Coordinated Control and Spectrum Management for 5G Heterogeneous Radio Access Networks

Grant Agreement No. : 671639
Call: H2020-ICT-2014-2

Deliverable D6.2
Final Report on Technical Validation

Version:	3.0
Due date:	31.3.2018
Delivered date:	7.6.2018
Dissemination level:	PU

The project is co-funded by

Authors

Markku Savela, Tao Chen (VTT); Chia-Yu Chang, Navid Nikaein, Konstantinos Alexandris (EUR); Roberto Riggio (FBK); Akhila Rao, Rebecca Steinert (SICS); Antonio Cipriano, Dorin Panaitopol (TCS); Dimitry Marandin (CMA); Sundar Daniel Peethala (UDE); George Agapiou (OTE); Adrian Kliks, Lukasz Kulacz (PUT); Jaakko Ojaniemi, Heikki Kokkinen (FS), Fang-Chun Kuo, Kostas Pentikousis (TP)

Editors

Dimitri Marandin (CMA), Navid Nikaein (EUR)

Coordinator

Dr. Tao Chen

VTT Technical Research Centre of Finland Ltd

Tietotie 3

02150, Espoo

Finland

Email : tao.chen@vtt.fi

Disclaimer

Acknowledgement

This report is funded under the EC H2020 5G-PPP project COHERENT, Grant Agreement No. 671639.

Version history

Version	Date	Remarks
1.0	1/10/2017	ToC created
1.1	1/11/2017	ToC finalized
1.2	16/2/2018	Experiment descriptions added
1.3	2/3/2018	First version of results added
1.4	20/3/2018	Revised section
1.5	1/5/2018	First complete draft ready for review
1.6	10/5/2018	Review comments addressed
2.0	15/5/2018	Nearly-ready version
2.1	25/5/2018	Revised with track changes
3.0	7/6/2018	Final version

Executive summary

The aim of this deliverable is to provide the descriptions of the experiments in the COHERENT project and the analysis on the obtained experimental results. The deliverable describes two reference implementations of the COHERENT C3 as an anchor for all experiments: the FlexRAN controller and the 5G-EmPOWER platform. Using COHERENT C3, six proof-of-concepts (PoC) have been built during the project and their descriptions are provided including used hardware, high-level description of the main software components, test scenarios and evaluation results.

Specifically, these PoCs have different targets:

- PoC-1 RAN Monitoring and Proactive Control: Gather network state information through Network Information Function (NIF) and aggregate them into radio access technology (RAT) agnostic network graph for RAN monitoring and proactive control;
- PoC-2 RAN Inter-app Communications: Inter-app communications of RAN control apps;
- PoC-3 RAN Sharing: RAN slicing for Public Safety (PS) application;
- PoC-4 Load Balancing: An open programmable platform used for LTE and Wi-Fi load balancing scenarios;
- PoC-5 Spectrum management: Spectrum Management Application (SMA) with the purpose of dynamically switching on/off the phantom-cell and spectrum sharing between the operators and vertical sectors;
- PoC-6 Distributed Antenna System: Distributed Antenna System (DAS) with coordinated UE pairing for improving/maintaining per user throughput and coverage extension.

These PoCs are evaluated in different wireless technologies in order to address multiple use cases and carry out the assessment of the COHERENT concepts. Practically, quantifiable evaluation results are provided in all experiments to be aligned with 5G-PPP KPIs and objectives. They represent as a basis for dissemination and exploitation activities. Particularly, the experimental setups have been presented in multiple events.

Table of contents

List of abbreviations

2G	2nd Generation Mobile Networks
3D	Three-Dimensional
3GPP	Third Generation Partnership Project
4G	4th Generation Mobile Networks
5G	5th Generation Mobile Networks
5GPoAs	5G Point of Access
5G PPP	5G Infrastructure Public Private Partnership
AI	Air Interface
ANDSF	Access Network Discovery and Selection Function
API	Application Programming Interface
ARPU	Average Revenue Per User
AS	Access Stratum
ASA	Authorised Spectrum Access
BBU	Baseband Unit
BM-SC	Broadcast Multicast Service Centre
BS	Base Station
BSS	Base Station System
C3	Central Controller and Coordinator
CA	Carrier Aggregation
CAL	CHARISMA Aggregation Level
CAPEX	Capital Expenditure
CBRS	Citizens Broadband Radio Service
CC	Component Carrier
CDN	Content Delivery Network
CEPT	European Conference of Postal and Telecommunications Administrations
CESC	Cloud-Enabled Small Cell
CESCM	Cloud Enabled Small Cell Manager
CMOS	Complementary Metal Oxide Semiconductor
CN	Core Network
CoMP	Coordinated Multipoint
CP	Control Plane
CPE	Customer Premises Equipment
C-RAN	Cloud Radio Access Network
CriC	Critical Communications
D2D	Device-to-Device
DC	Data Centre
DAS	Distributed Antenna System

