between the UE and the LMF and this is compliant with the standard. However, the LMF requests positioning methods based on RTT and TDOA measurements, which are not supported by the UE. Hence, in the second phase, the LMF communicates with the auxiliary proxy to obtain simulated measurements. Then the estimation is obtained by processing the simulated measurements through a localization algorithm.

With the setup complete, a series of tests were conducted to validate the localization procedure and the end-to-end integration. Initially, the basic functionalities of LMF were tested, specifically the reception of an LPP and NRPPa messages. The results showed that the LMF can successfully encode/decode the LPP and NRPPa messages and handle the relative messages by advance in the localization procedure.

Subsequently, the functionality of the several localization algorithms which have been applied in the specific module inside the LMF were tested. Since the purpose of the test was to verify the correct functioning of the localization algorithms, the RTT and TDOA measurements were simulated. As a result, the LMF returns the estimation with these simulated TDOA/RTT measurements as inputs.

Another test is focused on verifying the connectivity between the AMF and LMF, specifically assessing the functions required for the downlink transmission of LPP and NRPPa messages. The enhancements made to the AMF enable it to successfully forward these messages to the UE and gNB.

The connectivity between the LMF and UE was tested to validate the correct exchange of LPP messages in the uplink direction. Different UE devices were used during these tests to ensure compatibility. For a Google Pixel 6, the results were successful (Figure 34): the LPP message generated by the LMF during the estimation procedure was correctly received by the UE, and the corresponding response was properly sent back to the LMF. However, for Huawei and OnePlus devices, the tests were unsuccessful (Figure 35); although the LPP message from the LMF was received by the UE, the devices did not respond to the LMF. This lack of response was attributed to the absence of protocol support on these devices. Therefore, future large scale testing would focus on measuring the KPIs on Google Pixel 6 Devices.

```

5GMM capability
Element ID: 0x10
Length: 1
0... .. = Service gap control (SGC): Not supported
.0.. .. = IP header compression for control plane CIOt 5GS optimization (5G-IPHC-CP CIOt): Not supported
..0. .... = N3 data transfer (N3 data): Not supported
...0 .... = Control plane CIOt 5GS optimization (5G-CP CIOt): Not supported
.... 0... = Restriction on use of enhanced coverage support (RestrictEC): Not supported
.... .1.. = LTE Positioning Protocol (LPP) capability: Supported
.... .1. = ATTACH REQUEST message containing PDN CONNECTIVITY REQUEST message for handover support (H0 attach): Supported
.... ..1 = EPC NAS supported (S1 mode): Supported

```

Figure 34 - LPP Message generated at the LMF during the estimation procedure for a Google Pixel 6 (Successful Test)

```

5GMM capability
Element ID: 0x10
Length: 1
0... .. = Service gap control (SGC): Not supported
.0.. .. = IP header compression for control plane CIOt 5GS optimization (5G-IPHC-CP CIOt): Not supported
..0. .... = N3 data transfer (N3 data): Not supported
...0 .... = Control plane CIOt 5GS optimization (5G-CP CIOt): Not supported
.... 0... = Restriction on use of enhanced coverage support (RestrictEC): Not supported
.... .0.. = LTE Positioning Protocol (LPP) capability: Not supported
.... .1. = ATTACH REQUEST message containing PDN CONNECTIVITY REQUEST message for handover support (H0 attach): Supported
.... ..1 = EPC NAS supported (S1 mode): Supported

```

Figure 35 - LPP Message generated at the LMF during the estimation procedure for a Huawei/OnePlus device (Unsuccessful Test)

2.8 KPIS RELATING TO RIS

This Section contains KPIS relevant to testing with Reconfigurable Intelligent Surface (RIS) technologies:

- RIS Capacity Enhancement (**RCE**)
- RIS Update Rate (**RUR**)

Testing of RIS technologies within 6G-SANDBOX has thus far been conducted within the Malaga platform and within the premises of the Queen's University of Belfast.

2.8.1 CAPACITY ENHANCEMENT

The methodology for **RCE** was developed over the course of work done in the 6G-SANDBOX Malaga platform, specifically in a controlled field trial in Plaza de la Constitución (Málaga, Spain), targeting signal reflection, channel gain enhancement, coverage reliability, and energy efficiency in a dense urban non-line-of-sight (NLoS) scenario. In this specific case, a passive reflector was used, not a RIS, but the methodology can be applied to testing both a passive and active RIS as well. The experimental procedure was designed to collect data regarding 4 key factors, both with and without the assistance of the passive reflector.

- **Alignment & Verification:** Reflector alignment with laser tools and Beam Index (BI) validation ensured that the path between the Base Station (BS) and the UE was true and reflector assisted.
- **KPI Logging:** Core metrics were logged for ~50s under stable reflector-assisted conditions. This was done through the use of the Nemo Outdoor tool [17]. Said metrics are as follows:
 - Received Signal Reference Power (RSRP)
 - Signal to Interference plus Noise Ratio (SINR)
 - Throughput
 - Transmit Power
 - Beam Index (BI)
- **Dynamic Reconfiguration Assessment:** Reflectors were physically removed mid-trial, forcing the BS to reconfigure beam direction and adapt power levels. This allowed measurement of reconfiguration speed, resilience, and energy adaptation mechanisms in real-world conditions.
- **Comparative Analysis:** Post-removal KPIs were compared with assisted measurements to quantify reflector-induced capacity enhancements.

These 4 factors together provide the information that constitutes the **RCE** KPI.

The Malaga field trial employed a commercial 5G mmWave BS (n258 band), a set of passive reflectors, and a user equipment (UE) placed in an alleyway without direct line-of-sight (LoS) to the base station (BS) [18]. The BS itself was connected back to a 6G-SANDBOX Trial Network deployed at the Malaga platform. Reflector geometry and measurement distances were standardized, with an incident angle of 45° from the BS and a reflected angle of 0° towards the UE, ensuring consistent test conditions across devices. A diagram of the example setup can be seen in Figure 36.



Figure 36 - Overview of the outdoor measurements using a cellular mmWave system with a Reflector to enhance connectivity in a narrow, NLoS urban environment

The final measure and conclusion with regards to **RCE** is taken by looking at all of the collected results as a whole. Figure 37 shows the results relating to Signal Reflection and Channel gain. The measured metrics are the RSRP, SINR and throughput. Each reflector showed a marked and consistent improvement in terms of the RSRP, with

the most notable case being reflector B increasing the average RSRP by around 20dB. These metrics can also be used to measure the coverage and reliability. Reflector D is seen to have increased the minimum throughput in its case from 0 (ergo, connection failure) to 361.3mbps and increased the average throughput significantly as well.

Ref.	Metric	w/ Reflector			w/o Reflector		
		RSRP	SINR	Thrpt.	RSRP	SINR	Thrpt.
A	Max.	-52.9	29.1	386.5	-54.7	29.9	488.0
	Min.	-56.0	25.1	365.3	-70.0	18.0	239.7
	Avg.	-54.1	28.3	378.5	-60.5	27.2	376.9
B	Max.	-40.0	29.1	389.2	-57.1	29.6	462.5
	Min.	-57.1	25.3	330.4	-66.0	26.7	330.7
	Avg.	-41.1	28.4	370.3	-61.0	29.0	414.9
C	Max.	-51.6	29.6	404.1	-56.5	29.9	429.2
	Min.	-56.5	26.3	357.0	-71.1	24.1	278.5
	Avg.	-51.9	28.8	372.2	-60.1	28.2	389.7
D	Max.	-51.5	28.4	407.5	-55.5	29.8	426.3
	Min.	-55.7	26.5	361.3	-72.5	23.2	0.0
	Avg.	-52.7	27.7	394.8	-60.2	27.2	249.1
E	Max.	-54.0	28.4	447.9	-55.0	28.7	470.0
	Min.	-56.1	27.1	359.0	-64.5	23.5	278.6
	Avg.	-54.7	27.8	399.5	-61.5	26.6	424.9
F	Max.	-50.1	30.3	461.9	-57.6	27.9	453.0
	Min.	-50.9	29.0	428.0	-67.4	21.4	337.5
	Avg.	-50.5	29.5	443.2	-60.5	25.8	393.0

Figure 37 - RSRP, SINR and Throughput Performance for the Malaga RIS Field Trial with and without reflectors A,B,C,D,E,F

By looking at the transmitting power of the UE as shown in Figure 38, we can understand the impact the reflectors have on the energy efficiency. Figure 38 shows that transmit power is consistently and significantly reduced with all tested reflectors, leading to a significant increase in energy efficiency in the network.

Ref.	Metric	w/ Reflector	w/o Reflector
A	Max.	-7.00	1.70
	Min.	-12.70	-10.30
	Avg.	-11.21	-4.75
B	Max.	1.60	1.70
	Min.	-24.40	-4.80
	Avg.	-23.30	-4.30
C	Max.	-6.90	-1.70
	Min.	-14.10	-6.90
	Avg.	-13.91	-5.30
D	Max.	-8.30	2.30
	Min.	-13.90	-9.00
	Avg.	-12.67	-5.54
E	Max.	-9.20	-1.80
	Min.	-11.70	-7.20
	Avg.	-10.86	-3.69
F	Max.	-14.60	-3.70
	Min.	-15.20	-8.00
	Avg.	-15.11	-5.22

Figure 38 - UE Transmission power performance for the Malaga RIS Field Trial

The direct effect of the introduction of RIS reflectors can be seen when the data is observed as a time series. This can be seen in Figure 39 and Figure 40, where the impact on SINR, RSRP and Transmit power when a RIS is removed is very noticeable.

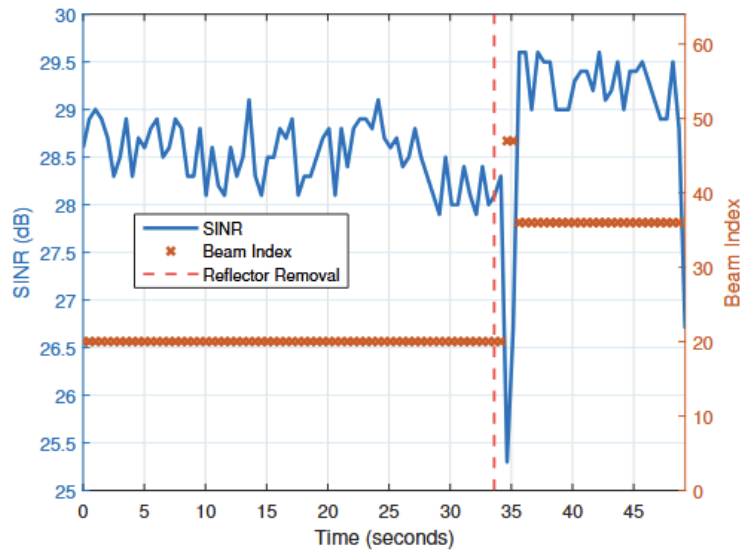


Figure 39 - Time series plot for SINR and Base Station Beam Index for Malaga RIS Field Trial with point of reflector B removal marked

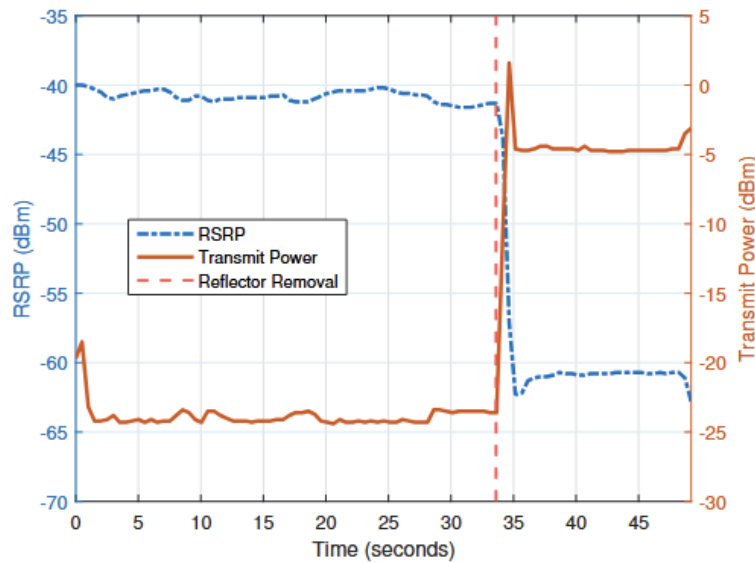


Figure 40 - Time series plot for RSRP and Transmit power for Malaga RIS Field Trial with point of reflector B removal marked

All of these results/analyses come together to provide the **RCE** 6G-SANDBOX KPI. Any component within this described methodology could be replaced with a perspective DUT – be it the RIS, the UE, the base station, or any of the components comprising the TN. The methodology can also be executed at any of the 6G-SANDBOX platforms given the correct conditions and physical environment exist.

2.8.2 UPDATE RATE

The practical measurements regarding RIS in the context of 6G-SANDBOX described in Section 2.8.1 were done with the use of a passive RIS, but 6G-SANDBOX also performed some testing with regards to active RIS, specifically the measurement of **RUR**. The active RIS in question was designed by QUB and is comprised of a controlling unit (microcontroller or FPGA) and the actual PCB with switching elements (PIN or varactor diodes, sometimes more elaborate RF switches). In terms of component update rates, the diodes can switch in nanoseconds, while the microcontrollers/FPGAs are in the order of microseconds, meaning they are the key with regards to the update rate of the RIS itself. The update rate of any active RIS is intrinsically linked to the

coherence time of the frequency it is communicating over. The coherence time is the time in which a channel can be assumed to be stationary and is therefore heavily influenced by the communication context and related to the Doppler spread. The equation used to determine coherence time in this context is shown below, with f_m being the relevant Doppler spread which is related to the communication frequency being used. The result of this equation is plotted as the blue line in Figure 41.

$$\sqrt{\frac{9}{16\pi} * \frac{1}{f_m}}$$

To test the update rate, the active RIS is connected to a Trial Network, in between a core and gNodeB on one end and a User Equipment (UE) on the other. The UE is walking at a constant speed of 1,4m/s. The frequency of communication is variable and in order to facilitate consistent connectivity, an active RIS should be able to update its elements at least once during the coherence time period. This means that a minimum update rate must be found where this interaction occurs. To measure the update rate of the RIS at different communication frequencies, an oscilloscope is hooked up to the shift registers modified by the RIS controller. A selection of diodes are also monitored and combined with the oscilloscope results to determine the update frequency. The test is run at a variety of different communication frequencies, with the results shown in Figure 41. In this case, this RIS crosses the boundary of 1+ updates per coherence time period at 29GHz, so the communication frequency in this context would need to be 29GHz or above.

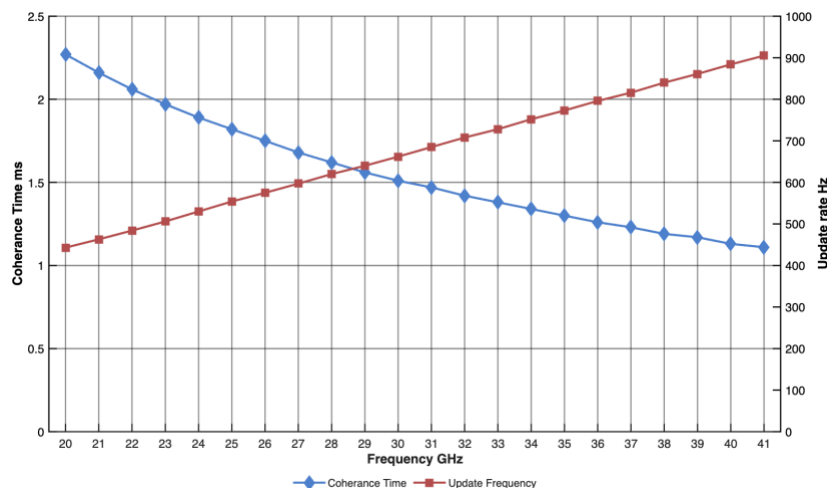


Figure 41 - Coherence time and RIS Update Frequency plotted for a UE walking at 1,4m/s

2.9 KPIs RELATING TO EMF EXPOSURE

This Section contains KPIs relating to Electromagnetic Field (EMF) exposure:

- Self EMF Exposure (**SEE**)
- Inter EMF Exposure (**IEE**)

These KPIs have different scopes, outlined below:

Self EMF Exposure

- 6G-SANDBOX adopts the definition of Self EMF Exposure from IEC 62209 [19]
- These measurements which are done with human phantoms are called generally as specific absorption rate (SAR) measurements. The SAR specification covers frequencies 4 MHz to 10 GHz [19].

- There is no possibility to measure self EMF exposure of 28 GHz base stations based on the SAR standard method.

Inter EMF Exposure

- 6G-SANDBOX adopts the definition of Inter EMF Exposure from IEC 62232 [19].
- These measurements are performed Over The Air (OTA) where the device under test (DUT) transmits RF signal, and the standard gain antenna is used for receiving the signal from measurement distance. Typically, tests are performed in RF chamber, but sometimes those can be performed in normal labs, as well.
- If EMF is measured over 6 GHz frequency, then the measurement method is based on power density, not absorbed RF power. The power density method describes RF exposure in terms of the density of absorbed power over area (W/m²) [20].

Within the context of 6G-SANDBOX, these KPIs are measurable at the Oulu platform, the methodology for which is detailed below.

2.9.1 EMF LIMITS FOR GENERAL AUDIENCE AND OCCUPATIONAL PERSONS

Inter EMF exposure for the 28 GHz base station needs to be measured in an OTA manner, and the first steps to establishing a methodology for EMF measurements is to understand what the standard is for safe limits. These differ depending on if the recipient is a member of the general public or someone who will be interacting with the emitting device in a regular, occupational manner. The EMF limits for the absorbed power over area used by 6G-SANDBOX are the following [21]:

- Occupational person limit: 5.0 mW/cm² or 50 W/m²
- General public limit: 1.0 mW/cm² or 10 W/m²

The effective isotropic radiated power (EIRP) is a metric which describes the transmission power level if the transmission is generated from single point and the radiation happens to all directions. The intent power flux is used if the device is not intended for use in proximity to the human body i.e. base stations, RF relay stations, and wireless backhaul units. The RF power flux density S of the radiation can be calculated as

$$S = \left(\frac{PG}{4\pi R^2} \right) = \left(\frac{EIRP}{4\pi R^2} \right),$$

where P is the input power to the antenna, G is the gain of the antenna compared to omni-directional antenna, EIRP is radiated transmission power in mW or in W towards the main lobe of the antenna and R is the measurement distance from the center of the radiator in cm or m. The unit of S is mW/cm² or W/m², and it is worth noting that the power flux density is independent of the operational frequency [22].

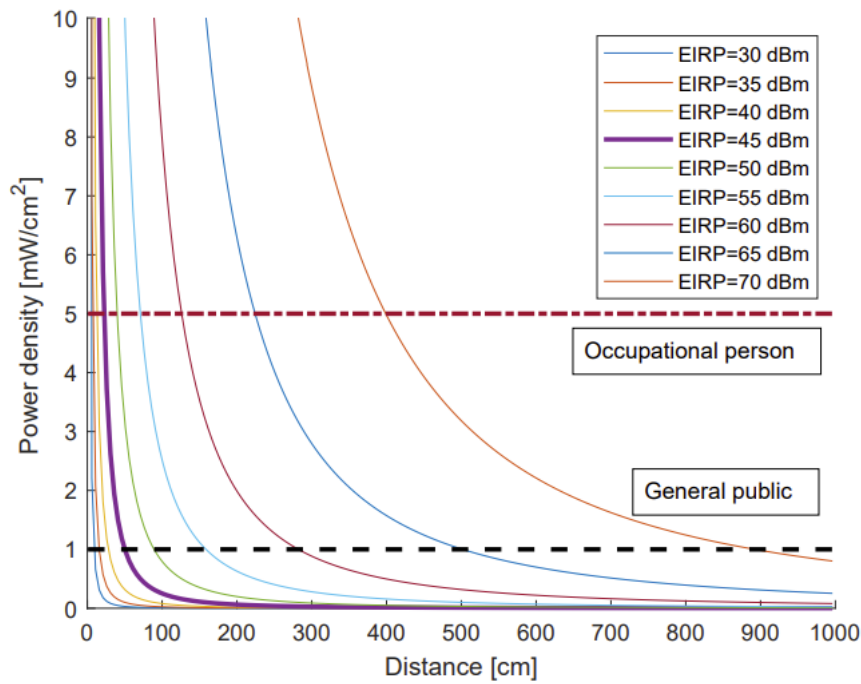


Figure 42: Safety limits for different radiated power levels with varied EIRP values.

2.9.2 UOULU EMF MEASUREMENT OF 28 GHz BASESTATION

The measurements for the EMF KPIs were performed at the University of Oulu laboratory equipped with a LiteON 5G mmW basestation (BTS) [23]. This base station operated a 5G NR signal with one component carrier of 100 MHz at 27.9-28.0 GHz frequency. The 5G NR signal was time domain duplex (TDD) configuration: DL:3, UL:1, Idle:1, total: 6 slots.

The block diagram of the EMF measurement setup is shown in Figure 43 where the base station is located on left and the OTA radio link to the standard horn antenna is in the middle. For prospective 6G-SANDBOX experimenter testing setups, any aspect of this block diagram could be replaced depending on the type of component the experimenter is putting under test. The free space path loss (FSPL) is marked next to the OTA link. The spectrum measurements were performed with a spectrum analyzer (Keysight UX50 GHz (N9040B)) in max hold mode with an RMS detector [24]. The used reference measurement antenna was LB-28-15-C-KF by A-info [25]. It covers frequencies 26 -40 GHz and the antenna horn size is 20.2 x 16.6 mm.

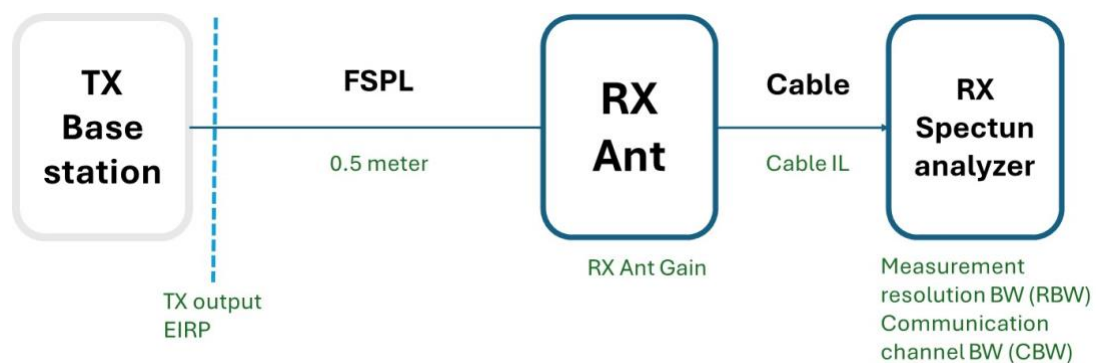


Figure 43: Block diagram of EMF measurement setup.

The EMF measurements were done in normal laboratory conditions where the BTS was mounted on robotic arm. The first measurements were performed with a 2-meter radio link. A photo of this setup is shown Figure 44(a).

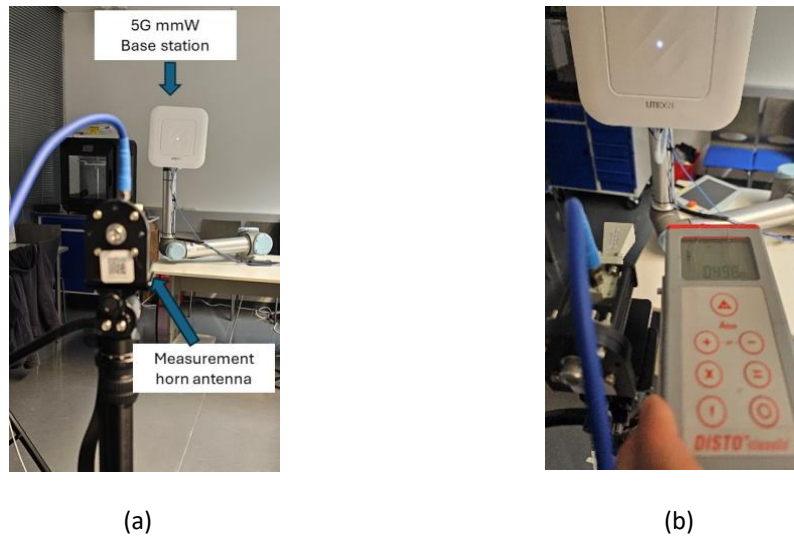


Figure 44: Measurement setups (a) Two meter OTA link and (b) 0.5-meter OTA link.

These measurements were not successful since no 5G signal was detected by the spectrum analyzer. The measurement distance was decreased down to 0.5 meters as shown in Figure 44(b) in following measurements and 5G signal was detected successfully. The measured spectrums from these measurements are shown in Figure 45. The wide band measurement is shown in Figure 45(a) and the narrowband results are shown in Figure 45(b).



Figure 45: Measured spectrum figures (a) Wide span spectrum and (b) 5G signal spectrum

The 5G signal power spectral density (PSD) was measured with markers from Figure 45 which have measurement bandwidth (BW) of 30 kHz. The measured PSD was projected over 100 MHz channel width. The free space path loss (FSLP), antenna gain and cabling losses were embedded from the results and a final measure for EIRP of 16.6 dBm was recorded. The measured EIRP power is 30 dB down from the data sheet value of 47 dBm.

The most likely cause of this discrepancy was the fact that the BTS was set to its normal operation mode for the duration of the measurement and connection to the UE. This meant that the BTS power control was activated and the BTS constantly changed the transmission power based on UE's channel indicator values. When measuring for a datasheet this measurement is performed in R&D or test mode which forces the BTS to constantly operate at maximum power, explaining the difference in the two results. The full set of results from this test can be seen in Table 6.

Table 6 - EIRP signal analysis based on measured 5G signal spectral power density of 28GHz Basestation

Measurement RBW (RMS)	30	kHz
Power on RBW on spectrum analyzer (SA)	-68	dBm
Signal Channel BW (CBW)	100	MHz
Ratio CBW/RBW	35,2	dB
Power on CBW on SA	-32,8	dBm
FSPL @ 0.5 m (Free space path loss)	55,7	dB
RX antenna gain	14,0	dB
Cable loss	5,0	dB
EIPR from LiteOn (continous TX)	13,9	dBm
TDD cycle (TX 3: RX 1) (DDDSU)	2,2	dB
EIRP from LiteOn with TDD TX	16,1	dBm

2.9.3 EMF RESULTS OF 28 GHZ BASE STATION

Based on these results, calculations for the EMF KPIs can be made, as shown in Table 7. The measured power density is 0.015 W/m² which is significantly lower than the general audience limit of 10 W/m². This measurement result indicates that there are no EMF violations from the used 5G base station. This EMF measurement result is according to the expected result that 28 GHz small cell base station is within safety limits of occupational persons and general public limits. This result is based on normal operation condition where the mobile terminal and base station are constantly altering transmission powers.

Table 7 - Power density based on measured transmission of 28GHz Basestation

EMF calculation of measurements			
R	Measurement distance	0,5	m
EIRP	RF power	16,6	dBm
		45,7	mW
		0,046	W
S	Power density	0,015	W/m ²

If, as previously discussed, the EIRP data sheet value of 47 dBm is used for EMF analysis then the general audience safety limit is 64 cm and the occupational persons limit is 29 cm as shown in Table 8. Given that this scenario is occurring when the base station is configured in an R&D or testing mode, the assumption can be made that this level of EMF exposure would never be experienced by the general audience, so the larger safety limit is not of significant consequence. It can be assumed that during operation within this mode the operator is knowledgeable and understands how to act around a device actively emitting higher than standard amounts of EMF radiation.