DLSCH	Downlink Shared Channel
DMO	Direct Mode Operation
DPI	Deep Packet Inspection
D-RAN	Distributed Radio Access Network
ECC	Edge Cloud Computing
eICIC	enhanced Inter-Cell Interference Coordination
eNodeB	Evolved Node B
eNodeB	Evolved Node B
eMBB	enhanced Mobile Broadband
eMBMS	enhanced Multimedia Broadcast Multicast Service
EMS	Element Management System
EPC	Evolved Packet Core
E-UTRAN	Universal Terrestrial Radio Access Network
EVM	Error Vector Magnitude
eV2X	enhanced Vehicle-to-X Communications
EPC	Evolved Packet Core
ESO	Essential Service Operator
ETP	European Technology Platform
ETSI	European Telecommunications Standards Institute
EU	European Union
FDD	Frequency-Division Duplex
FICORA	Finnish Communications Regulatory Authority
ForCES	Forwarding and Control Element Separation
GPP	General Purpose Processors
GSM	Global System for Mobile Communications
GW	GateWay
GWCN	Gateway Core Network
H2H	Human-to-Human
HAN	Heterogeneous Access Network
HMN	Heterogeneous Mobile Network
HSS	Home Subscriber Server
ICIC	Inter-Cell Interference Coordination
IEA	International Energy Agency
IEEE	Institute of Electrical and Electronics Engineers
IETF	Internet Engineering Task Force
IMT	International Mobile Telecommunications
IMU	Intelligent Management Unit
IoT	Internet of Things
IP	Internet Protocol

IRTF	Internet Research Task Force
ISI	Integral SatCom Initiative
ITRS	International Technology Roadmap for Semiconductors
ITU	International Telecommunication Union
ITU-R	ITU – Radio Communication Sector
IWF	Interworking Function
KPI	Key Performance Indicator
L1	Layer 1
L2	Layer 2
LSA	Licensed Shared Access
LSN	Last Sequence Number
LTE	Long-Term Evolution
LTE-A	Long-Term Evolution Advanced
M2M	Machine-to-Machine
MAC	Media Access Control
MANO	MANagement and Orchestration
MBMS	Multimedia Broadcast Multicast Service
MCS	Modulation and Coding Schemes
MEC	Mobile Edge Computing
MIMO	Multiple-Input Multiple-Output
MIoT	Massive Internet of Things
MMC	Massive Machine Communication
MME	Mobility Management Entity
MN	Mobile Network
MOCN	Multi-operator Core Network
MTC	Machine-Type Communication
MTD	Machine-Type Device
NEO	Network Operation
mmWave	Millimeter Wave
MNO	Mobile Network Operator
MOCN	Multi-operator Core Network
MoI	Ministry of Interior
MVNO	Mobile Virtual Network Operator
Naas	Network as a service
NAS	Non-Access Stratum
NEO	Network Operation
NFV	Network Function Virtualisation
NFVO	Network Functions Virtualisation Orchestrator
NFVI	Network Function Virtualisation Infrastructure

NGMN	Next Generation Mobile Networks Alliance
NRA	National Regulation Agency
NVS	Network Virtualisation Substrate
Ofcom	Office of Communications
OFDM	Orthogonal Frequency Division Multiplexing
OFDMA	Orthogonal Frequency Division Multiple Access
ONF	Open Networking Foundation
OPEX	Operational Expenditure
OSI	Open Systems Interconnection
OSS	Operations support systems
PCC	Primary Component Carrier
P-GW	PDN gateway
PHY	Physical Layer
PLMN	Public Land Mobile Network
PMR	Private (or Professional) Mobile Radio
PNF	Physical Network Function
PON	Passive Optical Network
PoP	Point of presence
PPDR	Public Protection and Disaster Relief
PRACH	Physical Random Access Channel
ProSe	Proximity Services
PTT	Push-To-Talk
PTS	Swedish Post and Telecom Agency
QoE	Quality of Experience
QoS	Quality of Service
RAN	Radio Access Network
RANaaS	Radio Access Network as a Service
RAT	Radio Access Technology
RAU	Radio Access Unit
RB	Resource Block
RF	Radio Frequency
RN	Relay Node
RT	Radio Transceiver
R-TP	Radio Transmission Point
RRH	Remote Radio Head
RRM	Radio Resource Management
RRU	Radio Remote Unit
RTC	Real-Time Controller
SA	Service and System Aspects

SAS	Spectrum Access System
SC	Small Cell
SCC	Secondary Component Carrier
SCS	Service Capability Server
SDK	Software Development Kit
SDN	Software Defined Network
SD-RAN	Software-Defined Radio Access Network
S-GW	Serving gateway
SID	Service Integration Driver
SLA	Service Level Agreement
SLP	SUPL Location Platform
SMA	Spectrum Management App
SME	Small and Medium Enterprises
SON	Self Organizing Network
SONAC	Service-Oriented Virtual Network Auto-Creation
SUPL	Secure User Plane Location
TCO	Total Cost of Ownership
TDD	Time-Division Duplex
TEDS	TETRA Enhanced Data Service
TETRA	Terrestrial Trunked Radio
TN	Transport Node
TR	Technical Report
TDM	Time Division Multiplexing
TR	Technical Report
UAV	Unmanned Aerial Vehicles
UE	User Equipment
UM	Unacknowledged Mode
UMTS	Universal Mobile Telecommunications System
UP	User Plane
V2X	Vehicle-to-X Communications
VIM	Virtualised Infrastructure Manager
VM	Virtual Machine
VNF	Virtual Network Function
VNFD	VNF Descriptors
VNFM	VNF Manager
vBBU	virtual Baseband Unit
vBSC	virtual Base Station Controller
VIM	Virtualised Infrastructure Manager
VNF	Virtual Network Function

VNO	Virtual Network Operator
VPL	Vehicular Penetration Loss
Wi-Fi	Wireless Fidelity
WiMAX	Worldwide Interoperability for Microwave Access
WLAN	Wireless Local Area Network

List of figures

List of tables

1. Introduction

1.1 Motivation, Objectives and Scope

This deliverable provides the final description of the experiments in the COHERENT project and presents the obtained experimental results. This deliverable is related to all Tasks in WP6, namely *Testbed Integration and Setup* (T6.1), *Execution of the Experiments* (T6.2), *Analysis of the Results and Technical Impact* (T6.3) with more effort associated with T6.3. The previous deliverable in WP6 (Deliverable D6.1 [1]) defines the steps required in order to build an integrated COHERENT testbed, which serves as input for the current deliverable.

COHERENT testbed integration is related to both the data plane and control plane procedures that are required to carry out the demonstrators, evaluations and assessments of the technical approach. The data plane and network technologies (LTE and Wi-Fi networks) provide transport services to the modules built on top. An abstract network view (or network graph as defined in Deliverable D2.2 [2]) exposes information of the physical and virtualized networks to the control plane. In the control plane the logical operations are carried over the converged multi-RAT network. In COHERENT, a logically centralized entity called Central Controller and Coordinator (C3) is in charge of logically centralized network-wide control and coordination among network entities in RAN, based on centralized network view (see [2] for more details). Information related to both operational and configuration data is exposed as *network graphs.*

This deliverable describes two reference implementations of the COHERENT C3 as anchors for all experiments: the FlexRAN platform and the 5G-EmPOWER platform. Using COHERENT C3, the six proof-of-concepts (PoCs) are used to support the integrated COHERENT testbed:

- **PoC-1 RAN Monitoring and Proactive Control**: Gather network state information through Network Information Function (NIF) and aggregate them into radio access technology (RAT) agnostic network graph for RAN monitoring and proactive control;
- **PoC-2 RAN Inter-app Communications:** Inter-app communications of RAN control apps;
- **PoC-3 RAN Sharing:** RAN slicing for Public Safety (PS) application;
- **PoC-4 Load Balancing:** An open programmable platform used for LTE and Wi-Fi load balancing scenarios;
- **PoC-5 Spectrum management:** Spectrum Management Application (SMA) with the purpose of dynamically switching on/off the phantom-cell and spectrum sharing between the operators and vertical sectors;
- **PoC-6 Distributed Antenna System:** Distributed Antenna System (DAS) with coordinated UE pairing for improving/maintaining per user throughput and coverage extension.