Table 8 - EIRP signal analysis based on measured 5G signal spectral power density

LiteOn BTS datasheet information			
EIRP	Maximum RF EIRP power	47	dBm
		50119	mW
		50,12	W
	General audience range	0,6315	meter
R		63,2	cm
	Occupational person range	0,2824	meter
R		28,2	cm

These results can be applied to EMF measurement in many different contexts, including the two contexts required by 6G-SANDBOX, those being the two listed KPIs of **SEE** and **IEE**. Relevant measurements and conclusions for these KPIs would be achieved through changing the measurement distances and devices under test based on the specific KPI being measured.

2.10 KPIS RELATING TO SENSING

This Section contains KPIS relating to the field of Sensing:

- Real Scenario Images (**RSI**)
- Audio Capture (**ACA**)
- Depth Capture (**DCA**)
- Pose and Orientation of XR Device (**POX**)
- Velocity of Device (**VOX**)
- Object Detection (**OBD**)
- Egocentric Segmentation (**EGS**)
- Augmented Reality Rendering (**ARR**)
- Virtual Reality Rendering (**VRR**)

The testing of the Sensing KPIS for 6G-SANDBOX is provided by the **University of Oulu** in its *Multimodal Sensing Lab*, which is an interactive space equipped with Meta Quest 2 and Meta Quest 3 headsets, Azure Kinect RGB-D sensors, and dedicated compute infrastructure for real-time and offline computer vision processing [26] [27] [28]. This environment was used to design the methodology for and execute examples of a series of XR experiments specifically targeting the Sensing based KPIS defined for Extended Reality (XR) devices in 6G-SANDBOX Deliverable D2.1 [29].

The Sensing KPI family covers both raw sensor signals and higher-level perception outputs required to build an accurate model of the environment around an XR user. Within 6G-SANDBOX, the infrastructure documented here provides sensing modalities can be accessed, logged, and exploited in a controlled way during XR experiments, and that can be used to provide quantitative information on user motion and scene dynamics. This sensing layer can also support other KPI families, for example by providing ground truth for EMF exposure experiments or communication-centric tests that involve realistic user behavior.

2.10.1 6G-SANDBOX SENSING ARCHITECTURE/METHODOLOGY

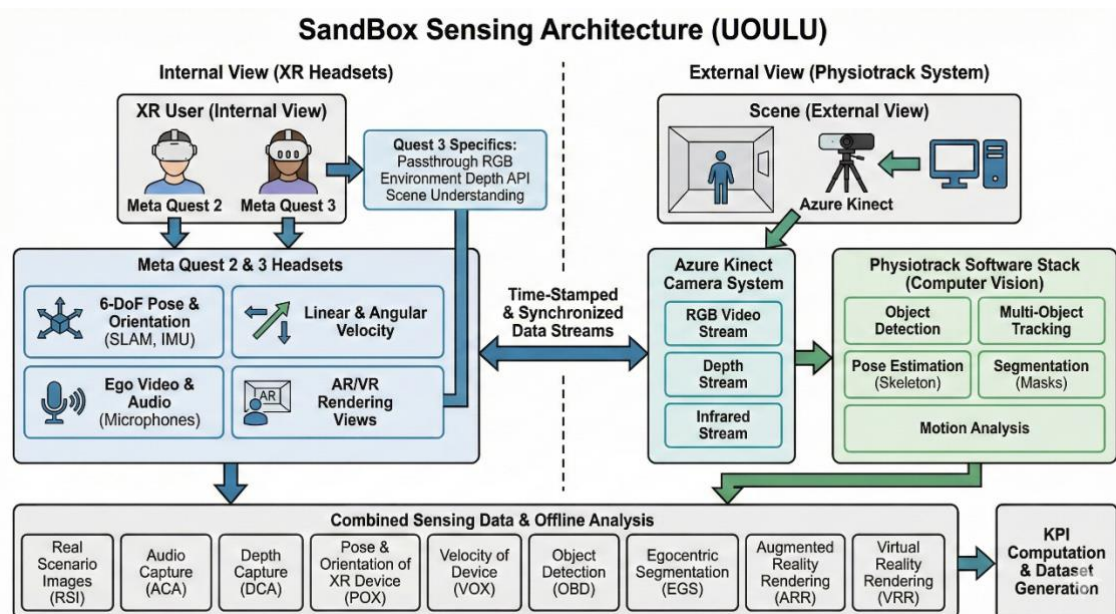


Figure 46 - Diagram showing the structure of the 6G-SANDBOX Sensing Architecture present at UOULU

The sensing architecture developed by UOULU combines two complementary sources of information, as depicted in Figure 46:

- an **internal view** provided by commercial XR headsets (Meta Quest 2 and Meta Quest 3), and

- an **external view** provided by a dedicated camera system and a computer vision software stack, Physiotrack [30].

The Meta Quest devices are used as sensor platforms in addition to their role as XR endpoints. Both Quest 2 and Quest 3 provide real-time 6 Degree of Freedom (DoF) pose and orientation information about the headset via Simultaneous Localization and Mapping (SLAM) and inertial sensors, as well as linear and angular velocity. They also support recording of video from the user point of view, together with audio from the headset microphones. On the Quest 3, passthrough RGB and an environment Depth API are available, and the platform exposes basic scene understanding in the form of room meshes, planes, and simple semantic labels. The devices naturally satisfy and can be used to validate the **ARR** and **VRR** KPIs, since they generate and display AR/MR views and fully virtual scenes internally.

The external view is built around one or more RGB cameras, including Azure Kinect devices that provide synchronized RGB-D and infrared streams. These cameras observe the user and the surrounding environment from an allocentric viewpoint. The video streams are processed by Physiotrack, which implements a modular pipeline for detection, tracking, pose estimation, segmentation, and motion analysis using state-of-the-art deep learning models. Typical configurations use a single Azure Kinect mounted in front of or above the interaction space, with at least one user wearing a Quest headset in the field of view.

All relevant streams from the headsets and the external cameras are time-stamped and recorded. Tracking data from the Quest devices, video and audio, external RGB-D, and all derived computer-vision outputs can therefore be re-aligned offline and used for KPI computation or to generate annotated datasets.

2.10.2 QUEST HEADSET-BASED SENSING

A number of the sensing KPIs can be directly satisfied just by the sensor suite on the XR devices, in this case the Meta Quest 2 and 3. Figure 47 shows both devices side to side.



Figure 47 - The Meta Quest 2 and 3 headset devices, the primary internal sensing devices used in 6G-SANDBOX's Sensing lab at UOULU. They provide 6-DoF pose and velocity, video, audio and AR/VR Rendering. The Meta Quest 3 also offers environment depth sensing

For **RSI** and **ACA**, both devices provide ego-view recordings in compressed video format together with headset microphone audio. These recordings correspond directly to the KPI definition of a compressed image or video of the real scenario from the perspective of the XR user and can be analyzed frame by frame if necessary.

POX and **VOX** are also supported by both devices. The Quest tracking system exposes the 6-DoF pose of the headset in several tracking spaces and provides linear and angular velocities derived from the IMUs and SLAM. These quantities are available in real time during experiments and are logged for potential later analysis.

For **DCA**, the situation is more heterogeneous. Quest 2 does not provide a per-pixel depth image but offers a reconstructed scene and mesh representation that encodes geometry. For **DCA** purposes, depth can be obtained by applying depth-from-RGB models to ego recordings. Quest 3, in contrast, provides an explicit Depth API that returns environment depth maps and thus supports real-time depth sensing. Both devices expose limited scene

semantics (for example walls, floors, and tables), but they do not provide a general-purpose object-detection service and thus cannot provide **OB** by themselves. Similarly, there is no native output corresponding to a full egocentric body mask of the user, so **EG** also cannot be provided.

As previously mentioned, both devices natively provide **ARR** and **VRR**. In passthrough or mixed-reality mode, the devices generate AR/MR views where virtual elements are overlaid on the real-world camera image. In VR mode, they render fully virtual scenes. These images can also be recorded as ego videos and used as part of the Sensing KPI reporting.

Overall, the headsets themselves can provide a significant portion of the 6G-SANDBOX sensing KPIs. This means that the testing methodology is potentially flexible and could be used outside of the UOULU premises, at least for certain KPIs. Table 9 depicts the KPI coverage obtained from the use of exclusively Quest2 and Quest3 devices and their available APIs.

Table 9 - Coverage of the 6G-SANDBOX Sensing KPIs by the Meta Quest 2 and 3 acting independently

KPI	Characteristics
RSI	Quest 2: Partially supported – Ego view can be recorded as RGB video with overlays and processed offline; no access to raw camera RGB frames for sensing. Quest 3: Supported – Passthrough Camera API provides a real time RGB stream for sensing, and ego video can also be recorded.
ACA	Quest 2: Supported – Headset microphone provides an audio stream and is recorded together with ego video. Quest 3: Supported – Same model as Quest 2; audio can be used as a sensing signal and for offline analysis.
DCA	Quest 2: Partially supported – Scene / mesh APIs provide a reconstructed environment geometry, and offline depth estimation is possible from ego RGB (for example with Depth Anything v2); no direct per pixel depth stream. Quest 3: Supported – Depth API exposes real time environment depth maps suitable for sensing; offline depth estimation from ego RGB is also possible.
POX	Quest 2: Supported – 6 DoF headset pose and orientation provided by SLAM and IMU tracking in defined tracking spaces. Quest 3: Supported – Same sensing model with improved tracking and anchor support.
VOX	Quest 2: Supported – Linear and angular headset velocity available via tracking and IMU APIs. Quest 3: Supported – Same, with OpenXR fields for velocities.
OB	Quest 2: Partially supported – Scene understanding yields a limited set of semantic classes (walls, tables, etc.); generic object detection typically relies on offline processing of ego video. Quest 3: Supported – Generic object detection can run on passthrough RGB in real time, in addition to scene semantics.
EG	Quest 2: Partially supported – Egocentric segmentation can be obtained offline by processing recorded ego video; no native real time egocentric body mask output. Quest 3: Supported – Passthrough RGB enables real time egocentric body segmentation with custom models, providing a body mask as a sensing signal (even though there is no dedicated platform API for the mask).
ARR	Quest 2: Supported – XR/AR content rendered directly in the headset; the composited ego view can be recorded. Quest 3: Supported – Same, with depth based occlusion and richer scene understanding.

VRR	Quest 2: Supported – Standard stereo VR rendering to the headset displays; ego VR video can be recorded. Quest 3: Supported – Same VR rendering model with higher resolution and improved optics.
-----	---

2.10.3 EXTERNAL SENSING WITH PHYSIOTRACK

Physiotrack is an open-source software framework developed by UOULU to provide external, scene-centric sensing around the XR user. It connects to external cameras and, in the 6G-SANDBOX configuration, to Azure Kinect, which supplies aligned RGB, depth, and infrared channels. Figure 48 shows the Physiotrack inference pipeline.

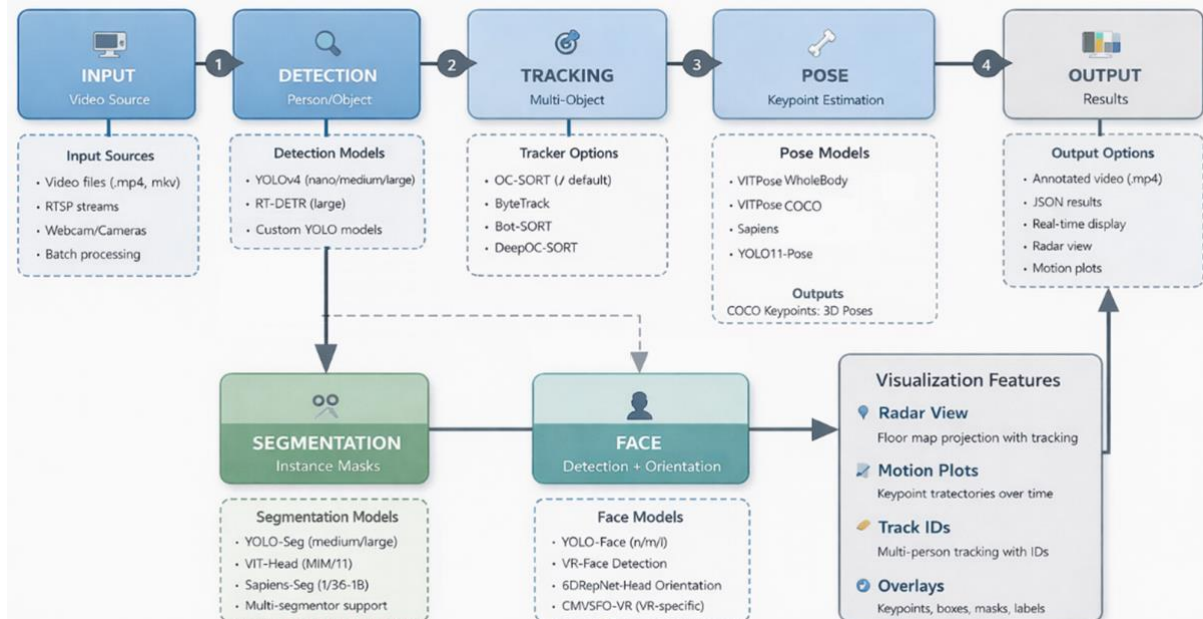


Figure 48 - Physiotrack full inference pipeline

At the detection and tracking stage, Physiotrack uses recent object detectors (for example, the YOLO and RT-DETR families and VR-specific variants) to identify persons and objects in each frame and to output bounding boxes with confidence scores. Multi-object tracking algorithms (such as OC-SORT, ByteTrack, BoT-SORT, and DeepOC-SORT) maintain consistent IDs over time, which makes it possible to follow multiple users and objects over long sequences and to extract their trajectories. External models such as Depth Anything v2 can provide depth images estimated from their monocular counterparts.

For human pose estimation, Physiotrack integrates several interchangeable models, including ViTPose, Sapiens, and YOLO-Pose variants. These models estimate 2D skeletons for each tracked person, with output formats ranging from COCO-style 17-joint skeletons to WholeBody skeletons with 133 joints. An optional face analysis module detects faces and estimates head orientation (yaw, pitch, roll), which is useful in experiments where user attention or gaze direction is of interest.

For **EGS**, Physiotrack runs instance segmentation models that produce pixel-level masks for persons and headsets. In the current configuration, these masks are allocentric (from the external camera viewpoint) rather than egocentric, but they still enable precise reasoning about which pixels in the scene belong to the user, the headset, or other objects.

In addition, Physiotrack implements floor-mapping and motion-plotting tools. A homography is used to project tracked persons onto a 2D floor plane, creating a radar-style top-down view of the scene. Time-series plots of

selected key points (for example wrist or ankle positions) are generated to analyze motion patterns, including periodic movements.

From the perspective of the 6G-SANDBOX Sensing KPIs, Physiotrack therefore provides **RSI, DCA, OBD, POX, VOX**, and a slightly different version of **EGS**.

Physiotrack does not render VR content and only overlays AR-style visualization on the external video for analysis. AR and VR rendering KPIs remain associated with the headset. Table 10 depicts the KPI coverage of the external video tracking system.

Table 10 - KPI Coverage of external video tracking system

KPI	Characteristics
RSI	Supported – External RGB video is ingested as a live sensor for detection, tracking, pose, and segmentation, and recordings can be analysed offline. Depth Anything v2 can be applied to recorded RGB for additional cues.
ACA	Partially supported – Audio can be recorded with an external microphone and time aligned with video, but the main sensing pipeline is based on visual streams rather than real time audio processing.
DCA	Supported – Kinect Azure provides real time metric depth (and infrared) aligned with RGB. Depth Anything v2 can be applied offline to recorded external RGB for dense depth estimation.
POX	Supported – Headset pose and orientation are estimated externally from RGB D using detection and pose estimation in the camera frame.
VOX	Supported – Linear velocities and approximate angular velocities are derived from tracked trajectories and pose changes over time for the headset and body joints.
OBD	Supported – Real time generic object detection (for example YOLO, RT DETR and variants) runs on the external RGB stream, providing bounding boxes, labels, and IDs.
EGS	Partially supported – Real time body and headset masks are produced from the external camera viewpoint (allocentric segmentation); masks do not come from the headset's egocentric view.
ARR	Partially supported – AR style overlays (bounding boxes, skeletons, masks, radar view) are rendered on top of the external video stream for analysis, but no AR content is rendered in the XR headset.
VRR	Unsupported – The system analyses real world camera streams and does not render fully virtual scenes.

2.10.4 COMBINING EXTERNAL SOLUTIONS AND HEADSETS – THE 6G-SANDBOX SENSING LAYER

Both of these types of solutions work in concert to provide 6G-SANDBOX's sensing layer solution package and cover all necessary KPIs. With regards to KPIs relating to sensing and signals (image, audio, depth, pose, and velocity), both solutions work in concert. Ego video and audio from the Quest headsets provide the user's point of view, while external RGB and depth from Azure Kinect provide an allocentric view of the same scene. Depth sensing is available both from Quest 3 (environment depth) and from Azure Kinect (metric depth and infrared) and can be further enhanced with offline depth-from-RGB models applied to recorded sequences. Pose and

velocity are available from two independent sources: the headset's internal SLAM and IMU system, and the external tracking from Physiotrack. This redundancy enables cross-validation and more robust motion analysis.

For KPIs relating to scene comprehension, Physiotrack supplies generic object detection, tracking, and segmentation at scene level, while the Quest devices provide ego-view recordings that can be used for egocentric analysis. Egocentric segmentation can be obtained by running body-segmentation models on Quest 3 passthrough images, using external segmentation masks from Physiotrack as supervision and validation signals. In this way, the 6G-SANDBOX sensing layer can address both allocentric and egocentric variants of the segmentation KPI.

For KPIs relating to rendering, the Quest headsets provide the AR/MR and VR images delivered to the user. These can be recorded and linked with the sensing streams to relate network or EMF conditions to what the user actually sees. Physiotrack contributes complementary overlays and dashboards that visualize detections, skeletons, masks, and trajectories on the external video for experiment monitoring. Table 11 depicts the KPI coverage obtained by a combined sensing system including both headset-based system and external video-based system with computer vision models.

Table 11 - Combined Sensing KPI coverage of the 6G-SANDBOX Sensing Layer (Meta Quest 2 / Meta Quest 3 + Physiotrack).

KPI	Characteristics
RSI	Supported – Real time RGB sensing from the external camera (PhysioTrack) combined with ego RGB recording and, for Quest 3, passthrough RGB sensing from the headset.
ACA	Supported – Ego audio from Quest microphones combined with (optional) external audio recording; multiple audio channels can be recorded to then be analysed and aligned with visual data.
DCA	Supported – Real time depth from Kinect Azure and Quest 3 Depth API, complemented by offline dense /accurate depth estimation from Depth Anything v2 on both ego and external RGB video.
POX	Supported – Internal SLAM based pose and orientation from the headset (optionally) fused with external pose from PhysioTrack for robust, cross validated XR device pose sensing.
VOX	Supported – Velocity information from headset IMUs and from externally observed motion via PhysioTrack, enabling redundant kinematic sensing and validation.
OBD	Supported – Scene level generic object detection from PhysioTrack, with the option to add ego view detection on Quest 3 passthrough and offline detection on ego recordings.
EGS	Supported – Egocentric body masks can be produced from Quest 3 passthrough with custom models, while PhysioTrack provides external body and headset masks that act as supervision and validation from an allocentric view.
ARR	Supported – AR and MR content rendered in the headset by Quest devices, combined with external AR style overlays and analytics from PhysioTrack on the scene video.
VRR	Supported – VR rendering provided by the Quest devices, complemented by PhysioTrack's external sensing of users and environment during VR experiences.

Figure 49 shows a qualitative example of the measurement of sensing KPIs at work, combining ego-view data and external data in a synchronous way. The external RGB view shows headset wearer detection, skeletonization and segmentation in 2 distinct moments in time. The visualization includes the external video as a main frame with inset panels for depth information, ego video, floor-map radar view and time series plot of wrist motion.

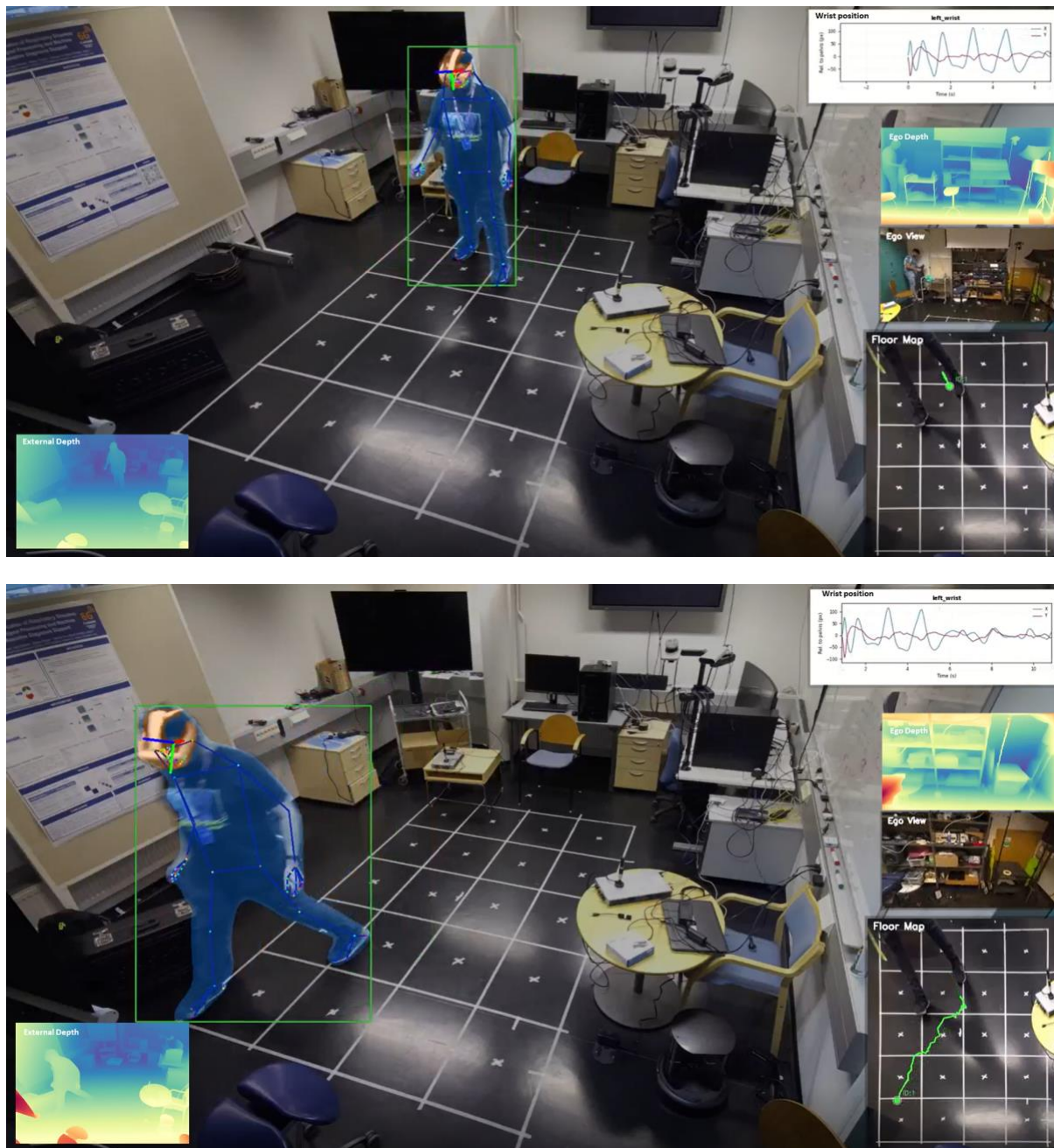


Figure 49 - The 6G-SANDBOX Sensing layer interface shown at two distinct points in time. The external camera view is shown in the main frame, with other inputs shown as inset panels

Each and any of the compromising components on both sides of the sensing layer may be replaced by an experimenter for the purposes of comparative testing of a desired component. Alternatively, the entire sensing layer can be used as a basis to build a methodology for the testing of an external component, network protocol, or whatever the experimenter desires in concert with the rest of the 6G-SANDBOX infrastructure present at UOULU.

2.11 KPIs RELATING TO ANOMALY DETECTION

This Section contains KPIs relating to testing of Anomaly Detection:

- Anomaly Detection Precision (**ADP**)

Anomaly detection (AD) components are implemented to act as security measure on two different interfaces in the core network, those being the N2 and the N6 interface. These anomaly detection algorithms are AI/ML-based and are trained to identify specific type of attacks. By implementing anomaly detection models over 2 different interfaces, it becomes possible to:

- Identify attacks coming from malicious User Equipments (UE) and compromised base stations that can have access to the Non-Access Stratum (NAS) protocol.
- Identify and mitigate attack originated from malicious UEs on critical infrastructures (slice security) and block them in case of denial of service (DoS).

2.11.1 ANOMALY DETECTION ON N2 INTERFACE

Replay attacks on the N2 interface represent a significant security threat in 5G Core networks, particularly targeting the communication between the gNB and the Access and Mobility Management Function (AMF). The N2 interface carries critical NGAP (NG Application Protocol) signaling messages responsible for mobility management, session establishment, handovers, and UE registration procedures. According to the Security Assurance Specifications (SCAS) for the AMF, replay attacks can exploit weaknesses in signaling validation mechanisms by capturing legitimate NGAP messages and retransmitting them at a later time [31]. This may lead to unauthorized session re-establishment, false mobility events, duplicated signaling procedures, resource exhaustion, or even service disruption within the core network.

The challenge with replay attacks on the N2 interface is that the replayed messages often appear syntactically correct and originate from previously valid signaling exchanges. Traditional rule-based security mechanisms and signature-based intrusion detection systems are frequently insufficient to detect these sophisticated attacks because the signaling content itself may not contain malformed parameters or obvious anomalies [32]. Instead, the attack manifests through subtle deviations in signaling behavior, timing patterns, sequence inconsistencies, message repetition frequency, and contextual correlations across multiple network entities.

To address this security gap, an AI-based Anomaly Detection (AD) system becomes highly important and strategically valuable. By leveraging machine learning techniques and behavioral analytics, the proposed anomaly detection solution can continuously monitor N2 signaling traffic and identify abnormal patterns associated with replay attempts in real time. Unlike static detection approaches, AI-based systems can learn the normal operational behavior of AMF signaling procedures and dynamically detect deviations that may indicate malicious replay activity, even when the replayed messages are technically valid according to protocol specifications.

The implementation of an AI-based anomaly detection mechanism significantly enhances the security posture of the 5G Core network by enabling proactive threat detection and faster mitigation of replay attacks before they impact network stability or service availability. Furthermore, such a solution aligns with SCAS security objectives by strengthening the AMF's resilience against advanced signaling threats and improving the overall trustworthiness of the N2 communication plane. In highly dynamic 5G environments, where signaling complexity and traffic volume continue to increase, intelligent anomaly detection becomes essential to ensure secure, reliable, and uninterrupted network operations.

2.11.1.1 EXPERIMENTATION SETUP

The setup created to test the AD is described in Figure 50. There are three main components involved in the TN:

- LoadCore 5G Core Network Simulator [33]: Used to generate regular and malicious traffic at the N2 interface.

- Anomaly Detection Component: Main component designed and implemented as a part of this experimentation effort with the purpose of detecting replay attacks and distinguish them from normal network behavior.
- 5G Core Network – Open5GS: Used as System Under Test (SUT).

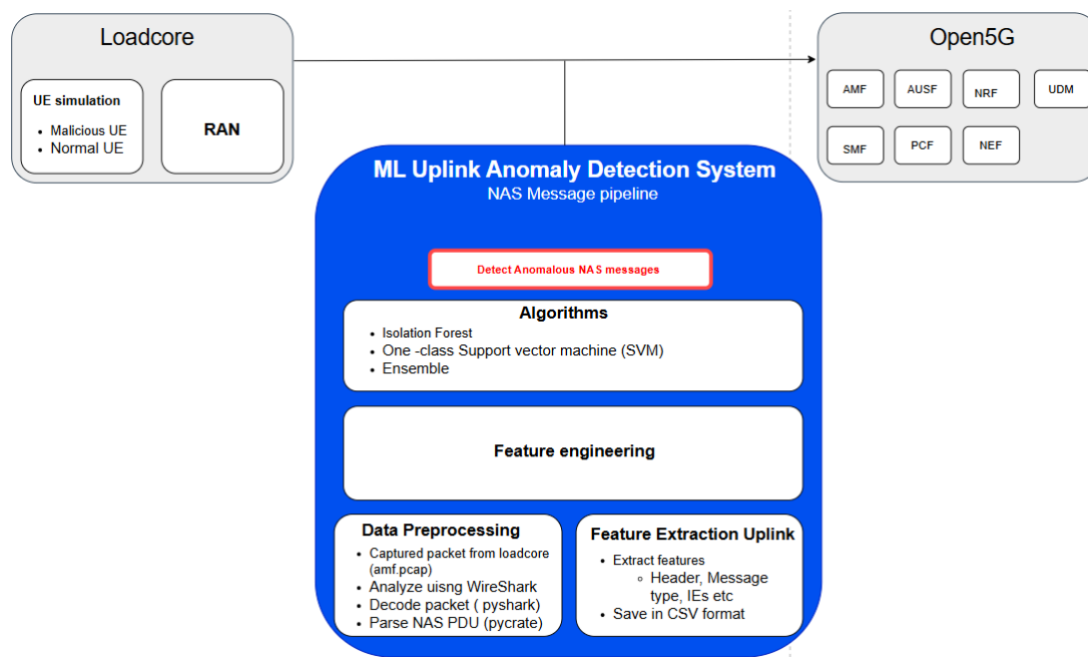


Figure 50: Setup to collect dataset and run experiments.

Figure 50 illustrates the architecture and processing pipeline of an **ML-based Uplink Anomaly Detection System (N2 Interface)** designed to detect anomalous or malicious NAS signaling messages in a 5G network environment. The system operates in the uplink direction, monitoring traffic generated by (UEs) and Base Station (gNBs) toward the 5G Core network.

On the left side of the diagram, the **LoadCore** platform is used to emulate the radio access environment. It generates both **normal UEs** and **malicious UEs**, enabling the creation of realistic network traffic scenarios for training and evaluating the anomaly detection models. The RAN (Radio Access Network) forwards the NAS signaling traffic toward the 5G Core.

On the right side, the **Open5GS** Core Network contains the primary 5G Core Functions, including:

- AMF (Access and Mobility Management Function)
- AUSF (Authentication Server Function)
- NRF (Network Repository Function)
- UDM (Unified Data Management)
- SMF (Session Management Function)
- PCF (Policy Control Function)
- NEF (Network Exposure Function)

Open5GS works as SUT and provide realistic control plane traffic from network functions used on the experimentation deployment.

The central component of the architecture is the **ML Uplink Anomaly Detection System**, which acts as an intelligent security analysis pipeline for NAS signaling messages exchanged between the UE and the AMF.

The workflow of the anomaly detection system is divided into several stages:

Data Preprocessing: Captured packets are collected from the AMF using packet capture files (amf.pcap). These packets are analyzed using Wireshark and decoded using Python-based packet analysis tools. The NAS PDUs (Protocol Data Units) are then parsed to extract relevant signaling information.

Feature Extraction: During this phase, important NAS signaling attributes are extracted from the decoded packets. These features include:

- NAS headers
- Message types
- Information Elements (IEs)
- Protocol-specific parameters

The extracted information is then organized and stored in CSV format to facilitate machine learning processing.

Feature Engineering: The extracted raw features are transformed and optimized to improve the learning capability of the machine learning models. This stage may include normalization, encoding, feature selection, and statistical correlation analysis to better characterize normal and anomalous NAS signaling behavior.

Machine Learning Algorithms: The system applies three AI/ML-based anomaly detection techniques:

- **Isolation Forest:** This model builds on random trees to isolate data points. It uses unlabeled data. Its use here is due to anomalies being easier to isolate because they have shortened paths.
- **One Class Support Vector Machine:** The novelty detection for SVM is used to learn what normal looks like and then detect anomalies on new data.
- **Optimized Ensemble:** The Ensemble is a weighted voting system, that combines both model scores in order to achieve a better overall performance. ($w1 \cdot \text{score1} + w2 \cdot \text{score2}$).

These algorithms are trained primarily on normal network behavior and are capable of identifying deviations that indicate potentially malicious or abnormal NAS signaling activity.

Anomaly Detection: The final objective of the pipeline is to **detect anomalous NAS messages** in real time. By continuously analyzing uplink signaling traffic, the system can identify suspicious UE behavior, signaling abuse, replay attempts, malformed NAS procedures, or signaling floods that could target the AMF or impact the stability of the 5G Core network.

2.11.1.2 RESULTS

The different types of message classifications are defined thusly:

- TP indicates malicious NAS messages that the system correctly flagged as abnormal
- FP are non-malicious NAS message falsely detected as an abnormal.
- FN represent abnormal NAS message that is wrongly sense as a normal.
- TN represent legitimate NAS message that is correctly classified.

The metrics used to analyze performance in terms of Anomaly Detection are as follows:

Precision score: $TP / (TP + FP)$, A high precision score indicates that most of the abnormal messages are being detected.

Recall: $TP / (TP + FN)$ If most of the abnormal NAS messages are being caught, then the Recall score will be high, otherwise it will be low.

F1-score: The F1-score gives a weighted harmonic mean of the precision and recall, where an F-beta score reaches its best value at 1 and worst score at 0 ($2 \cdot \text{Precision} \cdot \text{Recall} / (\text{Precision} + \text{Recall})$). It gives a balance between the Precision and the Recall scores.

Each of the model has a score for Precision, Recall and F1-score. The Ensemble takes the average scores of both models for detection.

As Figure 51 shows, the best performance is given by the Optimized Ensemble model, with F1 values of 100 – meaning fully correct identification of all types

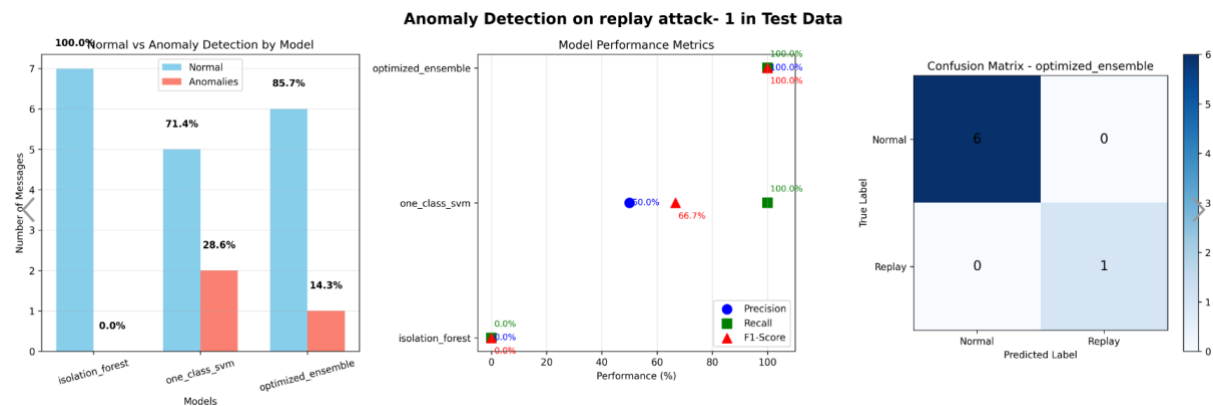


Figure 51: Results of the N2 Layer Anomaly Detection testing, showing the differing performances from the 3 implemented models.

2.11.2 ANOMALY DETECTION ON N6 INTERFACE AND NEAR TO REAL-TIME REACTION

Denial-of-Service (DoS) and Distributed Denial-of-Service (DDoS) attacks on the N6 interface represent one of the most critical security threats to the 5G Core data plane infrastructure. The N6 interface connects the User Plane Function (UPF) to external Data Networks (DNs), carrying massive volumes of user traffic and enabling connectivity to Internet services, enterprise applications, and mission-critical communication systems. In this context, malicious UEs can intentionally generate abnormal or excessive traffic flows toward the N6 interface, overwhelming UPF resources and causing severe degradation of network performance, service availability, and slice isolation mechanisms.

Unlike signaling-plane attacks, DoS and DDoS attacks on the N6 interface directly target the data plane capacity and processing capabilities of the UPF. Attackers may use compromised or malicious UEs to generate high-rate traffic floods, abnormal packet bursts, malformed flows, or coordinated distributed attacks that consume bandwidth, CPU, memory, and packet-processing resources within the UPF. In network slicing environments, these attacks become even more dangerous because excessive traffic generated within one slice may negatively impact the performance, latency, and reliability of other slices sharing the same infrastructure. This creates a major security concern for critical communication services such as public safety, industrial automation, healthcare, autonomous transportation, and ultra-reliable low-latency communications (URLLC), where service continuity and isolation are essential.

Traditional security mechanisms based on static thresholds, predefined signatures, or conventional traffic filtering are often insufficient to detect modern DoS/DDoS attacks in highly dynamic 5G environments [34]. Attack traffic generated by UEs may initially resemble legitimate high-volume traffic patterns, making it difficult to distinguish malicious behavior using conventional rule-based approaches. Furthermore, distributed attacks originating from multiple UEs can evolve rapidly and adapt to network conditions, increasing the complexity of detection and mitigation.

To address these challenges, the implementation of an AI-based Anomaly Detection (AD) system integrated directly with the UPF provides a highly effective and intelligent security mechanism for protecting the N6 interface. By continuously analyzing user-plane traffic behavior, flow characteristics, traffic distribution patterns, packet rates, and UE activity profiles, the AI-based detection system can identify abnormal traffic conditions associated with DoS and DDoS attacks in real time. Machine learning algorithms enable the system to learn

normal traffic behavior of the slice and dynamically detect deviations that indicate malicious activity, even when the traffic appears protocol compliant.

The integration of the anomaly detection framework with the UPF is particularly important because it enables immediate and automated mitigation actions. Once malicious UEs are identified, the UPF can directly enforce traffic blocking, flow rejection, or session termination policies against the attacking UEs, preventing the malicious traffic from propagating further into the network or impacting other slices. This closed-loop security mechanism significantly reduces attack response time and minimizes the risk of large-scale service degradation.

2.11.2.1 EXPERIMENTATION SETUP

This setup (Figure 52) is composed of three main functional blocks: CyPerf traffic generation, LoadCore 5G core emulation, and the N6 anomaly detection platform. The objective is to generate malicious traffic flows, transport them through a realistic 5G user-plane environment, detect anomalies at the N6 interface, and trigger mitigation actions at the UPF.

At the top of the architecture, CyPerf is used as the traffic generation platform [35]. Two CyPerf agents emulate the client and server sides of the communication. The DUT network between the two agents is intentionally disabled, meaning CyPerf is primarily being used as a controlled attack and traffic emulation source.

The generated traffic is injected into the LoadCore environment, which emulates a complete 5G core network. The setup includes multiple 5G network functions such as AMF, AUSF, UDM, **UDR**, PCF/PCRF, SMF/PGW-C, and UPF/PGW-U. This allows the malicious traffic to traverse a realistic 5G architecture before reaching the external network.

The UE is configured with passthrough mode and user-plane functionality enabled. The traffic is forwarded through the emulated RAN/gNodeB. From the RAN, the traffic reaches the UPF through the N3 interface. The UPF acts as the user-plane anchor and enforcement point within the architecture. The N6 interface is the critical observation point in this setup because it represents the connection between the 5G core and the external data network. Since all user-plane traffic exits the core through N6, monitoring this interface enables visibility into the behavior of UE traffic after processing by the 5G network.

The anomaly detection system is deployed on a dedicated Ubuntu VM connected to the N6 traffic path. The VM hosts several components:

- Collector
- Detector
- Kafka [36]
- Kafka-monitor [37]

The **Collector** captures traffic or telemetry data from the N6 interface and forwards the information through **Kafka** for processing. The **Detector** analyzes the traffic behavior and searches for anomalies such as malicious flows, abnormal traffic rates, suspicious sessions, or attack signatures generated by CyPerf.

The overall workflow begins with CyPerf generating malicious traffic that is transported through the LoadCore 5G infrastructure as if it were legitimate UE traffic. The traffic traverses the RAN and UPF before exiting through the N6 interface. At this point, the monitoring platform captures and analyzes the traffic in real time.

When the detector identifies anomalous behavior, it informs the UPF through the defined control or policy mechanism. The UPF then applies mitigation actions against the malicious UE. Depending on the implementation, these actions may include traffic blocking, session termination, policy enforcement, or flow filtering. As a result, malicious traffic generated by the identified UE can be prevented from reaching the external network.

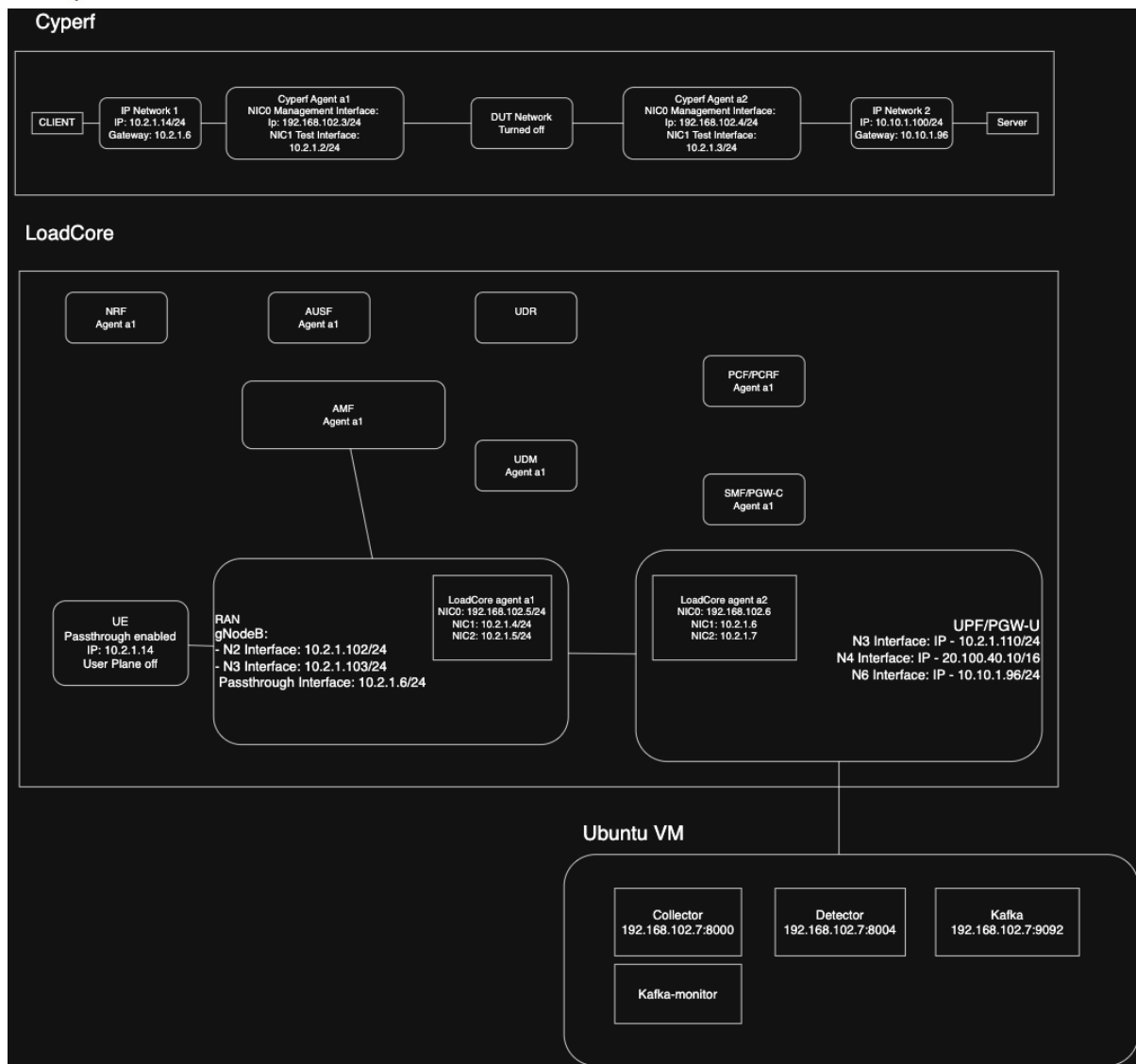


Figure 52: Experimentation setup for Anomaly Detection on the N6 Interface

2.11.2.2 USE CASE SELECTION AND TRAFFIC PROFILE DESCRIPTION

Autonomous driving was selected as the use case for validating the anomaly detection mechanism at the N6 interface because it represents a highly critical 5G application. Connected vehicles depend on reliable and secure low-latency communication, where abnormal or malicious traffic can directly impact service continuity and safety.

With the number of connected and autonomous vehicles growing on the road, the need to have communication infrastructure that supports them has grown with it. Vehicle-to-Everything (V2X) communication enables vehicles to exchange the information real-time with other vehicles, infrastructure, pedestrians and network. This communication creates modern Intelligent Transportation System (ITS). One communication type, Vehicle-to-Network, has enabled ITS to communicate with services that extend beyond local direct communication [38]. With V2N communication cars are able to receive direct communication from network. The technology enables us to drive vehicles remotely, download software updates over the air, manage traffic real-time, etc.

The safety critical nature of V2X communication brings low latency requirements on many of its traffic classes. ETSI ITS cooperative messages must be delivered within 100 milliseconds to remain operationally meaningful. In case of remote or tele-operated driving scenario latency requirements are more rigid, requiring less than 20 milliseconds of delay. The constraints introduce a fundamental vulnerability: Denial of Service or Distributed

Denial of Service Attacks. The attack does not need to bring down whole infrastructure; it just needs to push latency past the threshold at which communication becomes useless. This makes DoS/DDoS uniquely and disproportionately dangerous in V2X environments compared to conventional networks.

Since the N6 interface carries user-plane traffic between the 5G core and external application servers, it provides an ideal observation point to detect suspicious behavior affecting autonomous driving services. This makes the scenario suitable for evaluating real-time anomaly detection and UPF-based mitigation under realistic and mission-critical network conditions.

Table 12 provides a consolidated bitrate reference across all traffic classes. Note that in practice, many of these flows are concurrent: a single vehicle in a remote driving scenario simultaneously generates sensor data (UL), receives control commands (DL), streams video (UL), and exchanges CAMs (both)

Table 12 - Consolidated traffic profile for the autonomous drive use case [42] [43] [44] [45] [46] [47] [48]

Traffic Class	Protocol	Min Bitrate	Max Bitrate	Direction
ETSI ITS Messages (all)	UDP/BTP	1.2 kbps	240 kbps	Both
↳ CAM only	UDP/BTP	0.8 kbps	64 kbps	Both
↳ DENM only	UDP/BTP	2.4 kbps	32 kbps	Both
↳ CPM only	UDP/BTP	2.4 kbps	240 kbps	Both
↳ VAM only	UDP/BTP	1.2 kbps	32 kbps	Both
Remote Driving – Control	UDP	40 kbps	1,200 kbps	Downlink
Remote Driving – Video	UDP/RTP	2 Mbps	10 Mbps	Uplink
Remote Driving – Sensors	UDP	8 kbps	8 kbps	Uplink
OTA / PKI / Certs	TCP/TLS	~1 MB/session	2 GB+/session	Downlink
FCD / Probe Upload	TCP/MQTT	< 1 kbps avg	~10 kbps	Uplink
MEC Safety Warnings	TCP/MQTT	< 5 kbps	~50 kbps	Downlink
HD Map Upload	TCP/HTTPS	100 kbps burst	10+ Mbps burst	Uplink

2.11.2.3 ANOMALY DETECTION DESIGN AND MODELS EVALUATION

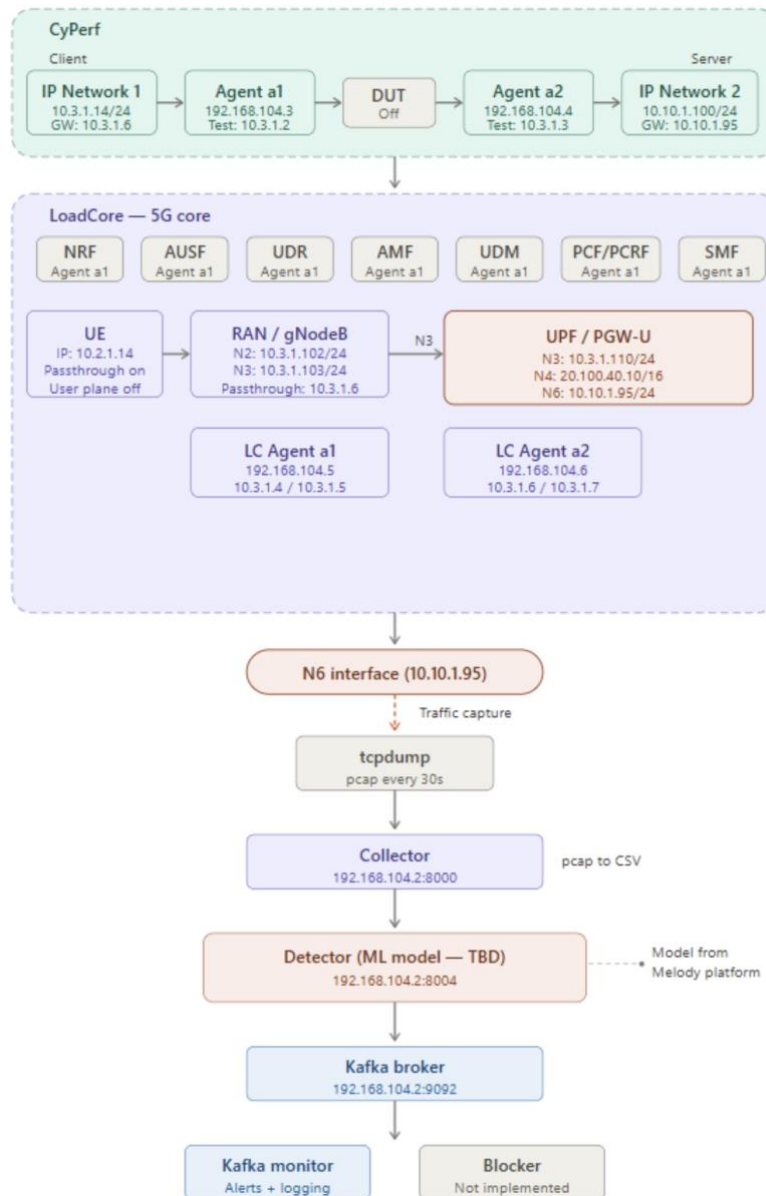


Figure 53: Full pipeline for Anomaly Detection in the N6 Interface

Figure 53 shows the full pipeline for detecting Anomalies in the N6 interface. Data generation is handled by both CyPerf and LoadCore to create realistic 5G network traffic and attacks. The Collector converts each pcap file into CSV format containing network flow records and forwards the CSV to the Detector. Each row in the CSV represents one network flow and the CSV includes extracted flow statistics that serve as input features for the ML model.

Flow-level features are extracted from the CSV data. These features are what the ML model uses to distinguish normal traffic from attacks. The feature set includes:

- **Volume features:** total forward packets, total backward packets, total bytes forward, total bytes backward, packets per second, bytes per second, flow bytes per second.
- **Timing features:** flow duration, forward inter-arrival time (mean, min, max, std), backward inter-arrival time (mean, min, max, std), active time, idle time.
- **Protocol features:** TCP SYN flag count, TCP ACK flag count, TCP FIN flag count, TCP RST flag count, TCP PSH flag count, SYN/ACK ratio, protocol type distribution.
- **Entropy features:** source IP entropy per time window, destination port entropy, packet size entropy.

- **Packet size features:** packet length mean, min, max, std (forward and backward), payload size distribution.

The Features are normalized before being fed to the model to ensure consistent scale across different feature types.

Model training is split in 3 steps as follows.

- Step 1 - Dataset upload: Labeled CSV datasets are uploaded to dataset management component. Each dataset is tagged with its type and version (e.g., "v2n_benign_v1", "v2n_udpflood_v1").
- Step 2 - Training: The algorithm used on this anomaly detection is the Iso Forest.
- Step 3 - Threshold/decision boundary setting: After training, the model's decision boundary or anomaly threshold is calibrated using validation data.

Models are evaluated offline, which tracks each experiment's configuration, parameters, logs, and results. The used evaluation metrics are:

- Accuracy: overall percentage of correct classifications
- Precision: percentage of flagged flows that are actually malicious
- Recall: percentage of actual attacks that are detected
- F1-score: harmonic mean of precision and recall
- Detection latency: time from attack start to first alert
- False positive rate: percentage of benign flows incorrectly flagged

The trained model is exported from offline environment and deployed into the existing detection pipeline.

2.11.2.4 RESULTS

An Isolation Forest model is trained exclusively on 15 minutes of benign V2N traffic (301 samples, 25 features, 30-second time windows). During testing, the model flags any traffic that deviates significantly from the learned baseline. Raw pcap captures are converted to CSV flow features using a custom Python script that computes 25 features per flow per time window, including volume metrics, timing statistics, packet size distribution, TCP flag counts, and source IP entropy. From this, the model identifies two possible types of traffic:

- **Benign traffic:** This is generated by 10 UEs generating V2N telemetry using Keysight LoadCore with Stateless UDP: 300-byte payloads at 10 packets per second per vehicle, based on 3GPP TS 22.186. Traffic flows through the emulated 5G core (UE → gNB → AMF → SMF → UPF) and exits at the N6 interface.
- **Attack traffic:** Synthetic UDP flood, SYN flood, and Slowloris attacks are generated with varying intensities to test the detector's ability to distinguish between normal V2N traffic and DoS attacks.