In this document, technical descriptions of PoCs are provided including used hardware, high level description of the main software components, test scenarios and evaluation results.

The components of each PoC use the input from other WPs and realize the relevant components of the COHERENT architecture described in Deliverable D2.2 [2]. The concept of COHERENT network abstraction with Network Graphs was a part of the WP3 activities. The WP4 activities are related to spectrum management and sharing, while in WP5 the partners combine their expertise in lower layers for mobile networks and software defined networking (SDN) to explore the programmable control methods and algorithm implementations for inter-cell coordination, resource allocation, and control plane resiliency. The results from WP3, 4, and 5 served as the technical base for the integrated testbed in WP6.

1.2 Structure of the Document

- **Chapter 2** describes the integrated testbed. First, the involvement of blocks of COHERENT architecture into PoC and addressed use cases are described in Section 2.1. Two reference implementations of the COHERENT C3 as anchors for all experiments: the FlexRAN platform and the 5G-EmPOWER platform are described in Section 2.2 followed by the description of the six PoCs (From Sections 2.3.1 to 2.3.6) correspondingly.

- **Chapters 3 – 8** provide evaluation results of six integrated PoCs. In every demonstration, several experiments are conducted and early results have provided the feedbacks to other work packages.
- **Chapter 9** summarizes the relevance of PoC to 5G PPP KPIs and other 5G KPIs.
- **Chapter 10** concludes the deliverable.

2. Integrated Testbed

This chapter starts with the COHERENT user cases, and then describes how COHERENT architecture is reflected in COHERENT testbeds. Two reference implementations of the COHERENT C3 are as anchors for all experiments: the FlexRAN controller and the 5G-EmPOWER platform are described in Section 2.2. They are the convergence points of all PoCs, and thus all PoC are integrated with either of these C3 realizations. Both C3 realizations include the Network Graph concept designed in WP3 and Section 2.2.3.1 and 2.2.3.2 elaborate on the realization details in both FlexRAN controller and 5G-EmPOWER correspondingly. Finally, six PoCs are described in Section 2.3.1 to 2.3.6. Detailed evaluation results of PoCs are given in Chapters 3 – 8.

2.1 Addressed uses cases and architectural components

The integrated testbed addresses following 5 main use cases:

- Spectrum management (PoCs: RAN visualization, spectrum management),
- RAN sharing (PoC: RAN Sharing, PoC: RAN Monitoring and Proactive Control),
- Load balancing (PoC: Load balancing),
- Distributed Antenna System (PoC: Distributed antenna system), and
- Proactive and/or remote control (Used in conjunction with other PoC)

These use cases have been carefully selected by the project partners because of their importance towards integrated 5G communications. A detailed description of each use case is provided in Section 6 of D6.1 [1]. Figure 1 provides a high-level view of the architectural components involved in each PoC. Note that each use case demonstrator is a collaboration action between groups of partners, who complement each other as they tackle a common problem. As per the project plan, WP6 partners share resources and knowledge and mutually benefit from working together.

2.2 C3 design and implementation

Two reference implementations of the COHERENT C3 have been developed within the project. Each of them focuses on different portions of the problem space. The FlexRAN platform focuses on the LTE domain and supports both real-time and non-real-time applications. Conversely, the 5G-EmPOWER platform supports both Wi-Fi and LTE (both open-source and commercial), but only non-real-time applications can be used on this platform. Both platforms support multi-tenancy in the form of RAN slicing and allow implementing spectrum management and mobility management applications, among other features. In this section we provide a short overview of the two platforms for a more detailed description we refer the reader to COHERENT D2.4 [20].

2.2.1 C3 design and implementation in 5G-EmPOWER Platform

5G-EmPOWER is a Multi-access Edge Computing Operating System (MEC-OS) supporting lightweight computing virtualization and heterogeneous radio access technologies [3] [4] [5]. A high-level view of the 5G-EmPOWER system architecture is depicted in Figure 2. The architecture is logically divided into three layers. The bottom layer consists of the physical and virtualized resources composing the data-plane. The second layer runs 5G-EmPOWER OS which controls the physical and virtual resources in the data-plane. The third layer deals with the actual slices.

5G-EmPOWER is a reference implementation of the C3, but it also targets Lightweight NFV Management and Orchestration and Multi-access Edge Computing. The details of these features are out of the scope of this document and are thus omitted.

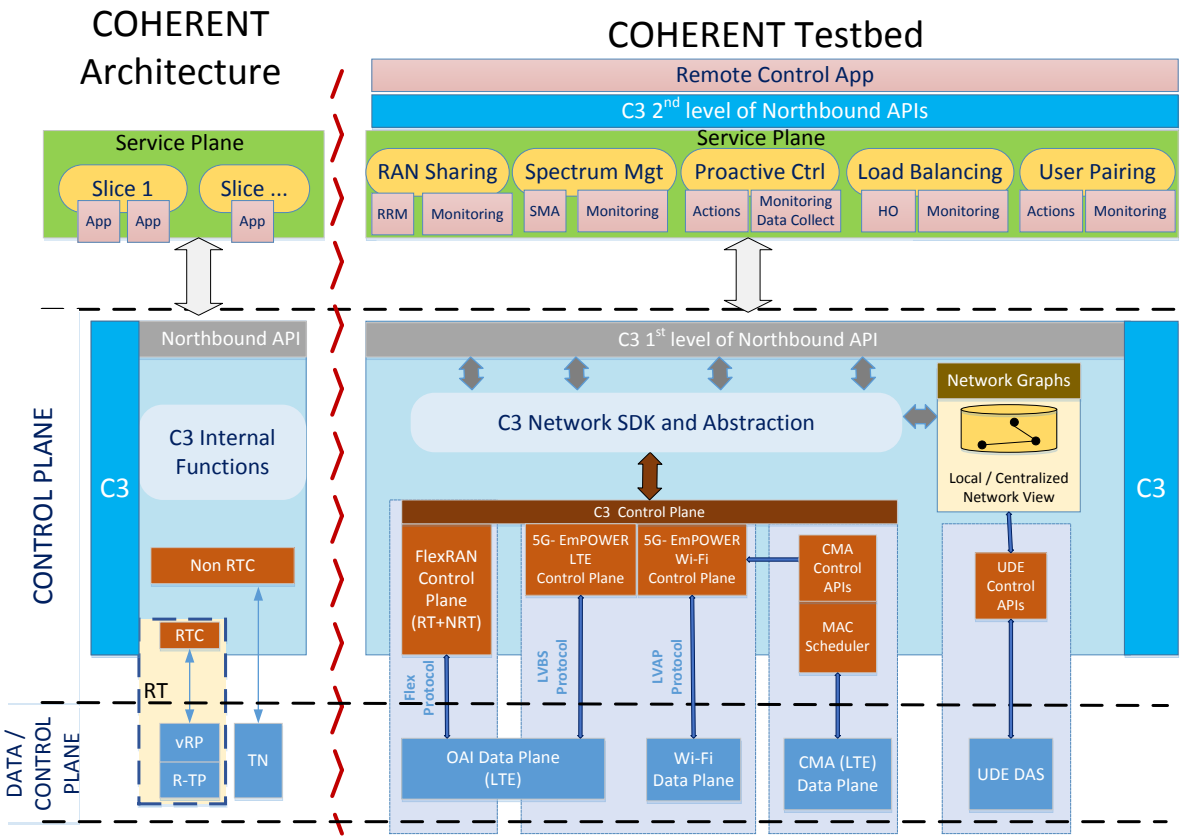


Figure 1: COHERENT Testbed and PoCs

The entire software stack has been released under a permissive APACHE 2.0 for academic use and is available at the official 5G-EmPOWER website: http://empower.create-net.org/. Documentation and tutorials are available in the associated GitHub project: https://github.com/5g-empower/5g-empower.github.io/wiki. Commercial use of the platform is regulated by the APACHE 2.0 License.

The communication between the 5G-EmPOWER MEC-OS and the data-plane device happens over a persistent TCP connection. The definition of the protocol and of the various messages is too verbose for this report and has been made available online under the same license as the rest of the platform: https://github.com/5g-empower/5g-empower.github.io/wiki/EmPOWER-Protocol.