The testing, the results of which can be seen in Figure 54 and Figure 55, shows that the model reaches a 99,5% accuracy with a false positive rate of only 1%

```

=== Overall Results ===

Benign: 301 | UDP flood: 301 | SYN flood: 301 | Slowloris: 301

      precision    recall  f1-score   support

  Benign         1.00      0.99      0.99         301
  Attack         1.00      1.00      1.00         903

 accuracy         1.00
 macro avg         1.00      1.00      1.00         1204
weighted avg         1.00      1.00      1.00         1204

Per-type detection:
Benign correct:    298/301 (99.0%)
UDP flood detected: 301/301 (100.0%)
SYN flood detected: 301/301 (100.0%)
Slowloris detected: 301/301 (100.0%)

Anomaly scores (lower = more anomalous):
Benign:    mean=-0.3302, min=-0.2061
UDP flood: mean=-0.0325, min=-0.1186
SYN flood: mean=-0.0621, min=-0.1304
Slowloris: mean=-0.0407, min=-0.1796

```

Figure 54: Overall results of the Anomaly Detection testing at the N6 Interface

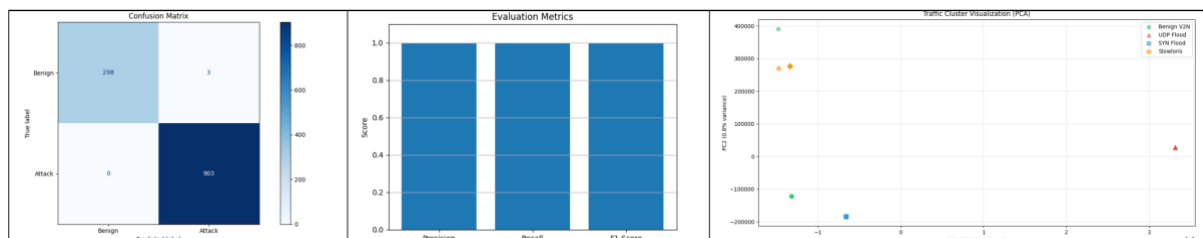


Figure 55: Graphical representations of the results of the Anomaly Detection testing at the N6 Interface

2.12 KPIs RELATING TO SECURITY AND CONFORMANCE TESTING

This Section contains KPIs relating to Security and Conformance testing:

- Security Conformance (**SEC**)

The security conformance testing KPIs are based on NF conformity with the 3GPP specification and in general those tests are referred to as Security Assurance Specification (SCAS) Tests. Within the auspices of 6G-SANDBOX those tests are aligned with the Release 18 of TS 33.512 (AMF), TS33.514 (NRF), and TS33.518 (UDM).

The key aspects of 5G SCAS testing are the following:

- **Threat Modeling and Analysis:** SCAS testing begins with a detailed threat modeling process, identifying potential security risks specific to 5G networks, including threats from malicious actors, system vulnerabilities, and potential attack vectors.
- **Security Architecture Review:** Involves a thorough examination of the 5G network architecture, including core network elements, virtualized infrastructure, and network slicing capabilities, to ensure adherence to security best practices and standards.
- **Protocol and Interface Testing:** SCAS testing evaluates the security protocols and interfaces used within 5G networks, such as authentication mechanisms, encryption protocols (e.g., AKA, EAP), and network interfaces (e.g., between radio access and core networks), to verify their effectiveness against intrusion attempts and data breaches.
- **Performance and Resilience Testing:** Beyond security protocols, SCAS testing assesses the performance and resilience of security measures under various operational conditions, including high traffic volumes, network congestion, and during emergency situations, ensuring continuous protection without compromising network performance.

- Compliance and Certification: SCAS testing ensures compliance with international security standards and regulatory requirements specific to 5G networks, facilitating the certification process for network operators and service providers.

2.12.1 TEST SETUP

The execution of SCAS testing made use of the following tools: OpenTAP Test Automation, LoadCore 5G Core Network Simulator, and SCAS Impairments Source Code [33] [46]. Open5GS was used as the testing target [47].

For the purposes of the test, a LoadCore configuration is established where the AMF (Access and Mobility Management Function) is configured as the DUT (Device Under Test). In this setup, the AMF is isolated to ensure focused testing. LoadCore is responsible for generating traffic, which originates from User Equipment (UEs) and the gNodeB (gNB). This configuration allows for a controlled environment where the performance and behavior of the AMF can be accurately assessed under various traffic conditions and with a special focus on security.

The SCAS testing set features many different sub-tests that test the security of individual network components. In the following Table 13, Table 14 and Table 15, these sub-tests and their requirements are described for the AMF, NRF and UDM respectively. For each component, a portion of these sub-tests were carried out.

Table 13 - Security Assurance Specification (SCAS) Testing for the Access and Mobility Management Function (AMF). Reference: TS33.512.

NF	Category	Sub Category	Sub Reference	Testing Name/Reference	Requirement Description
AMF-Specific Security Requirements and Related Test Cases	Authentication and Key Agreement Procedure	Synchronization Failure Handling	4.2.2.1.1 (A, B)	Synchronization failure handling	Upon receiving an authentication failure message with synchronisation failure (AUTS) from the UE, the SEAF sends an ausf_UEAuthentication_Authenticate Request message with a "synchronisation failure indication" to the AUSF.
			4.2.2.1.1 (A)	AUSF responds to Authentication message form DUT (AMF)	
			4.2.2.1.1 (B)	AUSF doesn't respond to Authentication message form DUT (AMF)	
		RES* Verification Failure Handling	4.2.2.1.2 (A, B, C, D)	RES* verification failure handling	The SEAF shall proceed with step 10 in Figure 6.1.3.2-1 and after receiving the Nausf_UEAuthentication_Authenticate Request message from the AUSF in step 12, proceed as described below: (1) if the AUSF has indicated in the Nausf_UEAuthentication_Authenticate Response message to the SEAF that the verification of the RES* was not successful in the AUSF, or (2) if the verification of the RES* was not successful in the SEAF, then the SEAF shall either reject the authentication by sending an Authentication Reject to the UE if the SUCI was used by the UE in the initial NAS message or the SEAF/AMF shall initiate an Identification procedure with the UE if the 5G-GUTI was used by the UE in the initial NAS message to retrieve the SUCI and an additional authentication attempt may be initiated. Also, (3) if the SEAF does not receive any Nausf_UEAuthentication_Authenticate Request message from the AUSF as expected, then the SEAF shall either reject the authentication to the UE or initiate an Identification procedure with the UE.
			4.2.2.1.2 (A)	The UE sends RR with SUCI to the DUT (AMF) and incorrect RES*	
			4.2.2.1.2 (B)	The UE sends RR with 5G-GUTI to the DUT (AMF) and incorrect RES*	
			4.2.2.1.2 (C)	The UE sends RR with SUCI to the DUT (AMF) and RES*	
			4.2.2.1.2 (D)	The UE sends RR with 5G-GUTI to the DUT (AMF) and RES*	
	N/A	Security Mode Command Procedure	4.2.2.3.1	Replay protection of NAS signalling messages	AMF shall support replay protection of NAS signalling messages between UE and AMF on N1 interface." as specified in TS 33.501 [2], clause 5.5.1
	N/A		4.2.2.3.2	NAS NULL integrity protection	NIA0 shall be disabled in AMF in the deployments where support of unauthenticated emergency session is not a regulatory requirement.
	N/A		4.2.2.3.3	NAS integrity algorithm selection and use	The AMF shall then initiate a NAS security mode command procedure, and include the chosen algorithm and UE security capabilities (to detect modification of the UE security capabilities by an attacker) in the message to the UE (see sub-clause 6.7.2 of the present document). The AMF shall select the NAS algorithm which have the highest priority according to the ordered lists." as specified in TS 33.501 [2], clause 5.5.2.
	N/A	Security in Intra-RAT Mobility	4.2.2.4.1	Bidding down prevention in Xn-handover	In the Path-Switch message, the target gNB shall send the UE's 5G security capabilities received from the source gNB to the AMF. The AMF shall verify that the UE's 5G security capabilities received from the target gNB are the same as the UE's 5G security capabilities that the AMF has locally stored. If there is a mismatch, the AMF shall send its locally stored 5G security capabilities of the UE to the target gNB in the Path-Switch ETSI 3GPP TS 33.512 version 16.3.0 Release 16 14 ETSI TS 133 512 V16.3.0 (2020-08) Acknowledge message. The AMF shall support logging capabilities for this event and may take additional measures, such as raising an alarm.
			4.2.2.4.2	NAS protection algorithm selection in AMF change	If the change of the AMF at N2-Handover or mobility registration update results in the change of algorithm to be used for establishing NAS security, the target AMF shall indicate the selected algorithm to the UE as defined in Clause 6.9.2.3.3 for N2-Handover (i.e., using NAS Container) and Clause 6.9.3 for mobility registration update (i.e., using NAS SMC). The AMF shall select the NAS algorithm which has the highest priority according to the ordered lists (see sub-clause 6.7.1.1 of the present document).
	N/A	5G-GUTI allocation	4.2.2.5.1 (A, B, C)	5G-GUTI allocation	A new 5G-GUTI shall be sent to a UE only after a successful activation of NAS security. The 5G-GUTI is defined in TS 23.003 [4]. Upon receiving Registration Request message of type "initial registration" or "mobility registration update" from a UE, the AMF shall send a new 5G-GUTI to the UE during the registration procedure. Upon receiving Registration Request message of type "periodic registration update" from a UE, the AMF should send a new 5G-GUTI to the UE during the registration procedure. Upon receiving Service Request message sent by the UE in response to a Paging message, the AMF shall send a new 5G-GUTI to the UE. This new 5G-GUTI shall be sent before the current NAS signalling connection is released. NOTE 1: It is left to implementation to re-assign 5G-GUTI more frequently than in cases mentioned above.
			4.2.2.5.1 (A)	5G-GUTI allocation	
			4.2.2.5.1 (B)	5G-GUTI allocation	
			4.2.2.5.1 (C)	5G-GUTI allocation	
	N/A	Security in Registration Procedure	4.2.2.6.1	Invalid or unacceptable UE security capabilities handling	If the REGISTRATION REQUEST message is received with invalid or unacceptable UE security capabilities (e.g. no 5GS encryption algorithms (all bits zero), no 5GS integrity algorithms (all bits zero), mandatory 5GS encryption algorithms not supported or mandatory 5GS integrity algorithms not supported, etc.), the AMF shall return a REGISTRATION REJECT message.

Table 14 - Security Assurance Specification (SCAS) Testing for the Network Repository Function (NRF). Reference: TS33.518.

NF	Category	Sub Category	Sub Reference	Testing Name/Reference	Requirement Description
NRF-Specific Security Requirements and Related Test Cases	N/A	NF Discovery Procedure	4.2.2.2.1	NF discovery authorization for specific slice	Upon receiving an authentication failure message with synchronisation failure (AUTS) from the UE, the SEAF sends an Nausf_UEAuthentication_Authenticate Request message with a "synchronisation failure indication" to the AUSF and the AUSF sends an Nudm_UEAuthentication_Get Request message to the UDM/ARPF, together with the following parameters: - RAND sent to the UE in the preceding Authentication Request, and - AUTS received by the SEAF in the response from the UE to that request, as described in clause 6.1.3.2.0 and 6.1.3.3.1. An SEAF will not react to unsolicited "synchronisation failure indication" messages from the UE. The SEAF does not send new authentication requests to the UE before having received the response to its Nausf_UEAuthentication_Authenticate Request message with a "synchronisation failure indication" from the AUSF (or before it is timed out). " as specified in TS 33.501[7], clause 6.1.3.3.2.

Table 15 - Security Assurance Specification (SCAS) Testing for the Unified Data Management (UDM). Reference: TS33.514.

NF	Category	Sub Category	Sub Reference	Testing Name/Reference	Requirement Description
UDM-Specific Security Requirements and Related Test Cases	N/A	User Privacy Procedure	4.2.1.1	De-concealment of SUPI from the SUCI based on the protection scheme used to generate the SUCI	The SIDF shall resolve the SUPI from the SUCI based on the protection scheme used to generate the SUCI." as specified in TS 33.501 [2], clause 5.8.2.
	N/A	Authentication and Key Agreement Procedure	4.2.2.1	Synchronization failure handling	When the UDM/ARPF receives an Nudm_UEAuthentication_Get Request message with a "synchronisation failure indication" it acts as described in TS 33.102 [9], clause 6.3.5 where ARPF is mapped to HE/AuC. The UDM/ARPF sends an Nudm_UEAuthentication_Get Response message with a new authentication vector for either EAP-AKA' or 5G-AKA depending on the authentication method applicable for the user to the AUSF, as specified in TS 33.501 [2], clause 6.1.3.3.2.
	N/A		4.2.2.2	Storing of authentication status of UE by UDM	The UDM shall store the authentication status of the UE (SUPI, authentication result, timestamp, and the serving network name) after authentication" as specified in TS 33.501 [2], clause 6.1.4.1a.

2.12.2 RESULTS

Table 16 presents the results for every test case executed on the Open5GS deployment. Passes are marked green whereas fails are marked in red, with a reasoning for the failure given in the final column. The results show that the AMF and NRF implemented on Open5GS and therefore on 6G-SANDBOX are not fully compliant with the 3GPP SCAS. Most of the errors are related to miss-implementation and lack of software that implements certain expected behaviour.

Table 16 - Results of all carried out SCAS tests on the AMF, NRF and UDM of the Open5GS Instance

Network Function	Document	Test Identification (Document Section)	Success Criteria	Failure Reasons
AMF	TS 33.512	4.2.2.1.1(A)	FAILED	AMF did not respond to SyncFailure from the UE
AMF	TS 33.512	4.2.2.1.1(B)	PASSED	
AMF	TS 33.512	4.2.2.1.2(A)	FAILED	Authentication failure in AUSF to AMF (response to

				Nausf_UEAuthentication_Authenticate) is not present
AMF	TS 33.512	4.2.2.1.2(C)	FAILED	Authentication failure in AUSF to AMF (response to Nausf_UEAuthentication_Authenticate) is not present
AMF	TS 33.512	4.2.2.1.2(E)	PASSED	
AMF	TS 33.512	4.2.2.3.1	PASSED	
AMF	TS 33.512	4.2.2.3.3	FAILED	The algorithm with high priority is not supported by the SecurityModeCommand.', 'output': {'SMCAAlgorithm
AMF	TS 33.512	4.2.2.6.1	PASSED	
AMF	TS 33.512	4.2.2.1.2(B)	FAILED	Registration Request using 5G-GUTI is not found in the trace.
AMF	TS 33.512	4.2.2.1.2(D)	FAILED	Registration Request using 5G-GUTI is not found in the trace.
AMF	TS 33.512	4.2.2.1.2(F)	PASSED	
AMF	TS 33.512	4.2.2.3.2(A)	FAILED	Authentication Request was not received by UE'
AMF	TS 33.512	4.2.2.4.1	FAILED	PathSwitchRequest frame does not found in the capture
AMF	TS 33.512	4.2.2.4.2	FAILED	HandOverRequest message is not present in the capture'
AMF	TS 33.512	4.2.2.5.1(A)	FAILED	Some 5G-TMSIs are duplicated by the AMF'
AMF	TS 33.512	4.2.2.5.1(B)	PASSED	
AMF	TS 33.512	4.2.2.5.1(C)	FAILED	Some 5G-TMSIs are duplicated by the AMF'
NRF	TS 33.518	4.2.2.2.1	FAILED	Normally to be compliant with 3gpp NRF should send back 403 Forbidden when the slices miss match, But the NRF sends 200 ok in both cases.
UDM	TS 33.514	4.2.1.1	PASSED	
UDM	TS 33.514	4.2.2.1	PASSES	
UDM	TS 33.514	4.2.2.2	PASSED	

2.13 KPIs RELATING TO AI/ML

This Section contains KPIs relating to AI/ML Validation:

- AI/ML Accuracy (**AMA**)
- AI/ML Effectiveness Improvement (**AME**)

The AI/ML KPI testing methodology for 6G-SANDBOX was developed during the second round of OCs at the Athens Testbed, during which one of the OCs was interested in testing a Federated Learning (FL) mechanism designed for accurate forecasting of energy consumption/generation for households through the integration of external data sources. This consisted of two main modules, one for the forecasting of generation and one for the forecasting of consumption. The FL mechanism would make use of both historical energy data alongside data collected from locally deployed weather stations and IoT agents in the households. The consumption

module endpoints collect only locally generated data, aggregating them into daily totals and training a separate ML-based forecasting model.

For the purpose of AI/ML validation within 6G-SANDBOX, a trial network is instantiated consisting of 3 main components:

- Open5GS to serve as the network core [47].
- The UERANSIM tool, which provides the trial network with multiple UEs and gNodeBs [48].
- A dedicated FL server

Two types of FL clients are integrated into the trial network. Existing clients (deployed on the experimenter's premises) have been configured to access the FL server via VPN over the Internet, while additional FL clients are initiated on the trial network, each served by a corresponding UE simulated through the use of UERANSIM. Besides the FL clients, a number of UEs are instantiated for the purpose of generating background traffic during experimentation. This trial network set up, in this case, permitted the validation of 5G's ability to accommodate the service of forecasting at a large scale.

Following proper model training, the results were subjected to an evaluation methodology. This included the calculation of metrics such as Mean Square Error (MSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE) and Symmetric Mean Absolute Percentage Error (SMAPE). These metrics compare the predicted result to the real result, thus providing the **AMA** KPI. These metrics are universal, and as such the methodology can be adapted for use to many other different AI/ML testing implementations.

The accuracy of an AI/ML model strongly depends on the availability of sufficient training data. In this case, from the available daily time series, 80 % of the samples are used as the training set; the remaining 20 % serve as the test/validation set. A supervised ML regression model is trained on the 80 % segment to predict next-day consumption given past daily totals (and implicitly the patterns reflected in them), and its performance is evaluated on the 20 % hold-out set using MAPE and SMAPE. Based on the findings when the same training procedure is repeated with progressively longer training sets (e.g. data representing a few months vs. a full year), the resulting MAPE decreases, demonstrating the positive relationship between the size of the training set and **AMA**. This process of modifying the scope of the training dataset and investigating its impact on **AMA** provides the **AME** KPI.

Figure 56 shows a graphical representation of the **AMA** KPI, comparing the actual results against predicted results as a measure of accuracy of the AI/ML model.

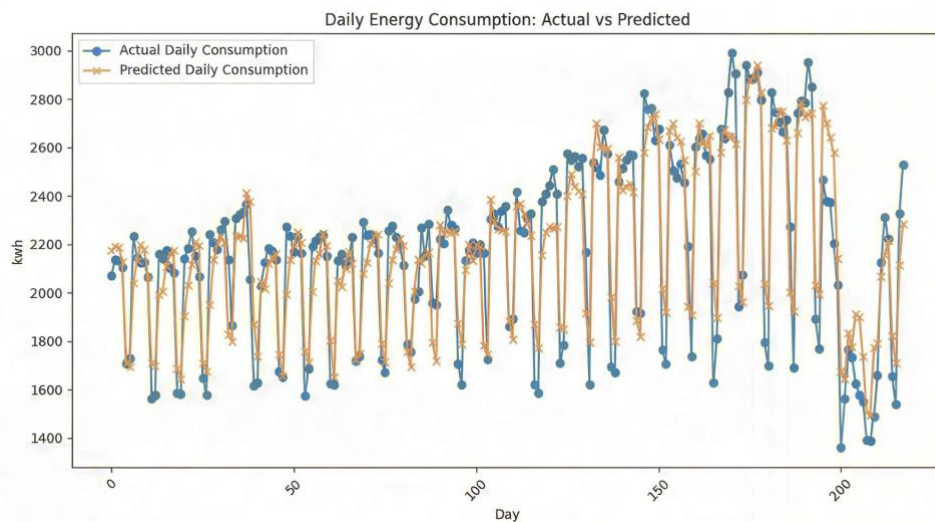


Figure 56 - Graph showing a comparison between Actual Energy Consumption vs Energy Consumption predicted by the AI/ML model in Athens AI/ML Experimentation

Figure 57 shows the Mean Absolute Percentage Error (MAPE) and its symmetric variant SMAPE which clearly indicates how the forecasting accuracy improves as the horizon of available data increases. This is a graphical representation of the **AME** KPI

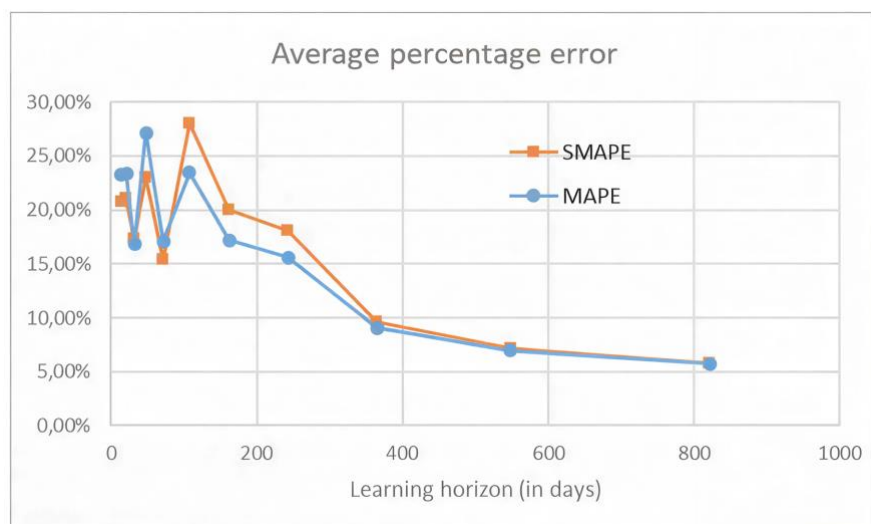


Figure 57 - Graph showing progression of Average Percentage Error as the amount of training data available to the AI/ML model was increased in Athens AI/ML Experimentation

3 REPORT ON CONTINUATION OF BENCHMARKING EFFORTS

Deliverable 5.3 covered some of the initial KPI measurement efforts in 6G-SANDBOX, most of which were then put into a context of benchmarking/calibrating the four platforms [49]. This was done by deploying the same Reference Trial Network (RTN) on each of the four different platforms with identical configurations and running a series of tests to generate a block of KPIs. Ideally, the four platforms would show if not identical then comparable results for each of these, confirming a base level of functionality that could be used in the future to validate platform performance. As Deliverable D5.3 showed, this was not the case, with the platforms showing significant performance deltas in all tested KPIs. Not only was this deemed unacceptable, but it was also an extremely surprising result, meriting further investigation.

3.1 PHASE 2

The initial results collected in Deliverable D5.3 were the result of the deployment of 4 Reference Trial Networks with identical parameters and allocations, but these deployments were still done essentially by hand. At the point in which D5.3 was being produced and the results generated, the TNLCM Trial Network deployment infrastructure was not yet fully operational and could not be used for the deployment effort. Therefore, engineers at each platform deployed the Reference Trial Network manually and without all of the infrastructure consistency that comes with TNLCM. For Phase 2 of the benchmarking investigation, TNLCM was ready and operational at all platforms and was used to deploy the Reference Trial Network at all four platforms. The full battery of tests described in D5.3 were performed again, on the Reference Trial Network with the exact same deployment parameters as described in D5.3 as well. The results for this second phase of benchmarking are shown in the tables below. Once again, for each KPI, the best performing platform is chosen as the Baseline, and any results that improve over the baseline are shown in **green**. For KPIs that were measured with varying amounts of UEs, the delta shown is the average delta across all User Equipment (UE) amounts. For ease of reference, the measured KPIs are as follows:

- Peak Data Rate (**PDR**)
- User Experienced Data Rate (**UDR**)
- Packet Loss Rate (**PLR**)
- User Plane Latency (**UPL**)
- Delay Deviation/Jitter (**DDJ**)
- Control Plane Latency (**CPL**)

Table 17 - Benchmarking Phase 2 Results for **PDR** KPI

Platform	Uplink Delta %	Downlink Delta %
Athens	49,34	39,24
Berlin	Baseline	
Malaga	37,25	9,46
Oulu	8,58	3,82
Overall Phase 1 Average	47,72	64,97
Overall Phase 2 Average	31,72	17,51

Table 18 - Benchmarking Phase 2 Results for **UDR** KPI

Platform	Uplink Delta %	Downlink Delta %
Athens	57,97	51,39
Berlin	Baseline	
Malaga	35,59	17,7
Oulu	53,82	52,02
Overall Phase 1 Average	0,56	14,07
Overall Phase 2 Average	49,13	40,37

Table 19 - Benchmarking Phase 2 Results for **PLR** KPI

Platform	Uplink Delta %	Downlink Delta %
Athens	45,79	42,6
Berlin	7,09	13,44

Malaga	Baseline	
Oulu	25,67	16,4
Overall Phase 1 Average	7,43	73,86
Overall Phase 2 Average	26,18	24,15

Table 20 - Benchmarking Phase 2 Results for **UPL** KPI

Platform	Uplink Delta %	Downlink Delta %
Athens	21,4	4,97
Berlin	31,38	5,85
Malaga	Baseline	
Oulu	25,67	16,4
Overall Phase 1 Average	1842,32	6217,22
Overall Phase 2 Average	26,29	9,07

Table 21 - Benchmarking Phase 2 Results for **DDJ** KPI

Platform	Uplink Delta %	Downlink Delta %
Athens	30,00	5,41
Berlin	Baseline	
Malaga	26,67	18,92
Oulu	16,67	10,81
Overall Phase 1 Average	3408,33	1505,56
Overall Phase 2 Average	24,45	11,71

Table 22 - Benchmarking Phase 2 Results for **CPL** KPI

Platform	Delta %
Athens	0,00
Berlin	0,00
Malaga	Baseline
Oulu	0,00
Overall Phase 1 Average	8,78
Overall Phase 2 Average	0,00

The efforts undertaken in Phase 2, the introduction of the TNLCM, the consolidation of all virtualization parameters and the standardization of the Vnet appear to have had a significant effect on the deltas between the platforms. In general, the truly massive deltas have completely disappeared, meaning all platforms are now at least comparable. In some cases, like for the three latency-based KPIs, the Phase 2 efforts have reduced the observed deltas dramatically, down to averages of 30% or lower. The deltas for the **PLR** show an improvement in downlink, but a widening of the performance gap in the uplink. Most interestingly, the deltas for the **UDR** appear to show a worsening performance differential compared to Phase 1.

Despite the general trend for deltas between Phase 1 and Phase 2 being downward, the continuing lack of parity has a few potential implications. Firstly, it would make the Reference Trial Network and the benchmarking methodology less useful as methods to validate potential new 6G-SANDBOX platforms. For example, based on

Table 19, a new platform showing a **PLR** result 45% worse than the baseline would have to be accepted, as these results show that to be within the expected level of variance. This could potentially obscure any underlying issues with said new platform. It also means that testing within 6G-SANDBOX is not platform agnostic – any continuation of a previous test campaign would have to be undertaken on the original platform it was executed on, or else all previous measurements would have to be retaken to account for the performance disparity.

It is with these two reasons in mind that the investigation was continued, with the goal to try and further the consortiums understanding of the underlying reasoning beneath the performance disparities and if they could be manipulated

3.2 PHASE 3

The testing described in Phase 2 constituted 4 deployments that, in terms of the virtualization layer, were as close to identical as was deemed to be possible. Therefore, any further investigations or parameter manipulation would need to go deeper, beyond the simple deployment parameters set within the TNCLM. Before such a campaign could be undertaken, it was necessary to understand the validity of the deltas measured in Phase 2. With the results as they are, it is impossible to discern what proportion of the deltas represent true disparity between the platforms and what proportion can be explained by the Measurement Uncertainty (MU) inherent in the measurement of the chosen test cases in the Reference Trial Network. In order to determine this MU for each KPI, the Reference Trial Network was deployed on three completely identical laptops which were configured as hosts within the TNCLM deployed in the Keysight Technologies premises in Belgium. The standard battery of benchmarking tests would be run on each deployment 6 times, and the variance in the results would be compared between all three. For each server, an average value for each KPI would be taken based on these 6 iterations. The MU would be presented as the coefficient of variation between each of the 3 hosts for a given KPI. If one of the platform parity measurement deltas fell below this uncertainty value, it could not be conclusively attributed to being caused by platform disparity. On the other hand, if a platform parity measurement delta was larger than the MU value, the differences in performance could still be stated to be due to platform disparity, and not due to an expected variance in the measurement.

Table 23 - Benchmarking Phase 3 Results for **PDR** KPI

	Uplink MU %	Downlink MU %
	2,18	2,72
Platform	Uplink Delta % - MU %	Downlink Delta % - MU %
Athens	47,16	36,52
Berlin	Baseline	
Malaga	35,07	6,74
Oulu	6,40	1,10
Averages		
Overall Phase 2 Average	31,72	17,51
Overall Phase 2 Average with MU Considered	29,54	14,79

Table 24 - Benchmarking Phase 3 Results for **UDR** KPI

	Uplink MU %	Downlink MU%
	0,45	10,51
Platform	Uplink Delta % - MU %	Downlink Delta % - MU %
Athens	57,52	40,88
Berlin	Baseline	
Malaga	35,14	7,19
Oulu	53,37	41,51

Averages		
Overall Phase 2 Average	49,13	40,37
Overall Phase 2 Average with MU Considered	48,68	29,86

Table 25 - Benchmarking Phase 3 Results for **PLR** KPI

	Uplink MU %	Uplink MU %
	1,83	1,75
Platform	Uplink Delta % - MU %	Downlink Delta % - MU %
Athens	43,96	40,85
Berlin	5,26	11,69
Malaga	Baseline	
Oulu	23,84	14,65
Averages		
Overall Phase 2 Average	26,18	24,15
Overall Phase 2 Average with MU Considered	24,35	22,40

Table 26 - Benchmarking Phase 3 Results for **UPL** KPI

	Uplink MU %	Downlink MU %
	18,14	13,55
Platform	Uplink Delta % - MU %	Downlink Delta % - MU %
Athens	3,26	0
Berlin	13,24	0
Malaga	Baseline	
Oulu	7,53	2,85
Averages		
Overall Phase 2 Average	26,29	9,07
Overall Phase 2 Average with MU Considered	8,15	0,95

Table 27 - Benchmarking Phase 3 Results for **DDJ** KPI

	Uplink MU %	Downlink MU %
	10,44	10,88
Platform	Uplink Delta % - MU %	Downlink Delta % - MU %
Athens	19,56	0

Berlin	Baseline	
Malaga	16,23	8,04
Oulu	6,23	0
Averages		
Overall Phase 2 Average	24,45	11,71
Overall Phase 2 Average with MU Considered	14,01	2,68

Table 28 - Benchmarking Phase 3 Results for **CPL** KPI

Platform	MU %
	0,98
Platform	Delta % - MU %
Athens	0,00
Berlin	0,00
Malaga	Baseline
Oulu	0,00
Averages	
Overall Phase 2 Average	0,00
Overall Phase 2 Average with MU Considered	0,00

For most of the KPIs in question, the measurement uncertainty determined by the Phase 3 testing is very low, in the sub 5% range. In these cases, the deltas determined over the course of Phase 2 can be said to still be valid and represent unexplained differences in performance between the seemingly identical virtual deployments of the 4 platforms. There are 5 KPIs where the MU is notably higher, those being the **UDR** in DL, **UPL** in both directions and the **DDJ** in both directions. All of these are in the 10% to 20% range and therefore have a greater impact on the ability to draw conclusions on their respective Phase 2 results. While the MU for **UDR** in downlink is quite high at 10,51%, the deltas drawn from Phase 2 are still relevant even when it is considered, being around 30% on average. The story is a little different for the **UPL** where both MU values are such that, when considered alongside the Phase 2 deltas, it becomes more difficult to say if these are still relevant deltas. A similar pattern can be seen in the **DDJ** results, albeit to a lesser extent.

It is worth noting that the combination of MU values with the Phase 2 deltas should be seen in an indicative sense, and not as exact realities. The differences in the hardware/setup between the MU tests and the Phase 2 testing are significant, however the Phase 3 MU results do give two clear conclusions: Firstly, that certain measured KPIs in the deployed RTNs have a significant level of variance even across identical hardware, which could be something that requires further investigation in the future and secondly, that the Phase 2 deltas (with the possible exception of the **UPL** deltas) still consist of valid representations of the difference in performance between the platforms and can be used as grounds to inform future work.

3.3 PHASE 4

The final phase of work undertaken in the benchmarking efforts within 6G-SANDBOX was an investigation into if certain underlying hardware factors could be manipulated to affect the performance of virtualized

deployments in predictable, systematic ways. The onus for this task was the results generated by Phase 2: if everything possible had been done in the virtual layer, the only way to try and bring the platforms closer to parity was to go deeper into factors below the virtual layer. 4 factors were chosen for this phase of the investigation:

- **Consistency:** While not technically a hardware factor, it was important to determine the consistency of generated results at a platform for two reasons, the first of which being to provide a good baseline of average performance for each KPI to compare against the results generated when modifying the other factors. Secondly, results here would also serve to inform the “trustworthiness” of a platform – if multiple tests in a row showed a consistent result for a given KPI, changes in the result of said KPI observed when modifying other factors could be definably said to be influenced by those modifications. If a platform showed a lack of consistency, this would become more difficult, and furthermore it may indicate an underlying issue with said platform.
 - To collect consistency results, the full suite of tests was run on the platform 8 times in a row. An average value for each KPI calculated based on these 8 datapoints. The RTN was set up with 20 UEs for all iterations
- **Interference:** Often referred to as “Noisy Neighbors” in this context, the presence of other VMs on the same host/machine as a target VM is known to be able to have a marked effect on the performance and consistency of the target VM. Work within Keysight has shown that this can occur even when the neighboring VMs are idle, which could pose significant issues within the context of 6G-SANDBOX.
 - To collect results about the effects of interference, different TNs/VMs were collocated with the RTN in different configurations, and the resultant KPIs from TC1-4 were collected. The RTN was set up with 20 UEs for these tests
- **CPU and Memory Allocation:** CPU core and amount of RAM are the two main resources granted to a VM upon deployment, and as such were obvious candidates for the Phase 4 investigation effort. The RTN has 3 constituent components and therefore 3 VMs to modify and investigate the results for.
 - To collect results about the effects of CPU and Memory allocation, first minimum values for both allocation types for all 3 VMs in the RTN were determined experimentally. Maximum values were set as 8 Cores and 16Gb for CPU and Memory allocation respectively. All test cases were then run on with each of VMs set to either minimum, maximum, or their default values. This was repeated for each VM, as well as a set of tests run for all of the VMs set to both their maximum and minimum configurations. The RTN was set up with 20 UEs for all of these tests.

Tests relating to these factors were performed at the Athens and Berlin platforms due to their availability. For the Adaptive Calibration Deltas in this section, deltas showing results better than the baseline are represented in **blue** and deltas showing results worse than the baseline are shown in **red**.

3.3.1 CONSISTENCY

As described above, consistency testing involved subjecting the RTN deployment at a given platform to 8 consecutive cycles of the full test set in a row, in order to both examine the consistency of the platform and, if said platform proved to be consistent, determine an average reference value for each of the KPIs. Both the Berlin and Athens platform participated in this testing, with notably different results. In the following tables, UL refers to the Uplink direction and DL to the Downlink direction.

Table 29 - Phase 4 Consistency Results for the Berlin Platform

KPI/Direction	Mean Value	Standard Deviation	Coefficient of Variation
PDR UL	1281,03Mbps	26,24	2,05%

PDR DL	1340,00Mbps	12,16	0,87%
UDR UL	32,29Mbps	0,29	0,90%
UDR DL	30,63Mbps	0,22	0,71%
PLR UL	6,70%	0,04	0,61%
PLR DL	5,48%	0,03	0,56%
UPL UL	108,78ms	2,83	2,60%
UPL DL	61,94ms	0,68	1,11%
DDJ UL	368,78us	26,01	7,05%
DDJ DL	431,33us	9,23	2,14%
CPL	0,31s	0,01	1,93%

Table 30 - Phase 4 Consistency Results for the Athens Platform

KPI/Direction	Mean Value	Standard Deviation	Coefficient of Variation
PDR UL	956,55Mbps	100,15	10,47%
PDR DL	1239,53Mbps	62,26	5,02%
UDR UL	20,39Mbps	3,41	16,74%
UDR DL	22,75Mbps	2,48	10,92%
PLR UL	7,79%	0,28	3,65%
PLR DL	6,36%	0,24	3,69%
UPL UL	88,02ms	8,35	9,49%
UPL DL	58,54ms	2,89	4,94%
DDJ UL	497,33us	39,08	7,86%
DDJ DL	486,6us	32,33	6,64%
CPL	0,32s	0,01	1,76%

In general, the Berlin platform shows better consistency than the Athens one in these results, with its highest Coefficient of Variation (COE) value being 7,05% for **DDJ** in Uplink. The COE values in Athens are higher across the board, even in KPIs where it outperforms Berlin, such as **UPL**. The difference is starkest in the throughput based KPIs, where Berlin shows a sub 1% COE while Athens sees values as high as 16%. This significant inconsistency in the Athens results makes it more difficult to draw conclusions from the data generated from it for the remainder of Phase 4.

3.3.2 INTERFERENCE

A number of different scenarios featuring different TN and VM configurations were deployed at each platform to act as interfering agents for the RTN. These were as follows:

- Scenario 1 (SC1): The RTN is joined by another instance of the RTN with the exact same parameters/allocation. The second RTN does not perform any actions and is idle throughout the entirety of the test cases being conducted on the original RTN.
- Scenario 2 (SC2): The RTN is joined by another instance of the RTN with the exact same parameters/allocation. The full suite of test cases is performed on both RTN instances.
- Scenario 3 (SC3): The RTN is joined by a “dummy” VM. Said VM uses a linux library (stress-ng) to artificially stress 38 (virtual, as hyperthreading is enabled by default) CPU cores while the test suite is performed on the RTN.
- Scenario 4 (SC4): The RTN is joined by a “dummy” VM which artificially generates 256Gbps worth of memory read/write load while the test suite is performed on the RTN

- Scenario 5 (SC5): The RTN is joined by a “dummy” VM which artificially stresses 6 CPU cores while the test suite is performed on the RTN. In this case, hyper-threading is turned off, meaning these are 6 distinct hardware cores.
- Scenario 6 (SC6): The RTN is joined by a “dummy” VM which artificially generates 256Gbps worth of memory read/write load while the test suite is performed on the RTN. In this case, hyper-threading is turned off.

In the following Table 31 and Table 32, the results for the main RTN are shown in the ACM Delta format: as percentage deltas from the respective platforms average ‘Baseline’ result. Figure 58 and Figure 59 show heatmaps of the data results, for ease of pattern identification and analysis. In Athens, due to restrictions in platform access, only SC1 and SC2 were performed.

Table 31 - Phase 4 Interference Results for the Berlin Platform

KPI/Direction	COV % Value	SC1 % Delta	SC2 % Delta	SC3 % Delta	SC4 % Delta	SC5 % Delta	SC6 % Delta
PDR UL	2,05	7,30	4,41	7,63	4,72	2,71	2,33
PDR DL	0,87	12,78	11,81	21,79	6,17	4,51	4,14
UDR UL	0,90	30,37	28,62	37,49	39,34	19,15	30,76
UDR DL	0,71	22,93	24,32	7,29	57,22	10,96	39,40
PLR UL	0,61	9,03	6,42	1,33	1,53	3,30	3,59
PLR DL	0,56	2,56	2,29	5,14	3,04	5,33	4,33
UPL UL	2,60	12,40	13,36	22,36	5,93	5,33	4,83
UPL DL	1,11	9,99	10,00	26,15	8,79	5,32	3,58
DDJ UL	7,05	5,44	23,22	55,78	14,97	5,94	5,99
DDJ DL	2,14	10,69	9,10	47,49	4,67	0,08	0,40
CPL	1,93	1,53	0,07	0,27	1,39	0,25	14,14

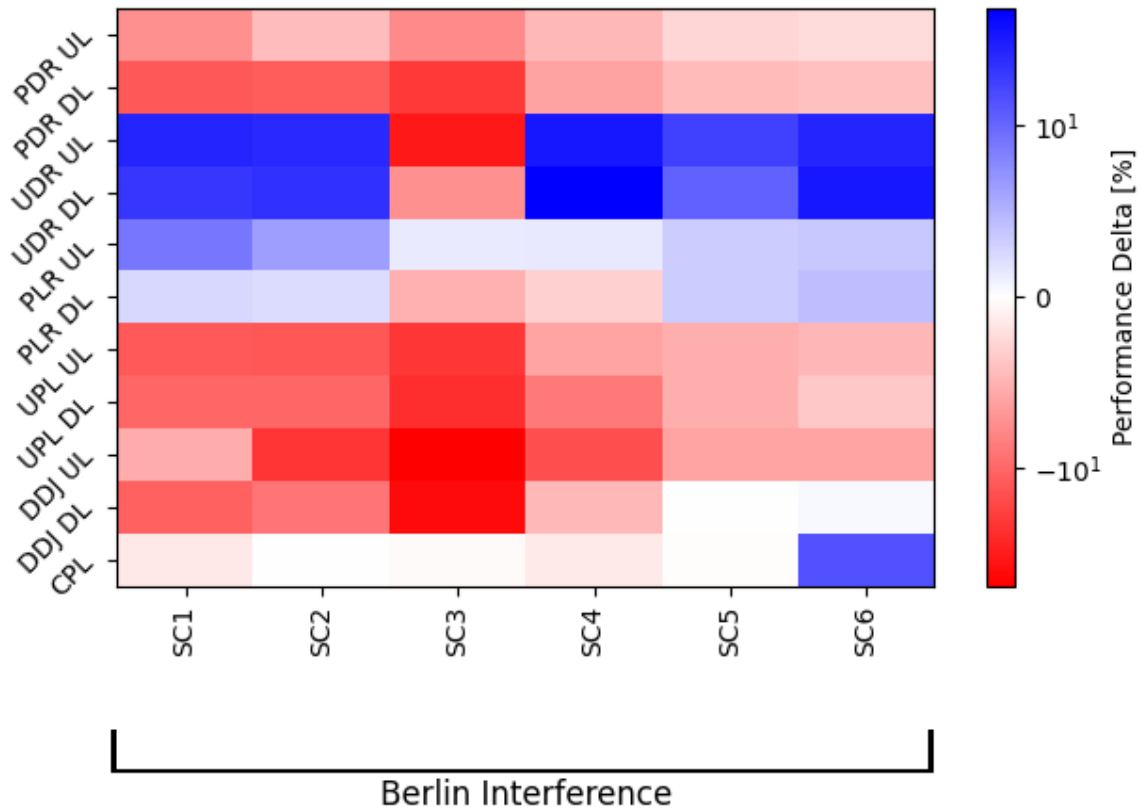


Figure 58 - Heatmap of Phase 4 Interference Berlin Deltas

Table 32 - Phase 4 Interference Results for the Athens Platform

KPI/Direction	COV % Value	SC1 % Delta	SC2 % Delta
PDR UL	10,47	11,92	19,47
PDR DL	5,02	3,03	2,97
UDR UL	16,74	7,17	1,16
UDR DL	10,92	6,98	3,24
PLR UL	3,65	0,42	3,42
PLR DL	3,69	7,62	1,46
UPL UL	9,49	21,53	24,98
UPL DL	4,94	9,72	13,52
DDJ UL	7,86	1,89	4,53
DDJ DL	6,64	6,97	13,15
CPL	1,76	1,33	0,53

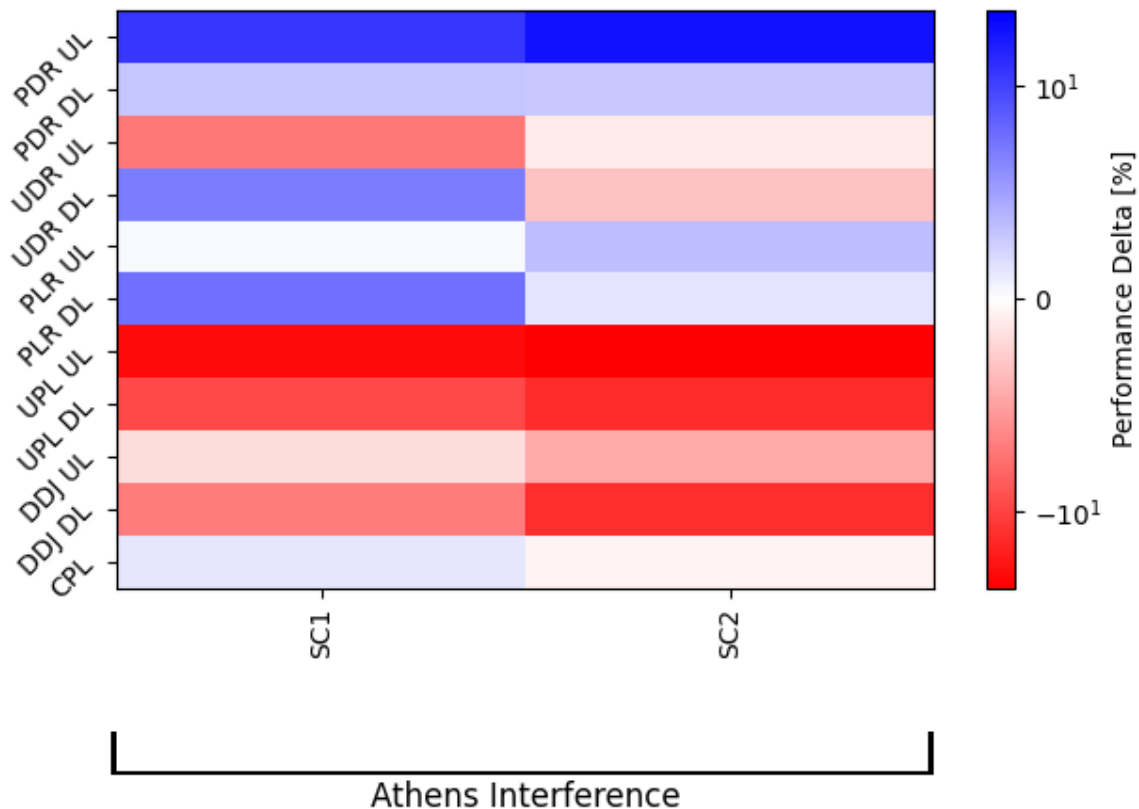


Figure 59 - Heatmap of Phase 4 Interference Athens Deltas

The results from Berlin are significantly more relevant for analysis in this case due to the significantly larger COV values in Athens. A lot of the deltas seen in Table 32 are impossible to draw conclusions from due to the fact that they are smaller or of a similar magnitude to the COV values. The **UPL** values are some of the only ones that exceed their COV counterparts and show a significant degradation in performance during both SC1 and SC2.

A lot more can be said about the results from Berlin seen in Table 31, where almost all the deltas exceed the relatively small COV values. Much like in Athens, the latency related KPIs see significant hits to their performance across the board, notably in SC3 with the heavy loading of the hyperthreaded CPU. SC3 in general seems to be the scenario that most heavily impacts a large majority of the KPIs in Berlin, with significant deltas across the board with the exception of **PLR** and **CPL**.

The most interesting result in Table 31 however is the results for the **UDR** and to some extent the **PLR**. Apart from the aforementioned SC3, **UDR** appears to show increased performance across the board when interference is present on the RTN's host machine. This performance increase is significant, with up to 57% for SC4 in the downlink direction. This is extremely unexpected – the addition of other VMs/TNs cohabitating with the RTN was not expected to make the RTN perform better.

As it stands at the time of writing, the consortium does not have a satisfactory explanation for this observed trend. Some early theories posited that it could be caused by power saving protocols within the host machine. The **UDR** can be described as a measure of minimum expected performance. With only one TN/VM deployed on a host machine, the host machine experiences a larger portion of downtime. These longer periods of downtime could cause power saving protocols on the host machine to kick in, which may result in the limiting of performance, especially at the lower end. Once more TNs/VMs are introduced onto the host machine, the

amount of downtime decreases as the average load on the host machine increases, which theoretically would discourage the engagement of power saving protocols and cause the bump in performance shown in the test results.

However, repeating the tests shown in this Section with all accessible power saving protocols disabled showed no appreciable difference compared to the previous results, meaning that if power saving protocols are the culprit behind the strange performance bumps noted in the results, said protocols are deeper than the ones that are accessible in BIOS options of a host machine. More research will be required to determine the exact cause of this behavior.

3.3.3 CPU ALLOCATION

To test the effects of CPU allocation, individual tests would be run with each of the 3 constituent VMs within the RTN set to a minimum value and a maximum value. The maximum was set to 8 allocated cores, while the minimum for each VM was determined experimentally as the minimum required for them to function within the RTN. While a given VM's allocation was changed, the other two VMs would retain their default values. Additionally, a test was run with all of the VMs allocated the maximum amount and another with the minimum. Table 33 shows the details in terms of CPU core allocation for each of the tested setups.

Table 33 - CPU Core Allocation Legend for CPU Core Allocation Testing

Setup ID	RAN VM Allocation	DN VM Allocation	Network Core VM Allocation
DN Min	4	1	4
DN Max	4	8	4
RAN Min	1	4	4
RAN Max	8	4	4
Network Core Min	4	4	1
Network Core Max	4	4	8
All Min	1	1	1
All Max	8	8	8

In the following Table 34 and Table 35, the results for the RTN are shown in the ACM Delta format: as percentage deltas from the respective platforms average 'Baseline' result. Figure 60 and Figure 61 show heatmaps of the data results, for ease of pattern identification and analysis.

Table 34 - Phase 4 CPU Allocation Results for the Berlin Platform

KPI/Direction	COV % Value	DN Min % Delta	DN Max % Delta	RAN Min % Delta	RAN Max % Delta	Core Min % Delta	Core Max % Delta	All Min % Delta	All Max % Delta
PDR UL	2,05	4,72	4,52	6,28	6,15	21,51	3,81	18,87	3,22
PDR DL	0,87	5,04	6,17	5,10	9,16	23,48	4,72	33,46	21,42
UDR UL	0,90	11,46	5,55	5,41	2,21	24,09	5,92	15,96	12,12
UDR DL	0,71	4,65	1,07	7,99	1,86	25,87	1,41	32,17	26,48
PLR UL	0,61	46,64	0,20	34,35	1,28	8,86	2,74	30,96	4,38
PLR DL	0,56	7,37	1,91	39,98	0,44	6,49	0,09	31,18	1,32
UPL UL	2,60	1,24	3,12	0,29	10,74	2,72	3,06	6,12	28,96
UPL DL	1,11	0,19	1,32	0,90	11,99	47,36	0,60	13,38	9,35
DDJ UL	7,05	9,00	4,67	8,47	5,21	125,07	4,40	102,83	2,23
DDJ DL	2,14	6,57	3,79	1,47	9,12	271,41	4,02	123,03	2,24

CPL	1,93	1,42	3,27	1,85	0,79	2,47	2,72	0,12	2,47
-----	------	------	------	------	------	------	------	------	------

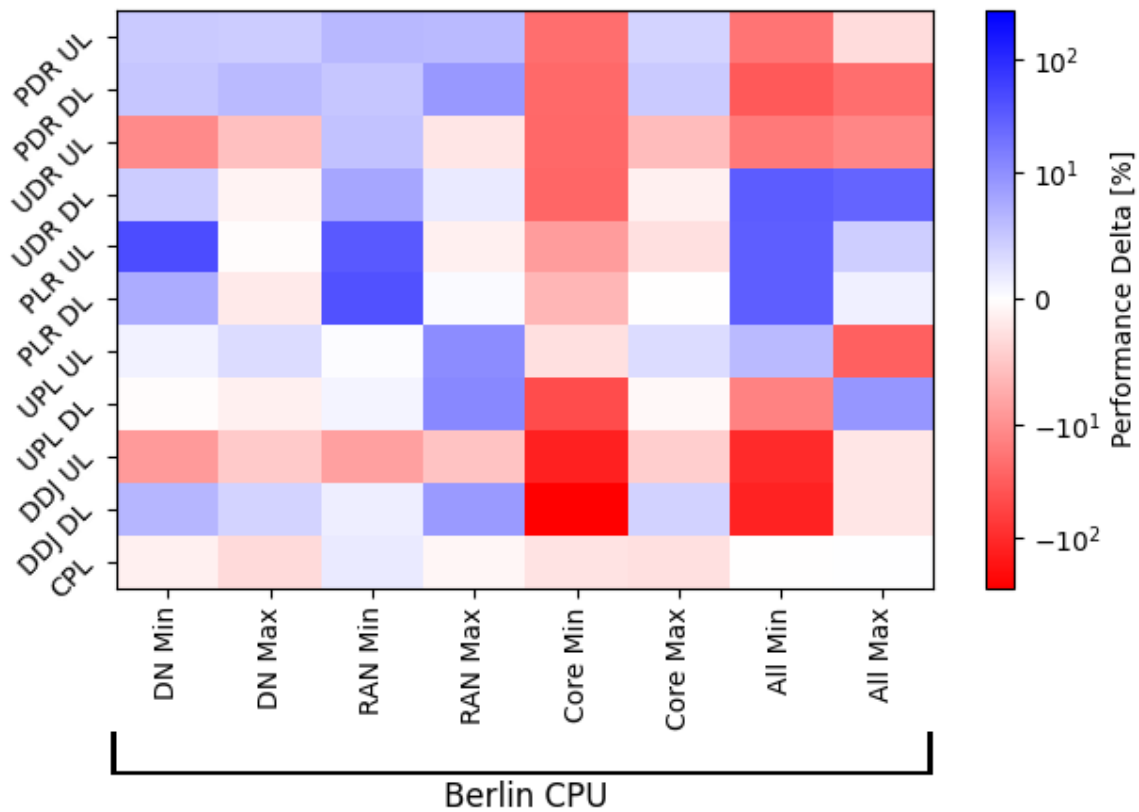


Figure 60 - Heatmap of Phase 4 CPU Allocation Berlin Deltas

Table 35 - Phase 4 CPU Allocation Results for the Athens Platform

KPI/Direction	COV % Value	DN Min % Delta	DN Max % Delta	RAN Min % Delta	RAN Max % Delta	Core Min % Delta	Core Max % Delta	All Min % Delta	All Max % Delta
PDR UL	10,47	14,33	8,03	7,69	2,81	40,99	10,43	46,75	5,74
PDR DL	5,02	14,12	0,41	2,72	6,31	43,76	4,80	52,09	0,38
UDR UL	16,74	2,82	4,57	6,15	0,00	8,24	18,89	18,52	6,37
UDR DL	10,92	3,05	0,70	12,94	10,00	14,22	4,93	3,12	5,57
PLR UL	3,65	49,39	1,02	33,81	2,73	6,01	2,09	28,22	2,29
PLR DL	3,69	3,69	1,66	37,68	1,61	1,76	0,69	47,58	1,43
UPL UL	9,49	5,37	1,22	0,49	2,87	15,71	6,54	16,29	1,83
UPL DL	4,94	5,29	3,10	4,90	2,67	67,56	1,75	77,40	1,16
DDJ UL	7,86	12,60	7,51	7,10	1,27	145,30	1,47	155,56	10,32
DDJ DL	6,64	24,33	3,21	0,33	2,75	292,31	0,53	314,30	8,34
CPL	1,76	0,91	1,61	0,38	2,92	2,23	2,11	2,10	2,45