The 5G-EmPOWER MEC-OS supports multi-tenancy in the form of RAN slicing. Users can use either the web-based dashboard or a REST interface in order to perform CRUD (Create, Retrieve, Update, and Delete) operations on a network slice. The technical details about how slicing is performed will be provided by WP5 Deliverable D5.2 [21] In this section, we will report on the slicing interface exposed by the 5G-EmPOWER OS.

The 5G-EmPOWER platform supports four types of data-plane devices:

1. **Wireless Termination Points (WTPs).** Each WTP consists of two components: one OpenVSwitch instance implementing the backhaul data; and one Agent implementing the 802.11 data–path. The Agent is implemented using the Click Modular Router. Click is a framework for writing multi–purpose packet processing engines and is being used to implement just the WTPs/Wireless Clients frame exchange, while all the decision logic is implemented at the EmPOWER Runtime. Communications between the Agent and the Runtime take place over a persistent TCP connection. Any Wi-Fi capable of running the OpenWRT OS can become an EmPOWER WTP. Instruction on how to configure a WTP can be found: https://github.com/5g-empower/5g-empower.github.io/wiki/Setting-up-the-WTP.

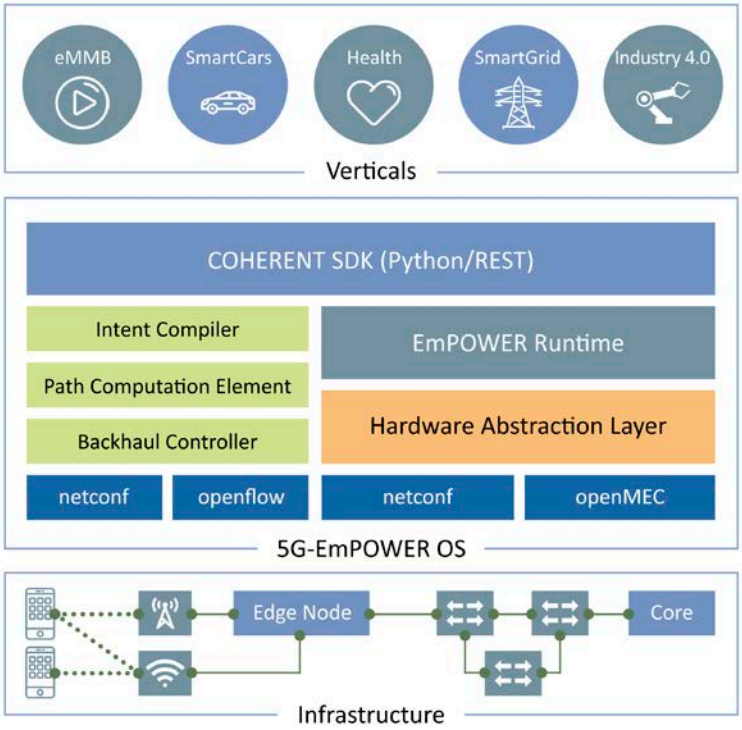


Figure 2: The 5G-EmPOWER System Architecture

2. **Click Packet Processors (CPPs**). These nodes essentially combine an OpenFlow switch with a general purpose x86 CPU. As the name suggests, CPPs leverage on multiple instances of the Click Modular Router in order to implement packet processing. Each CPP includes an OpenVSwitch instance, one or more Click instance, and one Agent. The latter is in charge of monitoring the status of each Click instance as well as of handling the requests coming from the controller. Any x86 machine capable of running Linux can become a CPP. Instructions on how to configure a CPP can be found online[1].

3. **Virtual Base Stations (VBSes).** These are essentially LTE eNodeBs. The 5G-EmPOWER platform currently supports both open and commercial LTE small cells. In order to communicate with the EmPOWER Runtime, the small cells must implement the EmPOWER Agent. Such Agent, which is part of the EmPOWER platform, has a modular architecture allowing it to be adapted to different platforms (both open source and commercial). Instruction on how to configure a VBS can be found online[2].

4. **Transport Nodes (TN).** These are the entities located between RTs and core network. A set of TNs is forming a backhaul/fronthaul network whose data plane can be configured by the 5G-EmPOWER. An OpenFlow switch is an example of Transport Node.