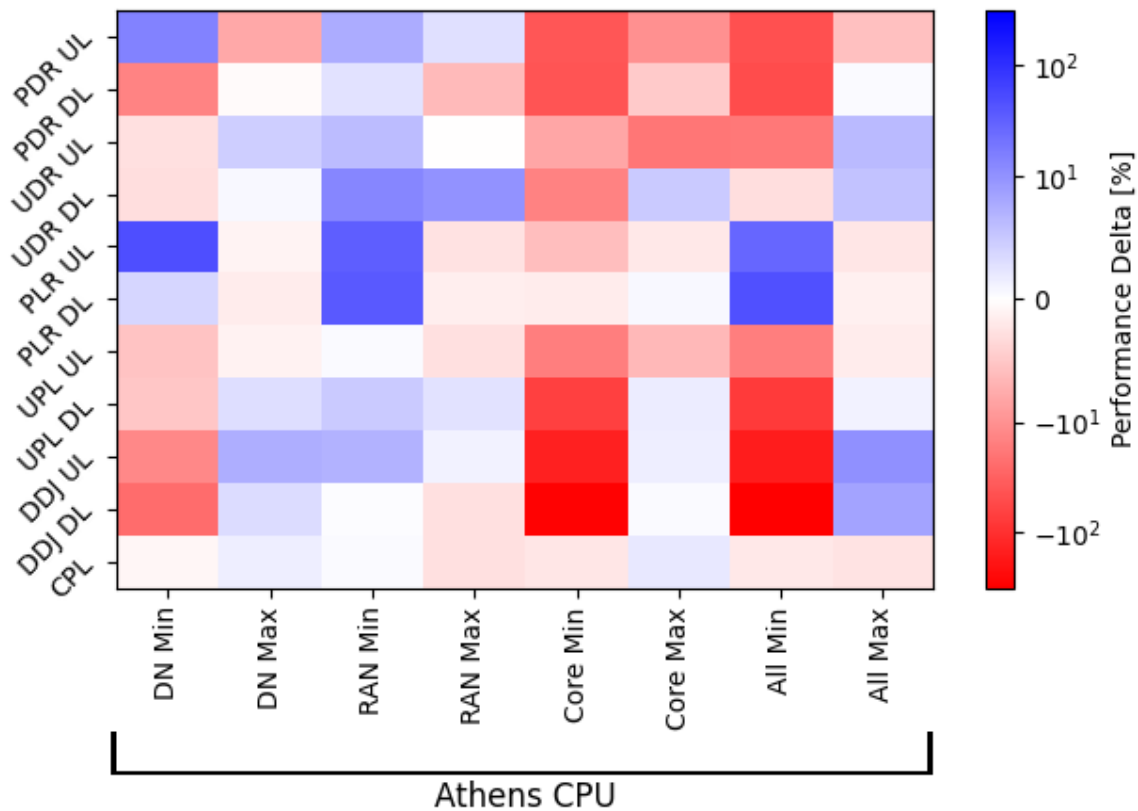


Figure 61 - Heatmap of Phase 4 CPU Allocation Athens Deltas

Once again, the higher COV values for Athens make it less useful for analysis than Berlin, but some patterns are present across both platforms. The Open5GS Core VM appears to be very sensitive to low CPU core allocations, as when the allocation is dropped to 1 all KPIs see significant degradation in performance across both platforms. Further testing showed however that this effect seems to dissipate quickly, as deltas were already approaching normal values once the allocation was changed to 2. Another pattern of note is the significant improvement in **PLR** values when the RAN and to some extent the DN are set to their minimum allocation values. The current theory explaining this counterintuitive effect is that, when multiple cores are assigned to these VMs, the processing of packets is shared between said cores, and the packets are recombined after processing. This introduces more opportunity for error and packet loss. When only one core is assigned to this task, no splitting and recombination occurs, theoretically leading to the improved performance observed. More research would be required to fully confirm this hypothesis, however.

The reduction of the RAN CPU allocation to its minimum value also appears to improve the performance of most of the KPIs across the board in both Athens and Berlin. This might have the same explanation as the aforementioned pattern in the **PLR** results, but once again more research is required to confirm this.

3.3.4 MEMORY ALLOCATION

The same type of tests described in the previous Section would also be run to determine the effects of differing memory (RAM) allocations for the 3 VMs. The minimum values were once again determined experimentally.

Table 36 - RAM Allocation Legend for Memory Allocation Testing

Setup ID	RAN VM Allocation	DN VM Allocation	Network Core VM Allocation
DN Min	8	4	4
DN Max	8	16	4
RAN Min	5	8	4
RAN Max	16	8	4
Network Core Min	8	8	2
Network Core Max	8	8	16
All Min	5	4	2
All Max	16	16	16

In the following Table 37 and Table 38, the results for the RTN are shown in the ACM Delta format: as percentage deltas from the respective platforms average 'Baseline' result.

Table 37 - Phase 4 Memory Allocation Results for the Berlin Platform

KPI/Direction	COV % Value	DN Min % Delta	DN Max % Delta	RAN Min % Delta	RAN Max % Delta	Core Min % Delta	Core Max % Delta	All Min % Delta	All Max % Delta
PDR UL	2,05	7,67	4,99	17,74	16,79	16,03	0,11	16,64	17,94
PDR DL	0,87	11,49	5,16	17,04	13,32	15,62	1,01	11,97	12,47
UDR UL	0,90	3,82	5,63	15,14	16,37	19,49	6,04	16,32	19,43
UDR DL	0,71	0,24	9,88	4,26	1,68	6,39	0,13	1,21	2,67
PLR UL	0,61	1,73	2,16	11,89	15,55	19,72	3,70	10,94	15,89
PLR DL	0,56	2,42	1,82	14,02	12,55	14,91	2,38	12,36	12,59
UPL UL	2,60	0,89	2,66	9,50	8,38	4,35	2,16	6,38	10,10
UPL DL	1,11	4,26	0,92	8,22	8,69	11,20	1,14	6,29	0,75
DDJ UL	7,05	7,11	7,92	5,75	9,01	28,53	3,04	18,50	3,86
DDJ DL	2,14	4,71	9,35	17,08	15,92	10,59	10,97	17,77	18,47
CPL	1,93	4,09	0,82	4,10	1,87	0,26	1,18	1,50	0,37

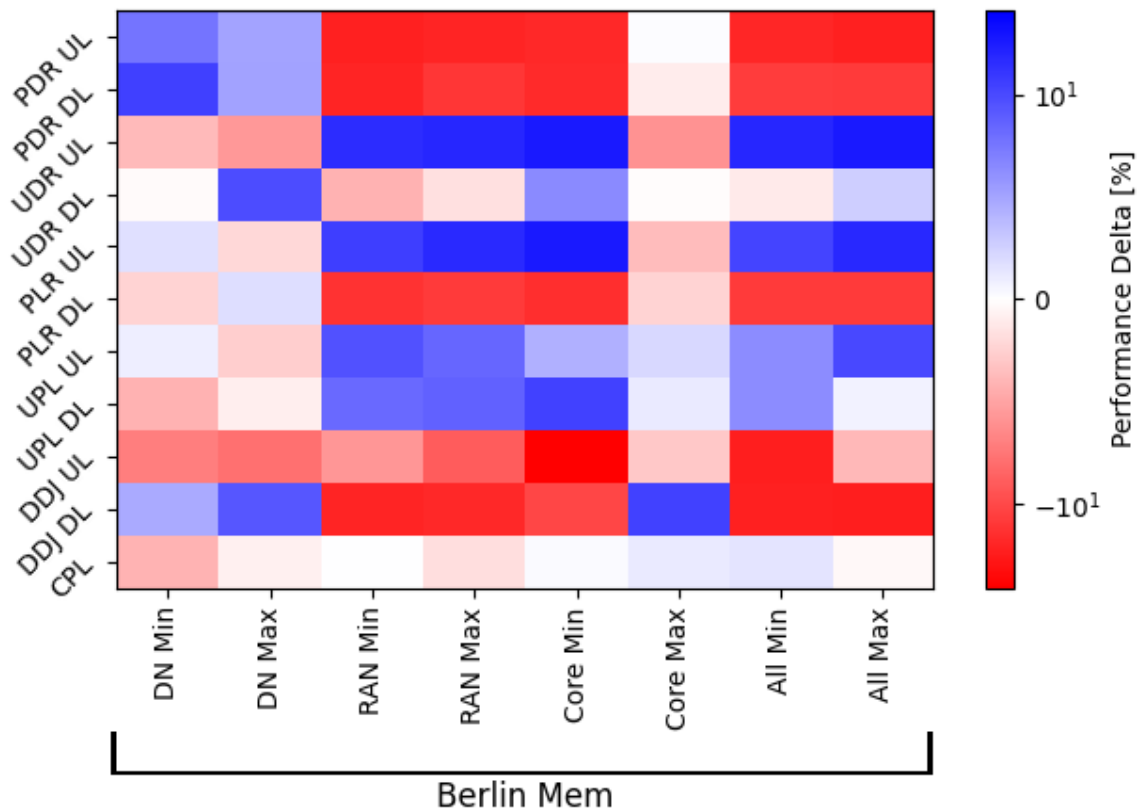


Figure 62 - Heatmap of Phase 4 Memory Allocation Berlin Deltas

Table 38 - Phase 4 Memory Allocation Results for the Athens Platform

KPI/Direction	COV % Value	DN Min % Delta	DN Max % Delta	RAN Min % Delta	RAN Max % Delta	Core Min % Delta	Core Max % Delta	All Min % Delta	All Max % Delta
PDR UL	10,47	1,58	4,60	9,20	10,68	8,66	10,22	6,63	0,00
PDR DL	5,02	4,47	3,23	2,54	22,10	0,10	2,08	5,04	5,75
UDR UL	16,74	7,40	5,31	8,99	2,43	0,55	16,79	1,85	6,61
UDR DL	10,92	17,93	1,66	46,06	4,23	0,39	1,34	6,15	0,18
PLR UL	3,65	0,05	1,47	0,27	1,51	0,01	0,59	0,78	0,72
PLR DL	3,69	1,16	0,16	2,19	1,94	0,95	1,27	1,30	2,44
UPL UL	9,49	1,66	4,95	2,97	3,65	5,50	3,07	4,12	4,31
UPL DL	4,94	3,27	5,41	6,13	1,54	3,70	0,18	1,29	2,22
DDJ UL	7,86	6,70	10,52	7,91	1,07	4,09	4,76	3,08	3,69
DDJ DL	6,64	3,00	9,17	6,70	1,52	3,41	5,01	1,15	1,56
CPL	1,76	1,79	0,25	1,60	0,60	1,73	1,83	1,54	4,06

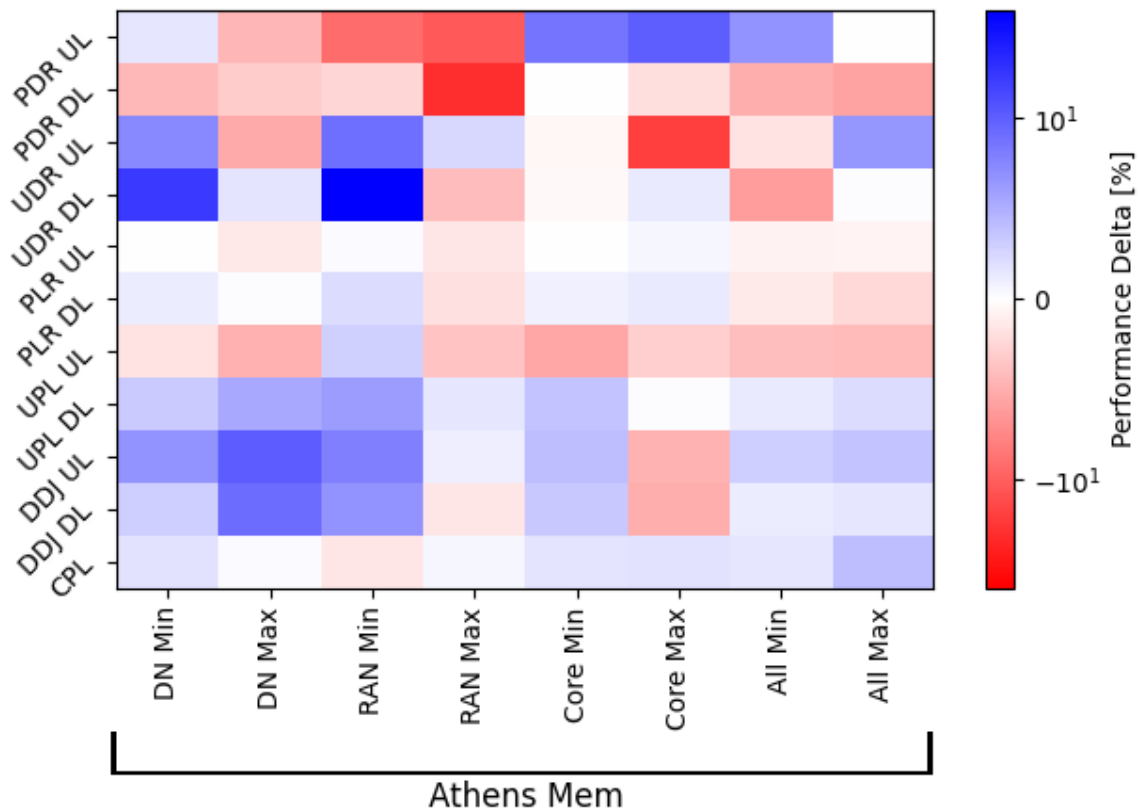


Figure 63 - Heatmap of Phase 4 Memory Allocation Athens Deltas

The results for the memory allocation testing offer perhaps the murkiest picture thus far. There are no clearly discernable general patterns either on a per KPI basis or on a per allocation basis. While there are some significant deltas in certain results (Core minimum/**PLR** in Berlin for example) these do not show consistency in either direction or across both platforms. Once again, similar to what was seen with the core allocation testing, there are significant amounts of counter-intuitive results – reductions in memory allocation leading to increased performance.

3.4 BENCHMARKING: FINAL ANALYSIS AND CONCLUSIONS

The initial questions that motivated this work as discussed in Deliverable D5.3 eventually germinated into two main lines of inquiry throughout the different phases of the investigation:

- Determine how close to parity the 4 6G-SANDBOX platforms were when all were running 4 “identical” virtualized deployments
- Determine what, if any, methods can be used to modify the performance of virtualized deployments beyond just the parameters in the virtualization layer

In both of these, significant surprises emerge when considering the entire body of work. Despite the fact that all four platforms have inherent differences, the virtualization structure of 6G-SANDBOX was expected to alleviate these differences to a significant degree, allowing for tests to be repeated across multiple platforms and for performance of a platform to be verified. Looking at the results shown in Deliverable D5.3 and Section 3.1 it is clear that this is not the case. The work in D5.3 shows the results of deploying four virtualized deployments in a “layman” sense – deployment was by hand, as was verification of all deployments being identical in the virtualized layer. The deltas seen are significant, to the point of the results being incomparable. When these

deployments are changed to be done in a more “professional”, systematic manner, the deltas are normalized to some extent, but their magnitudes are still such that they cannot be ignored, as seen in Section 3.1. This normalization went in both ways however, as some of the better results in Phase 1 such as the **UDR** were normalized up to the 20-40% range that most of the Phase 2 Deltas fall within. When we compare these results with those generated from Phase 3, only a couple of the deltas fall within the range of the measurement uncertainty and therefore cannot be used to describe performance disparity between the platforms. For most of the Phase 2 results, the MU can be used to explain only a small part of the observed delta, leaving underlying issues and differences between the platforms as the main culprit.

The issues encountered with the recreation of virtualized performance naturally led to a desire to go deeper, to attempt find more levers to pull and knobs to turn. The consortium’s expectation was that some of the more “obvious” parameters of interest such as hardware assignments and background noise would have clear and predictable impacts on the performance of virtualized deployments. These impacts could then be extrapolated into relationships and potential modifications that could then be used to understand how the platforms could be brought closer together.

Looking at Figure 64 and Figure 65, which collate all of the Phase 4 results into two heatmaps for the purpose of pattern identification, it is clear that direct relationships between factors and performance in the virtual layer are not consistently identifiable, much less proportional or expressible in terms of an equation or axiom. The most consistent overall pattern, visible in Figure 65, is the CPU core allocation to the network core, specifically when set to a minimum, which shows performance degradation across the board. No other individual factor or situation has such a consistent effect across all KPIs. The interference scenarios SC1 and SC2 appear to adversely impact the latency KPIs while improving the **UDR** and **PLR** across both platforms. As mentioned in Section 3.3.4 the memory allocation factors are the hardest to draw any conclusions from. Memory allocation to the DN VM appears to have a minimal effect on performance of any kind. The **PDR** performance appears to be adversely affected by both the maximum and minimum allocation of memory to the RAN VM, which in itself is an extremely surprising result. Looking at Figure 65 from the y-axis, the only clear pattern to be discerned is that the **CPL** KPI appears to be relatively resistant to all identified factors, showing no consistent performance delta.

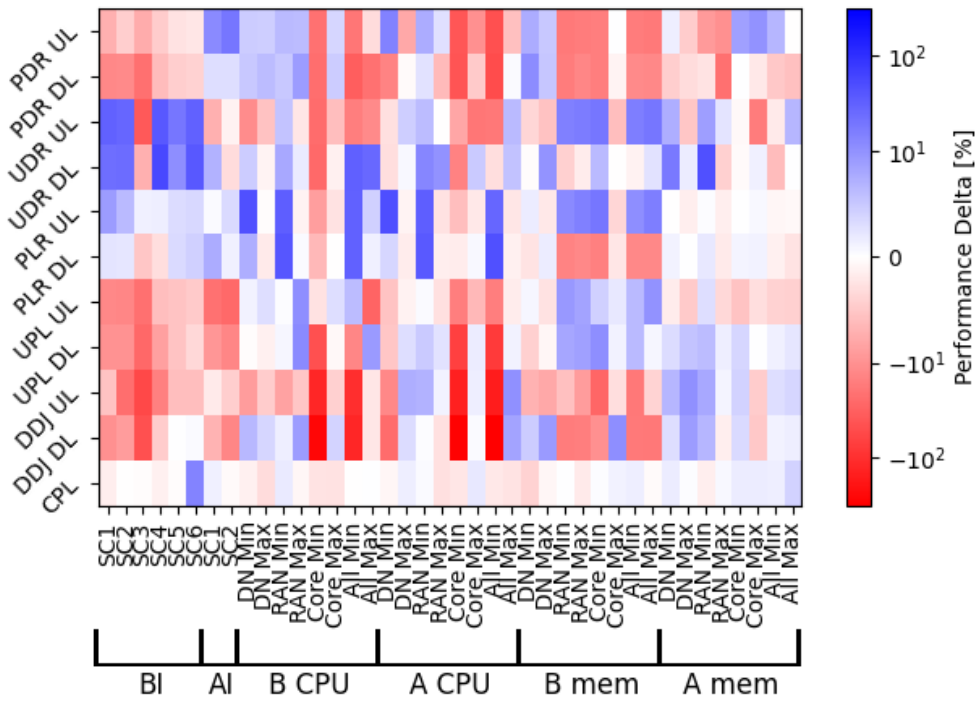


Figure 64 - Heatmap of all Phase 4 Deltas for both the Athens and Berlin Platforms, where Berlin is represented as “BI” and Athens as “A”. The interference test cases are referred to as “I”, the CPU allocation as “CPU” and the Memory allocation as “mem”

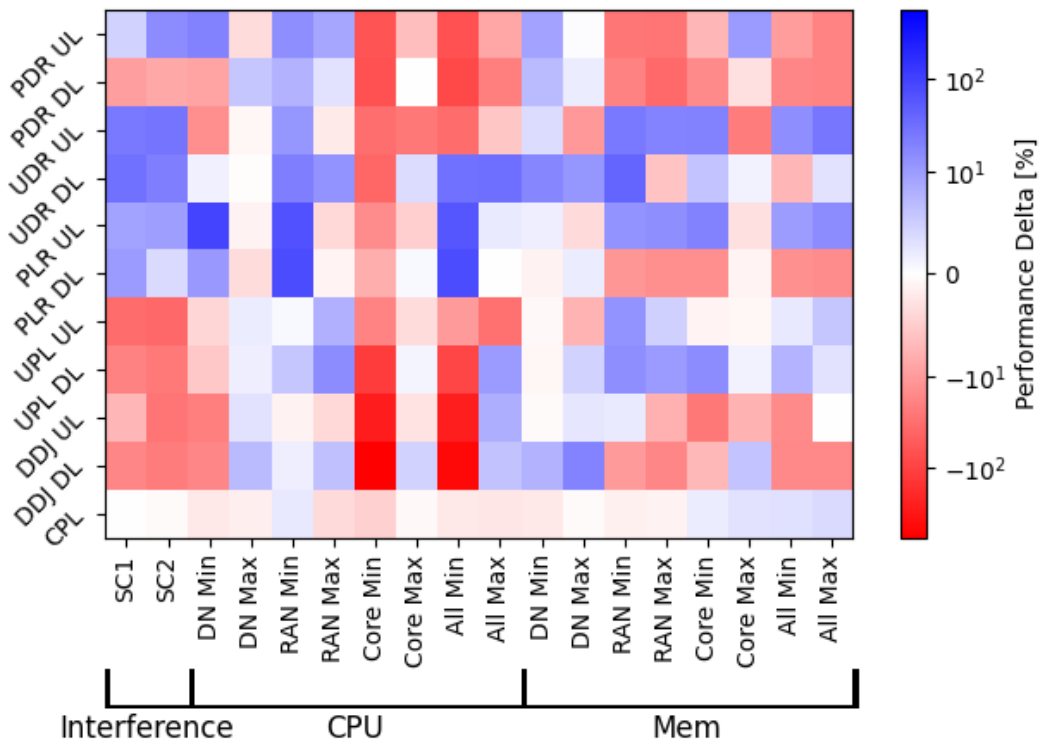


Figure 65 - Heatmap of Phase 4 Deltas for both Athens and Berlin summed together. Dark blue indicates an overall positive effect on both platforms, dark red indicates an overall negative, and light red/blue/white indicate either no or conflicting changes

While the lack of consistency here means that the desired results of Phase 4 were not achieved, the sparsity of patterns and consistent factors is in itself an extremely interesting observation. The sometimes seemingly random, sometimes counterintuitive effects of things such as the reduction of assigned hardware resources and the presence of other interfering deployments show us that the underlying connections between these different factors are far more complicated than first assumed. One of the sources of this complexity is certainly what these virtual deployments are trying to put together – networks that are pushing beyond the edge of 5G capabilities and into the expanding 6G, truly SBA space. Add onto this complexity the various systems implemented by 6G-SANDBOX such as a standardized deployment and networking methodology and the key question provided by the work in this paper begins to arise – How are virtualized network deployments expected to be modifiable by operators in an intuitive way if these relationships are abstracted to the point where changes to underlying factors are completely unpredictable?

From Phase 1 until Phase 4, the work described herein evolved from a relatively simple investigation into levels of disparity between multiple virtualized deployments, into an exploration and discovery of just how abstracted factors below the virtualization layer are in terms of having an impact on performance in the virtual layer. The two goals are intrinsically connected – achieving parity once the virtualization layer is exhausted of parameters is only possible by going deeper below, and if the factors deeper below are unwieldy and unpredictable it leads to deployments that have to be observed in vacuums with their tests being essentially unrepeatable. Considering the importance of measurement for the validation of performance, uncertainty in the results from deployment to deployment would have a distinct effect in terms of proving the capabilities of 6G enabled systems. These conclusions obviously have to come with the caveat and framing of being conducted within the auspices of 6G-SANDBOX. It is entirely possible that some of these abstractions between the hardware and virtual layers come as a result of the various different infrastructure wrappers present within 6G-SANDBOX deployments, and that without them the hardware-virtual connection would be simpler, and perhaps easier to observe relationships between. However, given the complexity of SBA network deployments, it is safe to assume that any future consumer grade deployment would have to have some measure of surrounding infrastructure to support its virtualized deployments and continued operation. While the effects of this perspective infrastructure would almost certainly not be the same as the results seen in this paper, the results here do indicate that a level of abstraction should be expected.

4 REPORT ON DIGITAL TWIN EFFORTS

The methodology and design of the scenario for the EXata Digital Twin (DT) of the Athens platform were initially presented in previous deliverables, specifically D4.2 [50]. The complementary proposal for formal digital twins by University of Malaga was also partially developed in WP4. However, in order to report the experimentation results on WP5, the most recent contributions in these two domains are now reported in D5.5. This section presents the latest updates to the methodologies and scenarios, together with the corresponding results and findings.

4.1 EXATA DIGITAL TWIN IN ATHENS

The **first step** focused on the deployment and configuration of the antenna infrastructure at the NCSR campus. As part of the Athens platform DT deployment in EXata, a new outdoor Radio Access Network (RAN) component at NCSR campus has been introduced: an octo-port multiband antenna, shown in Figure 66 and Figure 67 [51]. The antenna is 1.4 m long, provides a 65° azimuth beamwidth, and supports the 1427–2690 MHz and 3100–4200 MHz frequency ranges through four wideband ports. The corresponding radio and deployment parameters were documented to enable accurate replication within the Digital Twin environment.

The antenna and radio configuration used in the experimental setup is based on a 5G NR deployment operating in Time Division Duplexing (TDD) mode with a carrier frequency of 3489.42 MHz and a channel bandwidth of 50 MHz. The system employs a 2x2 MIMO configuration to support spatial diversity and improve radio

performance. Furthermore, the deployment is configured with a subcarrier spacing (SCS) of 30 kHz and a transmission power of 0 dBm, corresponding to a transmit power level of 1 mW. Regarding the TDD frame structure, the configuration follows a periodic pattern with a duration of 5 ms. Within each period, the radio resources are allocated to both downlink and uplink transmissions through a combination of slots and symbols. More specifically, the configuration includes 7 downlink slots, 2 uplink slots, 6 downlink symbols, and 4 uplink symbols. This allocation assigns a larger portion of the available radio resources to downlink traffic while still maintaining dedicated uplink transmission opportunities.



Figure 66 - The new antenna shown in situ on the NCSR Campus

```
* NR SA FDD or TDD cell */

#define NR_TDD          1 // Values: 0 (NR FDD), 1(NR TDD)
#define FR2              0 // Values: 0 (FR1), 1 (FR2)
#define FR2
#define UDC_TYPE         1 // Values: 0 (NO UDC), 1 (B2), 2 (A2), 3 (B4) (refer to application note to identify your UDC type)
#define NR_TDD_CONFIG    10 // Values: FR1: 1, 2, 3, 4 (compatible with LTE TDD config 2) FR2: 10
#define NR_BANDWIDTH     100 // NR cell bandwidth
#define IF_ATTENUATION   -9 // Value: 10*log10(N_PORTS_COMBINER) + external eventual IF attenuators
#define IF_ATTENUATION
#else
#define NR_TDD_CONFIG    2 // Values: FR1: 1, 2, 3, 4 (compatible with LTE TDD config 2) FR2: 10
#define NR_BANDWIDTH     50 // NR cell bandwidth
#define NR_BANDWIDTH
#endif
#define N_ANTENNA_DL      2 // Values: 1 (SISO), 2 (MIMO 2x2), 4 (MIMO 4x4)
#define N_ANTENNA_UL      1 // Values: 1, 2, 4
#define NR_LONG_PUCCH_FORMAT 0 // Values: 0 (automatic), 2, 3 or 4

/* define to 1 to enable periodic SRS with N_ANTENNA_UL ports. Uplink
SU-MIMO is also enabled if N_ANTENNA_UL >= 2. Not all UEs support
uplink SU-MIMO. */
#define USE_SRS          0
```

[illegible]

Figure 67 - Configuration parameters for the new antenna on the NCSRD Campus

As a **second step**, a dedicated measurement script was created within the Nemo application to define the measurement procedures, logging parameters, and network performance indicators to be collected during the campaign [52]. Using the KEYS measurement UE (Figure 68) running the configured Nemo script, Signal to Noise Ratio (SINR) measurements were then collected along the selected route within the NCSR campus.



Figure 68 - UE Provided by KEYS for use with NEMO application testing

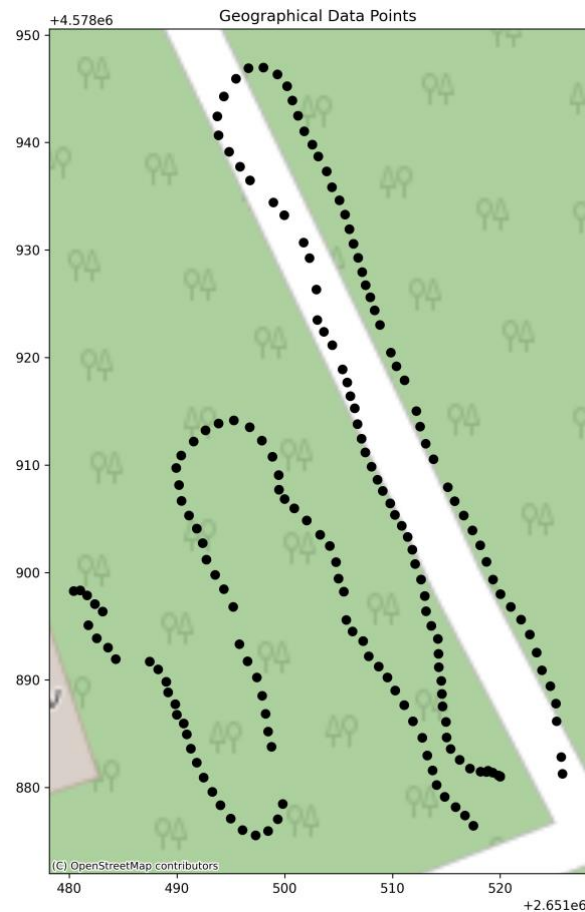


Figure 69 - Nemo Application export of the generated route based on the GPX data

Following the completion of the measurement campaign, the collected data was exported from the Nemo application. This process included both the signal measurements and the geographical coordinates corresponding to the traversed route (Figure 69). The exported datasets were subsequently processed and prepared for integration into the EXata simulation environment. This process included the incorporation of the antenna coordinates as well as the GPX route coordinates, enabling the accurate replication of the measurement path within the EXata scenario (Figure 70).

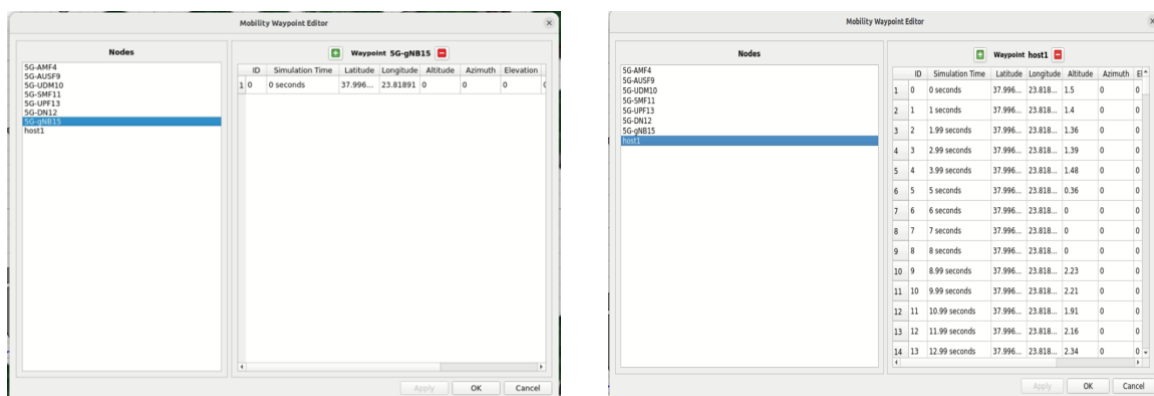


Figure 70 - Location of the new antenna and coordinates of the different waypoints imported into the EXata scenario

In addition, the antenna-related configurations were imported into EXata and applied to Node 15, which represents the antenna entity in the considered scenario. The final setup of the scenario, including the generated path, is illustrated in Figure 71.



Figure 71 - Final setup of the configuration of 5G SA deployment of the EXata Digital Twin and the path taken on the NCSR Campus

The results presented in Figure 72 compare the real measurements collected using the Nemo application with the corresponding measurements generated by the EXata Digital Twin scenario. In both cases, the measurements are mapped along the same trajectory within the NCSR campus, enabling a direct comparison between the real-world and simulated environments. The comparison indicates that the EXata scenario is capable of reproducing the general behavior and spatial trends observed in real measurements. In particular, similar variations in the SINR levels can be observed along different sections of the route, demonstrating that the DT captures the overall propagation characteristics of the deployment scenario. However, differences between real and simulated measurements are still present. This is expected, as the current iteration of the DT only incorporates the antenna-related configurations imported from the real deployment. More specifically, the simulation currently reproduces the antenna setup and propagation configuration, without yet including additional environmental and network-related factors such as detailed obstacle modelling, dynamic interference conditions, traffic load variations, or advanced radio optimization parameters. Therefore, the presented results should be considered as an initial validation step of the DT framework, demonstrating that the imported antenna configuration alone is sufficient to reproduce the general signal performance tendencies observed during the real measurement campaign.

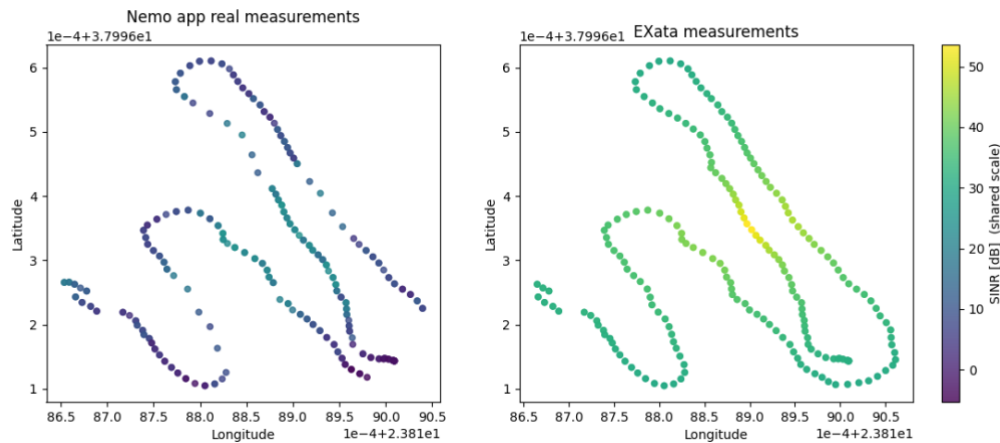


Figure 72 - Real-world measurements collected through the use of Nemo compared to the corresponding measurements generated by EXata

The next step is to import the antenna radiation patterns into EXata and perform a new set of experiments. This addition increases the realism of the simulations by accounting for the directional variation of antenna gain. It replaces the earlier simplifying assumption of isotropic antennas, which unrealistically assume uniform signal strength in all directions and may therefore overestimate link quality. As a result, variations in SINR are expected and should be interpreted as a consequence of more accurate antenna modeling rather than degradation in system performance. In the updated simulation results seen in Figure 73, the observed SINR differences arise from the use of direction-dependent antenna gains, which affect both the received signal strength and the interference levels. Since isotropic antenna models do not capture these directional effects, systematic differences between the two configurations on the order of 8 dB are expected.

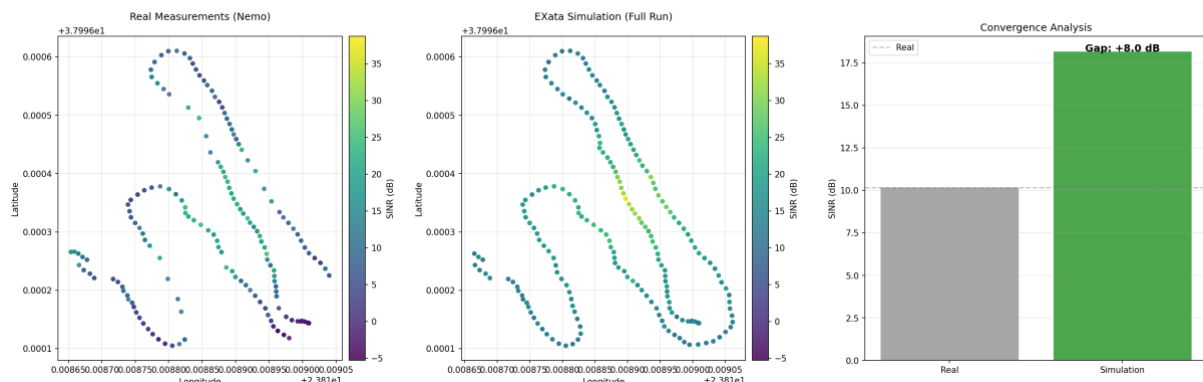


Figure 73 - Real-world measurements collected through the use of Nemo compared to the corresponding measurements generated by EXata. In this case, the antenna parameters are included in the simulation, as well as the radiation pattern files. The final graph shows the gap between the measured SINR values and the simulated ones

This experiment provides an initial understanding of the RAN Digital Twin and establishes a foundation for future studies aimed at increasing the realism and accuracy of the model. Additional parameters and environmental characteristics can be incorporated in subsequent experiments to further refine the simulation framework and improve its capability to reproduce the behavior observed in the physical network. By progressively integrating more realistic configurations and measurements, the RAN DT is expected to achieve closer alignment with real-world performance metrics and provide a more reliable representation of the operational environment.

4.2 FORMAL DIGITAL TWINS IN UMA

The evolution of communication networks toward 6G will significantly elevate the importance of protocols, shifting the focus toward the analysis of protocol implementations beyond traditional end-to-end Key Performance Indicators (KPIs). Traditionally, protocol engineering techniques have relied on theoretical models

of protocols, which are manually constructed by engineers. These models are subsequently processed using automated tools to perform correctness analysis, automated test case generation and execution, automated implementation, or performance evaluation.

Automata Learning has emerged as an efficient paradigm to automatically construct protocol models directly from the observation of their implementations by monitoring network traces. Packets extracted from these traces provide the necessary data to build ad-hoc state machines that represent selected aspects of the protocol's behavior. If these state machines or automata are constructed with sufficient information from the traces—incorporating, for instance, probabilistic aspects—the resulting automaton can be considered a Digital Twin of the protocol or network, which can then be utilized for runtime decision-making. For example, using a pre-built automaton to track protocol progression within a network can facilitate the detection of protocol anomalies or even automate network reconfiguration maneuvers.

The work carried out by the University of Málaga (UMA) within WP4 and WP5 of the 6G-SANDBOX project has resulted in a software framework designed to construct and analyze probabilistic automata from real-world networks. This framework applies Statistical Model Checking (SMC) to detect operational anomalies. The specific variant of automata proposed by UMA is termed the Formal Digital Twin (FDT).

This section summarizes the application of the FDT approach to analyze the Precision Time Protocol (PTP) via passive automata learning techniques. Specifically, the evaluation focuses on ptp4l, a widely adopted Linux implementation of PTP integrated into numerous real-time systems. The proposed methodology constructs an FDT of ptp4l across various network scenarios. To achieve this, UMA has implemented Passta, a specialized toolchain that automatically generates formal models of real systems using a passive automata learning algorithm. The inferred ptp4l FDT is evaluated to verify the protocol's compliance with the official standard, confirming its correct behavior in essential applications and thereby contributing to safe, dependable implementations within 6G networks.

All technical details of this work, as well as references to related literature are available in the relevant journal articles [54].

4.2.1 THE CASE STUDY: PTP PROTOCOL

With the deployment of 6G networks, the role of real-time communications is expanding significantly. A key pillar of 6G is Hyper-Reliable and Low-Latency Communications (HRLLC), a critical enabler for applications demanding instantaneous data exchange. HRLLC supports a wide array of time-sensitive systems, including autonomous vehicles, industrial automation, and remote surgery, all of which require precise synchronization to function safely. Even minor timing discrepancies in these systems can lead to catastrophic outcomes, such as systemic failures or severe safety hazards.

In this context, the Precision Time Protocol (PTP) stands out as one of the most critical protocols for clock synchronization. PTP operates on a leader-follower (master-slave) architecture, where a leader clock synchronizes the clocks of multiple follower devices across a network with nanosecond-level precision. To guarantee the reliability of these synchronized systems, verifying the correct operation of PTP is paramount. Given its critical role in emerging 6G applications, a thorough and systematic operational analysis under real-world conditions is required. A promising avenue to achieve this is the use of FDTs, which provide high-fidelity behavioral models of PTP suitable for formal verification.

As specified in the IEEE 1588 standard, PTP synchronizes devices by exchanging time-stamped messages that allow follower devices to adjust their clock frequencies and offsets. The synchronization procedure involves several sequential steps:

- The leader periodically broadcasts an Announce message to inform the network of its identity, quality, and clock status.
- The leader sends a Sync message, recording its exact transmission time (t_1).

- Upon receiving this message, the follower clock records its local reception time (t_2).
- The follower then transmits a Delay Request message back to the leader, logging its emission time (t_3).
- The leader receives the request, captures the arrival time (t_4), and returns this value inside a Delay Response message.

With these four timestamps (t_1 , t_2 , t_3 , t_4), the follower calculates the network propagation delay and its clock offset relative to the leader using the following formulas:

$$\text{Offset} = ((t_2 - t_1) - (t_4 - t_3)) / 2$$

$$\text{Delay} = ((t_2 - t_1) + (t_4 - t_3)) / 2$$

Following these calculations, the follower clock is adjusted accordingly. This process loops continuously to maintain tight synchronization. Since the analysis conducted in this work focuses on the follower side, a detailed overview of follower state behaviors is presented below according to the state machine semantics defined in the IEEE 1588–2019 standard (a diagram of the state machine is shown in Figure 74):

- **Initializing:** The initial state representing internal device configuration. Upon completion (Initialization_complete), the system transitions to Listening.
- **Listening:** The device attempts to join a PTP domain. It can loop back to itself if an Announce_receipt_timeout_expires event occurs, transition to Faulty if a Fault_detected event is triggered, or move to Uncalibrated via the bmc_slave event when the Best Master Clock (BMC) algorithm designates the device as a follower.
- **Uncalibrated:** A transient state where the follower prepares for synchronization. Once readiness requirements are met, the Master_clock_selected event triggers a transition to the Slave state.
- **Slave:** Represents the active, continuous synchronization process with the leader clock. If synchronization is lost, a Synchronization_fault event moves the system back to Uncalibrated.
- **Faulty:** A state entered upon any critical internal error. The system restarts and returns to Initializing once the fault condition is resolved (Fault_cleared).

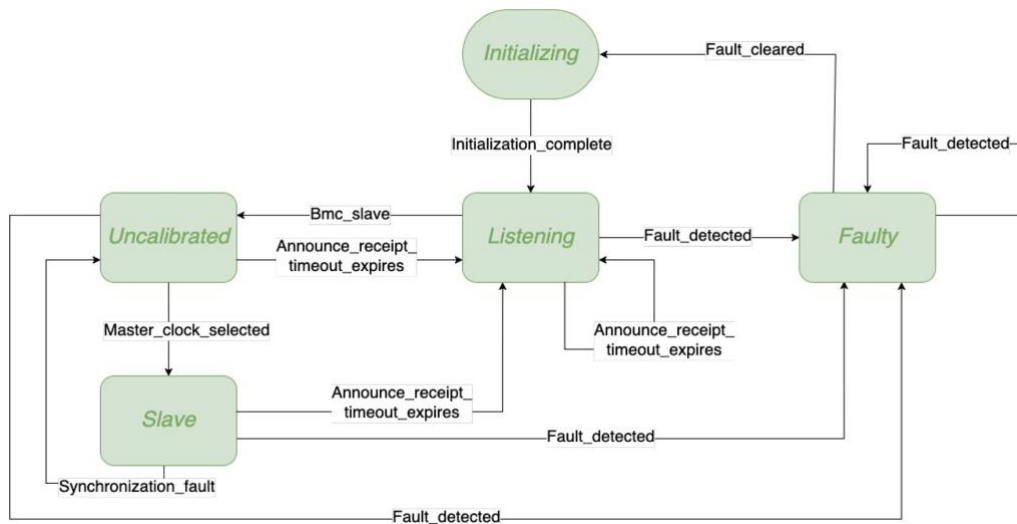


Figure 74 - PTP state machine of a slave device connected to one master.

4.2.1.1 KEY PROPERTIES FOR PTP

Based on the IEEE 1588 standard and the configuration parameters of the ptp4l implementation, a set of eleven core properties (P1–P11) was defined to capture the expected behavior of the protocol:

- **P1:** PTP transitions from Initializing to Listening upon the Initialization_complete event.

- **P2:** PTP transitions from Listening to Uncalibrated upon the `bmc_slave` event.
- **P3:** PTP transitions from Uncalibrated to Slave upon the `Master_clock_selected` event.
- **P4:** When a timeout expires, PTP transitions from Listening, Uncalibrated, or Slave back to Listening.
- **P5:** PTP transitions to Faulty upon a `Fault_detected` event from any state except Initializing.
- **P6:** PTP transitions from Faulty to Initializing upon the `Fault_cleared` event.
- **P7:** The Listening state has exactly three outgoing transitions: to Uncalibrated (on `bmc_slave`), to Listening (on `Announce_receipt_timeout_expires`), and to Faulty (on `Fault_detected`).
- **P8:** The Uncalibrated state has exactly three outgoing transitions: to Slave (on `Master_clock_selected`), to Listening (on `Announce_receipt_timeout_expires`), and to Faulty (on `Fault_detected`).
- **P9:** The Slave state has exactly three outgoing transitions: to Uncalibrated (on `Synchronization_fault`), to Listening (on `Announce_receipt_timeout_expires`), and to Faulty (on `Fault_detected`).
- **P10:** The Faulty state has exactly two outgoing transitions: to Initializing (on `Fault_cleared`) and to Faulty (on `Fault_detected`).
- **P11:** `ptp4l` bounds the residence time in the Faulty state using the `fault_reset_interval` parameter (16 seconds by default); thus, the implementation must remain in Faulty for at most 16 seconds.

4.2.2 AUTOMATED MODELING OF PTP AS A FORMAL DIGITAL TWIN

In order to generate the Formal Digital Twin of PTP, passive automata learning algorithms are applied to infer formal stochastic models from a System Under Learning (SUL) without actively perturbing its operation. Formal verification techniques, such as model checking, are then applied to these FDT to mathematically guarantee correctness.

The contributions made by UMA follow two distinct directions:

- **Methodological & Tool Framework:** UMA has developed a methodology supported by a robust toolchain (Passta) that extracts execution logs to infer Stochastic Real-Timed Automata (SRTA). These inferred models are automatically translated into UPPAAL timed automata format, enabling statistical model checking via UPPAAL SMC. The soundness of this transformation has been mathematically proven, ensuring that properties verified within the UPPAAL model are preserved by the automata learned by Passta.
- **Empirical Evaluation:** The framework was rigorously tested against the two latest major releases of the Linux PTP implementation (`ptp4l` v3 and `ptp4l` v4) under five distinct network scenarios. The results exposed several non-compliant behaviors relative to the IEEE 1588 standard, demonstrating the framework's effectiveness in uncovering hidden operational anomalies.

4.2.2.1 TOOLCHAIN OVERVIEW

The methodology is structured into four distinct, sequential stages (shown in Figure 75)

- **Stage 1 - Trace Extraction:** Raw execution logs collected from the `ptp4l` clients (SUL) are processed by an ad-hoc program called the Traces Extractor, converting raw strings into structured learning traces.
- **Stage 2 - Automata Learning:** Passta's trace processing module removes redundancies, and its learning module applies a passive learning algorithm to infer an SRTA representing the system's empirical behavior.
- **Stage 3 - Model Validation:** Because passive learning relies entirely on observation, an iterative learning-validation loop is established. A separate dataset (validation traces) is evaluated against the learned FDT. If a trace is rejected, it is reintroduced into the learning module to refine the model's structure and timing bounds until a predefined convergence threshold is met.
- **Stage 4 - Formal Verification:** The validated SRTA is parsed into a stochastic timed automaton format readable by UPPAAL SMC. Concurrently, the natural-language protocol properties (P1–P11) are formalized as Metric Interval Temporal Logic (MITL) formulas and analyzed via statistical model checking.

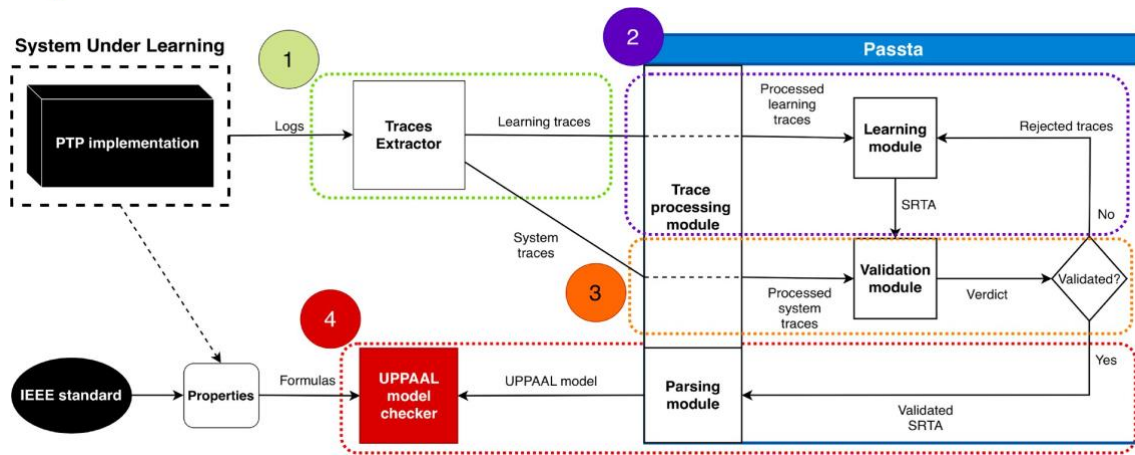


Figure 75 - Overview of the PTP Digital Twin Toolchain

Figure 76 and Figure 77 provide a graphical representation of a portion of the SRTA and the corresponding UPPAAL model that will support statistical model check-in.

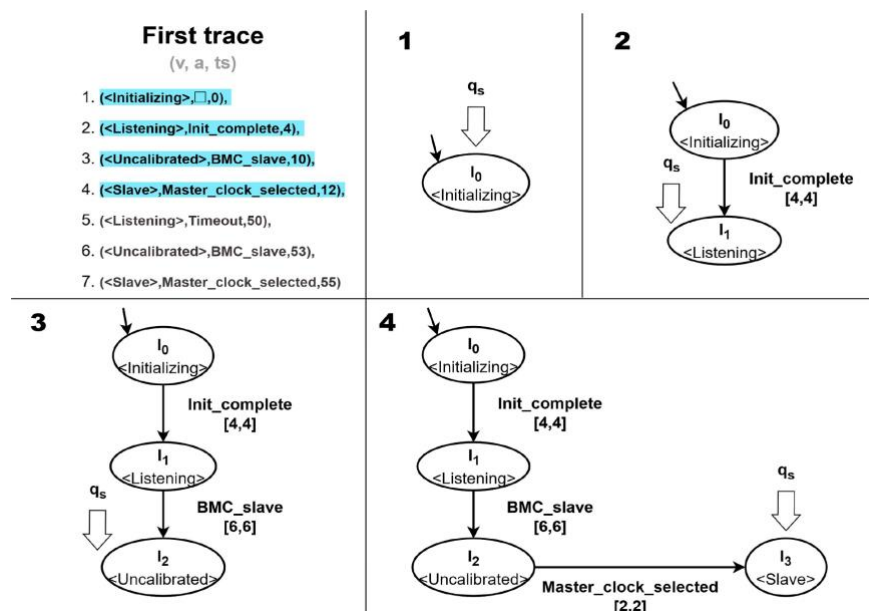


Figure 76 - Visual representation of Passta learning the first four observations of the first trace

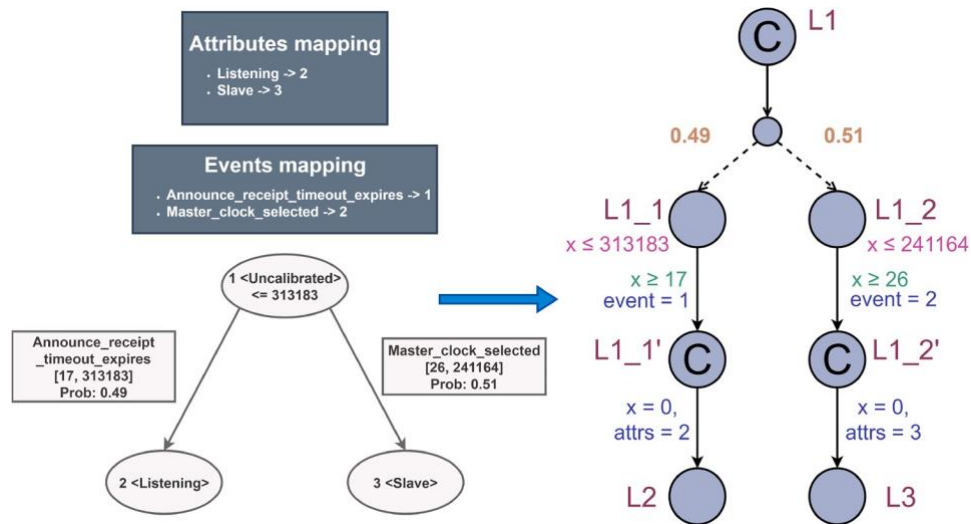


Figure 77 - Visual representation of the transfer of SRTA to UPPAAL

4.2.3 EXPERIMENTAL RESULTS

The workflow was applied to compare the behavioral profiles of ptp4l v3 and ptp4l v4 under a single leader-follower architecture. On the leader side, an Edge Grandmaster Clock (μFalcon-MX4) was utilized. On the follower side, two separate computers running Ubuntu 22.04.4 LTS executed their respective versions of ptp4l.

4.2.3.1 TRACE EXTRACTION AND MODEL VALIDATION

Five distinct network scenarios were engineered to trigger the behaviors specified in the PTP state machine:

- **Scenario 1 (Ideal Conditions):** Baseline execution under optimal network conditions with zero artificial delays or drops.
- **Scenario 2 (Egress Traffic Delay):** Random 1 ms delays are injected into the follower's egress traffic. Delay intervals last between 10 and 60 seconds, separated by random gaps of 10 to 60 seconds.
- **Scenario 3 (Network Disconnections):** Temporary disconnections are introduced by randomly shutting down the network interface following the same timing pattern as Scenario 2.
- **Scenario 4 (Announce Timeouts):** The follower parameter logAnnounceInterval is set to -5 (expecting a message every ~31.25 ms) while the leader continues broadcasting at the default rate, rapidly triggering announce timeouts.
- **Scenario 5 (Sync Timeouts):** Analogous to Scenario 4, the parameter logSyncInterval is forced to -5 to rapidly trigger synchronization timeout conditions.

Each scenario was executed at least 500 times, generating 300 traces for learning and 200 traces for validation (each trace spanning a 5-minute window). The stopping criterion required the FDT to recognize 100% of the learning traces and at least 90% of the validation traces. As shown in Table 39, stable convergence was achieved across all scenarios on the very first iteration.

Table 39 - Validation trace acceptance rates across different network scenarios

Version	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5
ptp4l v3	98.5%	94.0%	94.0%	96.0%	94.0%
ptp4l v4	95.0%	96.0%	94.0%	98.0%	94.0%

4.2.3.2 PROPERTIES VERIFICATION AND ANALYSIS

The properties were formalized into formulas specified with MITL logic [54]. The initial 11 properties are transformed into 11 MITL formulas (F1 to F11). Because UPPAAL SMC uses a simulation-based approach,

property verification outputs a probability value within the interval [0, 1] alongside a 99% confidence interval. Table 40 summarizes these verification results.

Table 40 - Satisfaction probabilities for MITL properties across ptp4l versions

Formula	Scenario 1 (v3 / v4)	Scenario 2 (v3 / v4)	Scenario 3 (v3)	Scenario 3 (v4)	Scenario 4 (v3 / v4)	Scenario 5 (v3 / v4)
F1	[0.99, 1.00] / [0.99, 1.00]	[0.99, 1.00] / [0.99, 1.00]	[0.99, 1.00]	[0.99, 1.00]	[0.99, 1.00] / [0.99, 1.00]	[0.99, 1.00] / [0.99, 1.00]
F2	[0.99, 1.00] / [0.99, 1.00]	[0.99, 1.00] / [0.99, 1.00]	[0.99, 1.00]	[0.99, 1.00]	[0.99, 1.00] / [0.99, 1.00]	[0.99, 1.00] / [0.99, 1.00]
F3	[0.99, 1.00] / [0.99, 1.00]	[0.99, 1.00] / [0.99, 1.00]	[0.99, 1.00]	[0.99, 1.00]	[0.99, 1.00] / [0.99, 1.00]	[0.99, 1.00] / [0.99, 1.00]
F4.1-3	[1.00, 1.00] / [1.00, 1.00]	[1.00, 1.00] / [1.00, 1.00]	[1.00, 1.00]	[1.00, 1.00]	[0.99, 1.00] / [0.99, 1.00]	[0.99, 1.00] / [0.99, 1.00]
F5	[1.00, 1.00] / [1.00, 1.00]	[0.99, 1.00] / [0.99, 1.00]	[0.99, 1.00]	[0.99, 1.00]	[1.00, 1.00] / [1.00, 1.00]	[1.00, 1.00] / [1.00, 1.00]
F6	[1.00, 1.00] / [1.00, 1.00]	[0.00, 0.01] / [0.00, 0.01]	[0.00, 0.01]	[0.00, 0.01]	[1.00, 1.00] / [1.00, 1.00]	[1.00, 1.00] / [1.00, 1.00]
F7	[0.99, 1.00] / [0.99, 1.00]	[0.99, 1.00] / [0.99, 1.00]	[0.99, 1.00]	[0.71, 0.73]	[0.99, 1.00] / [0.99, 1.00]	[0.99, 1.00] / [0.99, 1.00]
F8	[0.99, 1.00] / [0.99, 1.00]	[0.99, 1.00] / [0.99, 1.00]	[0.99, 1.00]	[0.96, 0.98]	[0.99, 1.00] / [0.99, 1.00]	[0.99, 1.00] / [0.99, 1.00]
F9	[1.00, 1.00] / [1.00, 1.00]	[0.99, 1.00] / [0.99, 1.00]	[0.00, 0.01]	[0.49, 0.51]	[0.99, 1.00] / [0.99, 1.00]	[0.99, 1.00] / [0.99, 1.00]
F10	[1.00, 1.00] / [1.00, 1.00]	[0.00, 0.01] / [0.00, 0.01]	[0.00, 0.01]	[0.00, 0.01]	[1.00, 1.00] / [1.00, 1.00]	[1.00, 1.00] / [1.00, 1.00]
F11 (16s)	[1.00, 1.00] / [1.00, 1.00]	[0.00, 0.01] / [0.00, 0.01]	[0.10, 0.11]	[0.00, 0.01]	[1.00, 1.00] / [1.00, 1.00]	[1.00, 1.00] / [1.00, 1.00]
F11 (17s)	[1.00, 1.00] / [1.00, 1.00]	[0.99, 1.00] / [0.99, 1.00]	[0.89, 0.91]	[0.99, 1.00]	[1.00, 1.00] / [1.00, 1.00]	[1.00, 1.00] / [1.00, 1.00]
F11 (18s)	[1.00, 1.00] / [1.00, 1.00]	[0.99, 1.00] / [0.99, 1.00]	[0.96, 0.98]	[0.99, 1.00]	[1.00, 1.00] / [1.00, 1.00]	[1.00, 1.00] / [1.00, 1.00]

The results reveal that while core transitions under nominal states behave correctly (F1–F5), significant deviations from the IEEE 1588 standard occur under adverse network conditions (Scenarios 2 and 3), tied directly to error-handling behaviors:

- **Violation of F6 & F10:** Neither version transitions from Faulty back to Initializing upon a Fault_cleared event. Instead, the learned models reveal that the implementations bypass this path entirely, resulting in low satisfaction probabilities.
- **Structural Divergences in ptp4l v4 (F7 & F8):** Under Scenario 3, ptp4l v4 displays extra, non-standard transitions out of the Listening and Uncalibrated states (e.g., direct transitions to unknown internal states '<>' on Fault_cleared), failing the strict constraints mandated by the standard. LinuxPTP changelogs confirm that these stem from an extra monitoring port introduced in v4.
- **Residence Time in Faulty State (F11):** Nominally configured to exit the Faulty state after 16 seconds, statistical evaluation shows that under traffic delays (Scenario 2), both implementations require slightly

longer (~17 s) to recover. Under network disconnections (Scenario 3), ptp4l v3 exhibits a delayed tail reaching up to 18.457 s. Conversely, ptp4l v4 displays a tighter distribution, strictly adhering to the threshold when relaxed to 17 seconds, which suggests that a legacy timing issue in v3 has been addressed in v4.

4.2.4 EVALUATION OF THE APPROACH

While the proposed passive FDT approach successfully exposes behavioral anomalies and version deviations, it is bounded by specific constraints:

- **Completeness:** Passive learning is inherently constrained by the variety of execution logs available. For instance, a Synchronization_fault event cannot be safely triggered without active system perturbation and is thus absent from the models.
- **Black-Box Limitations:** The framework identifies externally observable deviations but cannot pinpoint whether they stem from implementation bugs, deliberate design relaxations, or auxiliary logging mechanisms.
- **Generality:** Despite focusing on a single leader-follower topology, the Passta framework is system-agnostic. It can easily scale to complex topologies (multi-domain, multi-leader) or alternative time-sensitive protocols by simply adapting the trace extraction regex modules and defining the corresponding event abstractions.

5 FINAL KPI OVERVIEW

As part of the validation of success for the project, 6G-SANDBOX had a goal to measure at least one KPI for each of the KPI families defined in Deliverable 2.4. In order to not only achieve but surpass this expectation, an effort was made to measure every possible previously defined KPI. In the following Table 41, these KPIs are listed, with information relating to whether they were measured or not, the method by which they were measured, where evidence of said measurement can be found, and any notes regarding said KPI.

Table 41 - Final Overview of measured KPIs within 6G-SANDBOX

KPI Family	KPI	Measured?	Methodology	Evidence	Notes
Capacity (Network)	Peak Data Rate (PDR)	Yes	Wraparound Core testing using LoadCore Tool	Deliverable D5.3 Deliverable D5.5	Measured at all 4 platforms
	Area Traffic Capacity (ARR)	Yes	Network UE Simulation using UESIM	Deliverable D5.5	Measured in Malaga
	User Density/Connection Density (COD)	No			Lack of necessary infrastructure during measurement campaign
	RIS Capacity Enhancement (RCE)	Yes	Physical Testing in Malaga City Centre	Deliverable D5.5	
Capacity (User)	User Experienced Data Rate (UDR)	Yes	Wraparound Core testing using LoadCore Tool	Deliverable D5.3 Deliverable D5.5	Measured at all 4 platforms
Latency (Orchestration)	Service Provisioning Time (SPT)	Yes	Custom measurement script integrating with TNLCM deployment	Deliverable D5.5	Measured in Malaga

	Slice Provisioning Time (SLT)	Yes	Custom measurement script integrating with TNLCM deployment	Deliverable D5.5	Measured in Malaga
Latency (Connectivity Layer)	Control Plane Latency (CPL)	Yes	Wraparound Core testing using LoadCore Tool	Deliverable D5.3 Deliverable D5.5	Measured at all 4 platforms
	User Plane Latency (UPL)	Yes	Wraparound Core testing using LoadCore Tool	Deliverable D5.3 Deliverable D5.5	Measured at all 4 platforms
	Delay Deviation/Jitter (DDJ)	Yes	Wraparound Core testing using LoadCore Tool	Deliverable D5.3 Deliverable D5.5	Measured at all 4 platforms
	RIS Update Rate (RUR)	Yes	Internal Testing with prototype device at QUB	Deliverable D5.5	
Channel	Spectral Efficiency (SPE)	Yes	Network UE Simulation using UESIM	Deliverable D5.5	Measured in Malaga
	Received Signal Quality (RSQ)	Yes	Field Measurements using NEMO in Athens and Malaga	Deliverable D5.5 Deliverable D4.2	
EMF	Self EMF Exposure (SEE)	Yes	Device measurement at EMF lab in Oulu	Deliverable D5.5	
	Inter EMF Exposure (IEE)	Yes	Device measurement at EMF lab in Oulu	Deliverable D5.5	
	RIS Radiated Beam Directivity	No			No experiments relating to RIS Radiation analysis were conducted
	RIS Radiated Beam Peak Isotropic Power	No			
	RIS Radiated Beam Radiation Pattern	No			
	RIS Reflected/Refracted Beam EVM	No			
Reliability	Operational Network Reliability	Yes	Wraparound Core testing using LoadCore Tool	Deliverable D5.3	Measured at all 4 platforms
	Session Reliability	Yes	Wraparound Core testing using LoadCore Tool	Deliverable D5.3	Measured at all 4 platforms
	Packet Loss Rate (PLR)	Yes	Wraparound Core testing using LoadCore Tool	Deliverable D5.3 Deliverable D5.5	Measured at all 4 platforms
	Frame Loss Rate	No			Not measured specifically
Compute	Computational Resource Optimisation/Usage (CUU, RAU, ECU, SOL)	Yes	Measured with Scaphandra	Deliverable D5.5	

			powered tool interfacing with Open Nebula		
	Optimisation Periodicity	No			
Energy	Network/Device Energy Efficiency (DEE)	Yes	UXM based testbed setup	Deliverable D5.5	Measured in Malaga. Testbed developed in collaboration with 6G-VERSUS
	NFV Energy Efficiency (NEE)	Yes	Measured with Scaphandra powered tool interfacing with Open Nebula	Deliverable D5.5	
Security	Anomaly Detection Precision (ADP)	Yes	N2 and N6 ADP implemented and measured with ML models detecting synthetic attacks	Deliverable D5.5	
	Security Conformance (SEC)	Yes	Open5GS deployed through 6G-SANDBOX toolkit tested by applying 3GPP SCAS test cases	Deliverable D5.5	
	Tenant Data Privacy	No			No partner had the necessary infrastructure for this measurement
Localisation	Localisation Accuracy	No			Localisation was only measured in terms of validation of basic functionality, due to a lack of technology/necessary resources to go any further.
	Direction Accuracy	No			
	Localisation Integrity	Partial		Deliverable D5.5 Open Call Report	
Service Safety, Integrity and Maintainability	Service Safety, Integrity and Maintainability	No			Was deemed not suited for 6G-SANDBOX
AI Native Capabilities	AI/ML Accuracy (AMA)	Yes	Done with custom setup under auspices of Open Call	Deliverable D5.5 Open Call Report	
	AI/ML Training Effectiveness/Effectiveness improvement (AME)	Yes	Done with custom setup under auspices of Open Call	Deliverable D5.5 Open Call Report	
	AI/ML Precision	No			
	AI/ML Training Time	No			
	AI/ML Processing Times	No			

	AI/ML Processing Resources	No			
Sensing	6G-SANDBOX Sensing Suite (RSI, ACA, DCA, POX, VOX, OBD, EGS, ARR, VRR)	Yes	Custom setup in OULU sensing lab	Deliverable D5.5	

6 CONCLUSION

As a project whose aim was and is to produce the best and most feature complete 6G enabled testing platform, the panoply of measured KPIs covered in this Deliverable proves the success of 6G-SANDBOX in achieving its aim. Even in certain cases where a KPI has not been measured, the reasoning can be chalked up to a lack of desire or need from internal or external parties, not a lack of ability from the project's platforms.

The title of this deliverable and its stated purpose within the 6G-SANDBOX DoA is to be a final report on both KPI and KVI measurement and validation conducted over the course of the project. While the Deliverable covers a great many KPIs, no KVIs are covered. As a consortium, 6G-SANDBOX believes that KVIs are evaluated and assessed, not necessarily measured. They come from requirements set by soft sciences, which then interact with hard sciences through the means of requiring measurements from them in order to inform their assessment. Throughout the course of the project, 6G-SANDBOX was not approached by any open callers or prospective partners that had a KVI of interest they wished to evaluate that 6G-SANDBOX could apply its measurement methodologies to. The project itself is a measurement platform – it does not have any specific desires or needs with regards to KVIs, but instead exists and will continue to exist as a tool that future entities will be able to use for the generation of KPIs when assessing KVIs of import to them.

7 REFERENCES

- [1] Keysight Technologies, "P8800S UeSIM RAN Testing Toolset," 09 04 2025. [Online]. Available: <https://www.keysight.com/us/en/assets/3125-1168/technical-overviews/P8800S-UeSIM-RAN-Testing-Toolset.pdf>. [Accessed 08 06 2026].
- [2] Keysight Technologies, "RUSIM O-DU Testing Solution," 29 06 2021. [Online]. Available: <https://www.keysight.com/us/en/assets/3120-1436/data-sheets/RuSIM-UERU-Emulation-Solutions-Over-the-O-RAN-Fronthaul.pdf>. [Accessed 08 06 2026].
- [3] SRS RAN Project, "srsRAN Project Documentation," [Online]. Available: <https://docs.srsran.com/projects/project/en/latest/>. [Accessed 08 06 2026].
- [4] Keysight Technologies, "https://www.keysight.com/us/en/assets/3121-1108/data-sheets/CoreSIM-RAN-Core-Emulation-Solution-for-Wrap-Around-gNodeB-eNodeB-Testing.pdf," 04 04 2024. [Online]. Available: CoreSIM: Core Emulation Solution. [Accessed 08 06 2026].
- [5] F. J. a. S. H. a. L. M. a. R. J. a. C. M. G. a. M. F. a. C. F. a. Q. L.-C. H. a. Á.-M. C. S. a. S. D. a. K. E. J. a. F. S. Garcia, "6G DevOps: Facilitating AI and Yielding Systematic Insights in 6G Evolution," *IEEE Access*, vol. 13, pp. 176037-176046, 2025.
- [6] H. a. L. M. a. R. J. a. Á.-M. C. S. a. L.-C. H. Q. a. K. E. J. a. M. G. C. a. G. F. J. a. R. D. S. a. B. R. Sun, "Energy-Efficient Virtualized gNBs for Cloud-Native O-RAN: A Testbed-Based Study of CPU Resource Management in 5G/6G Networks," *IEEE Transactions on Green Communications and Networking*, vol. 10, pp. 1446-1461, 2026.

- [7] M. L. J. R. G. a. C. M. n. F. J. G. F. M. F. C. S. F. a. R. B. Haoxin Sun, "Efficient Resource Allocation for NextG O-RAN VNFs Based on Systematic Characterization," 08 07 2025. [Online]. Available: <https://www.techrxiv.org/doi/full/10.36227/techrxiv.175339344.49048595/v1>. [Accessed 08 06 2026].
- [8] Keysight Technologies, "6800C Series AC Power Source/Analyzers, 375-1750 VA," 19 04 2023. [Online]. Available: <https://www.keysight.com/us/en/assets/7018-05779/data-sheets/5992-2370.pdf>. [Accessed 08 06 2026].
- [9] Keysight Technologies, "Nemo Handy and Nemo Walker Air," [Online]. Available: <https://www.keysight.com/us/en/assets/3123-1442/data-sheets/Nemo-Handy-and-Nemo-Walker-Air.pdf>. [Accessed 01 06 2026].
- [10] CCI, "Octoport Multiband Antenna OPA-65R-LCUU-H6," 16 02 2016. [Online]. Available: <https://www.tempestns.com/wp-content/uploads/2023/08/CCI-OPA-65R-LCUU-H6-K-OctoPort-Multi-Band-Antenna-Datasheet.pdf>. [Accessed 02 06 2026].
- [11] Ericsson, "Radio 4408," 09 09 2019. [Online]. Available: https://www.winncom.com/docs/ericsson/Ericsson_Radio_4408_Datasheet.pdf. [Accessed 11 06 2026].
- [12] IEEE, "An Energy Efficiency Feature Survey of the Intel Haswell Processor," [Online]. Available: <https://ieeexplore.ieee.org/abstract/document/7284406>. [Accessed 19 03 2026].
- [13] Hublo, "Scaphandre Repository," [Online]. Available: <https://github.com/hubblo-org/scaphandre?tab=readme-ov-file>. [Accessed 19 03 2026].
- [14] Prometheus, "What is Prometheus?," [Online]. Available: <https://prometheus.io/docs/introduction/overview/>. [Accessed 19 03 2026].
- [15] 6G-VERSUS, "Deliverable D3.1 Experimentation Environment - Initial," 31 12 2025. [Online]. Available: https://6g-versus.eu/wp-content/uploads/2026/04/6G-VERSUS_D3.1_v1.0.pdf.
- [16] Keysight Technologies, "E7515W UXM Wireless Test Platform," [Online]. Available: <https://www.keysight.com/us/en/assets/3123-1873/data-sheets/E7515W-UXM-Wireless-Connectivity-Test-Platform.pdf>. [Accessed 19 03 2026].
- [17] Keysight Technologies, "Keysight N6700 Modular Power System Family," [Online]. Available: <https://www.keysight.com/us/en/assets/9018-04533/technical-specifications/9018-04533.pdf>. [Accessed 19 03 2026].
- [18] Free5GC, "Free5GC Open Source Core Network Implementation," Free5GC, [Online]. Available: <https://free5gc.org>. [Accessed 05 03 2026].
- [19] 3GPP, "3GPP TS 29.572 version 17.9.0 Release 17," 09 2023. [Online]. Available: https://www.etsi.org/deliver/etsi_ts/129500_129599/129572/17.09.00_60/ts_129572v170900p.pdf. [Accessed 05 03 2026].
- [20] Keysight Technologies, "Site Survey Solution for Private Networks - Application Note," [Online]. Available: <https://www.keysight.com/us/en/assets/3123-1823/application-notes/Site-Survey-Solution-for-Private-Networks.pdf>. [Accessed 11 02 2026].

- [21] Nokia, *ASMR AWGUC mmWave 2T2R n258 60dBm EIRP Technical Datasheet*, 2023.
- [22] IEC/IEEE, IEC62232 Determination of RF Field Strength, Power Density and SAR in the vicinity of Radiocommunication base stations for the purpose of evaluating Human Exposure to Radio Frequency Fields from Handheld and Body-worn wireless communication devices, BOG/ICES, 2020.
- [23] EMF Radiation Working Group, Telecommunication Engineering Centre, Department of Telecommunications, Ministry of Communications, Government of India, "TEC 31208:2023 Technical Report EMF Radiation from IoT/M2M Devices," 2023.
- [24] IEEE, "C95.1-2019/Cor 2-2020 - IEEE Standard for Safety Levels with Respect to Human Exposure to Electric, Magnetic, and Electromagnetic Fields, 0 Hz to 300 GHz - Corrigenda 2," IEEE, 2020.
- [25] M. Leinonen, "Over-the-air measurements, tolerances and multiradio interoperability on 5G mmW radio platform," University of Oulu, 27 11 2020. [Online]. Available: <https://oulurepo.oulu.fi/handle/10024/36548>. [Accessed 03 02 2026].
- [26] Lite-On Technology Corporation, "5G Total Solution," [Online]. Available: <https://www.liteon.com/en/product/networking-and-surveillance/networking-solutions/5g-total-solution>. [Accessed 16 02 2026].
- [27] Keysight Technologies, "N9040B UXA X-Series Signal Analyzer, Multi-touch," [Online]. Available: <https://www.keysight.com/us/en/assets/7018-04541/data-sheets/5992-0090.pdf>. [Accessed 03 02 2026].
- [28] A-Info, "LB-28-15-C-KF Datasheet," [Online]. Available: <https://www.ainfoinc.com/lb-28-15-c-kf-standard-gain-horn-antenna-26-5-40-ghz-15db-gain-2-92mm-female>. [Accessed 03 02 2026].
- [29] University of Giessen, "Meta Quest 2 Data Sheet," [Online]. Available: https://www.uni-giessen.de/de/studium/lehre/projekte/nidit/goals/quest2/specifications_quest-2.pdf. [Accessed 19 02 2026].
- [30] University of Giessen, "Meta Quest 3 Datasheet," [Online]. Available: https://www.uni-giessen.de/de/studium/lehre/projekte/nidit/goals/quest3/datasheet_quest-3.pdf.
- [31] Living Lab & Data Hub - Institute for the Future of Education - Tecnológico de Monterrey, "Azure Kinect DK Specifications," [Online]. Available: <https://ifeldh.tec.mx/sites/g/files/vgjovo1101/files/Azure%20Kinect%20DK%20Specifications.pdf>. [Accessed 19 02 2026].
- [32] 6G-SANDBOX, "Deliverable D2.1 Ecosystem Analysis and 6G-SANDBOX Facility Design," 30 06 2023. [Online]. Available: https://6g-sandbox.eu/wp-content/uploads/2023/07/6G-SANDBOX_D2.1_v1.5.pdf. [Accessed 19 02 2026].
- [33] C. A. C. Tharindu Ekanayake, "Physiotrack," [Online]. Available: <https://github.com/tharindu326/physiotrack/tree/main>. [Accessed 19 02 2026].
- [34] 3GPP, "TS 33.512 5G Security Specification Assurance (SCAS); Access and Mobility Management Function".

- [35] R. & C. Mitchell, "Behavior-rule based intrusion detection systems for safety critical smart grid applications. **IEE** Transactions on Smart Grid," [Online]. Available: <https://doi.org/10.1109/TSG.2013.2258438>. [Accessed 11 06 2026].
- [36] Keysight Technologies, "LoadCore - 4G/5G Core Network Test Solution," [Online]. Available: <https://www.keysight.com/us/en/assets/3120-1180/data-sheets/LoadCore-5G-Core-Testing.pdf>. [Accessed 26 05 2026].
- [37] H. C. d. R. C. B. J. M.-B. B. S. a. K. C. L. Dominato, "Tutorial on Communication between Access Networks and the 5G Core," 2022.
- [38] Keysight Technologies, "Keysight Cyperf: Scalable Network Application and Security Testing for Distributed Cloud," [Online]. Available: <https://www.keysight.com/us/en/assets/3120-1464/data-sheets/Keysight-CyPerf.pdf>. [Accessed 26 05 2026].
- [39] kafka, "AK 4.3.X," [Online]. Available: <https://kafka.apache.org/43/>. [Accessed 11 06 2026].
- [40] kafka, "AK 4.3.X / Operations / Monitoring," [Online]. Available: <https://kafka.apache.org/43/operations/monitoring/>. [Accessed 11 06 2026].
- [41] R. & K. C. & V.-G. F. & A. L. & A.-Z. J. Sedar, "A Comprehensive Survey of V2X Cybersecurity Mechanisms and Future Research Paths," *IEEE Open Journal of the Communication Society*, vol. 4, pp. 325-391, 2023.
- [42] CAR 2 CAR Communication Consortium, "Survey on CAM Statistics Technical Report C2CCC TR 2052," 2018.
- [43] CEPT ECC Working Group Spectrum Engineering, "ITS Message Sizes Contribution SE24-WI65-5-1," in *European Conference of Postal and Telecommunications Administrations*.
- [44] ETSI, "Intelligent Transport Systems (ITS); Vehicular Communications; Basic Set of Applications; Part 3: Specifications of Decentralised Environmental Notification Basic Service," 2014.
- [45] CEPT ECC Working Group Spectrum Engineering, "ITS Message Sizes Contribution SE24-WI65-5-1," in *European Conference of Postal and Telecommunications Administrations*.
- [46] J. A. F. a. L. M. P. A. M. Pereira, "Assessing the Impact of Message Generation Patterns in Vehicular Networks," in *8th International Conference on Vehicle Technology and Intelligent Transport Systems*, 2021.
- [47] ETSI, "Intelligent Transport Systems; Vulnerable Road Users Awareness; Part 3: Specifications of VRU Awareness BAsic Service," 2021.
- [48] 3GPP, "TR 26.985 V18.0.0 5G Media Streaming for Advanced Automotive and Extended Reality Use Cases; Traffic Models and Service Requirements," 2024.
- [49] OpenTAP, "What is OpenTAP," [Online]. Available: <https://doc.opentap.io>. [Accessed 26 05 2026].
- [50] Open5GS, "Welcome to Open5GS," [Online]. Available: <https://open5gs.org>. [Accessed 10 02 2026].
- [51] A. Gungor, "UERANSIM," [Online]. Available: <https://github.com/aligungr/UERANSIM>. [Accessed 10 02 2026].

- [52] 6G-SANDBOX, "Deliverable D5.3 - Intermediate Report on KPI and KVI Validation," 31 08 2024. [Online]. Available: https://6g-sandbox.eu/wp-content/uploads/2025/04/6G-SANDBOX_D5.3_V1.pdf. [Accessed 08 06 2026].
- [53] 6G-SANDBOX, "Deliverable D4.2 - Integration and Management for 6G Trial Networks - Final Report," 30 06 2025. [Online]. Available: https://6g-sandbox.eu/wp-content/uploads/2025/10/6G-SANDBOX_D4.2_v1.0.pdf. [Accessed 01 06 2026].
- [54] D. R. L. P. M.-d.-M. G. Rafael Lopez-Gomez, "Towards a formal digital twin of the PTP protocol using automata learning," *Journal of Logical and Algebraic Methods in Programming*, vol. 151, 2026.