2.2.2 C3 design and implementation in FlexRAN

The reference implementations of the COHERENT in FlexRAN/OAI platform are presented in this section. Firstly, the FlexRAN architecture is described with the special focus on real-time control aspect. Furthermore, we summarize FlexRAN API, which offers a set of functions that stand for the southbound APIs. Finally, we show the integration with OAI/FlexRAN control plane and Control and User Plane Separation (CUPS) based EPC for controlling the core network.

[1] *https://github.com/5g-empower/5g-empower.github.io/wiki/Setting-up-the-CPP*
[2] *https://github.com/5g-empower/5g-empower.github.io/wiki/Setting-up-the-VBS*

2.2.2.1 FlexRAN architecture

FlexRAN is a software-defined RAN (SD-RAN) platform that aims to support centralized real-time control for many RAN-related operations, e.g., one for monolithic 4G eNodeB, or multiple for a disaggregated 4G and 5G, as well as the capability of implementing new control functions based on NFV principles. It currently supports OpenAirInterface (OAI) and is released under Apache v2.0 license. FlexRAN control plane follows a two-layered design as depicted in Figure 3. Specifically, it is composed of a Real-time Controller (RTC) and RAN Runtime. RAN Runtime can act as a local agent with network view controlled by the RTC (master/slave relationship), or acting jointly with other local agents and the RTC. More, the RTC can be connected to a number of local agents; one for each RAN node. The control and data plane separation is provided by the RAN Runtime API which acts as the southbound API with FlexRAN control plane on one side and BS data plane on the other side. Such communication is provided by FlexRAN Protocol that facilitates the communication between the RTC and the RAN Runtime. RAN control can be developed allowing a bi-directional interaction between the controller and the local agents. Especially, RTC can issue control commands that define the RAN Runtime operation and the RAN Runtime can reply with the corresponding messages. Detail descriptions of FlexRAN can be found: http://mosaic-5g.io/flexran/

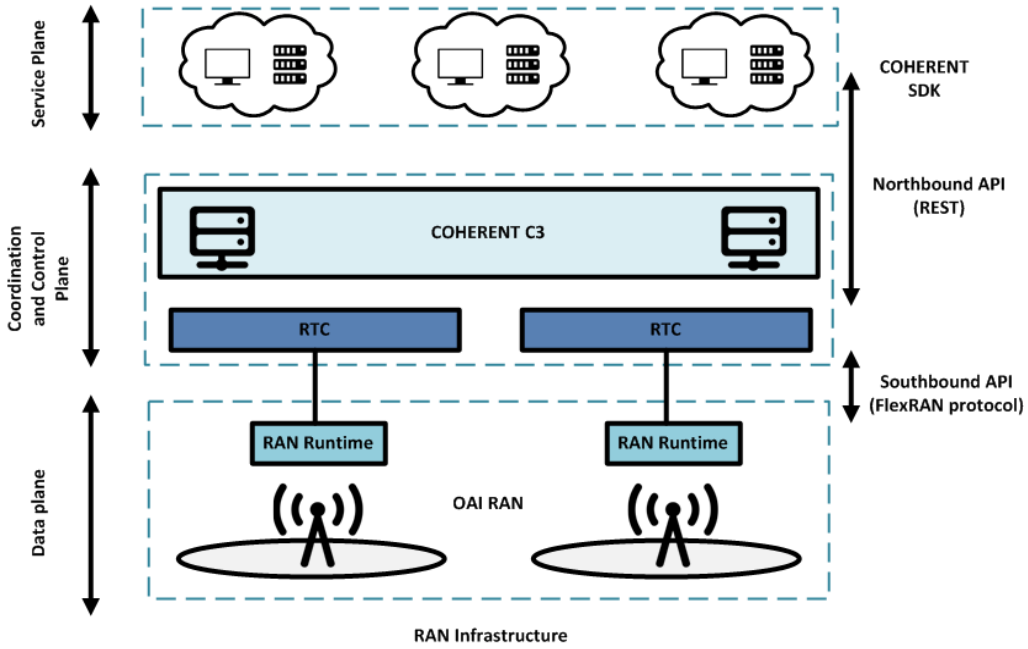


Figure 3: FlexRAN Architecture

On top of RTC lies a northbound API, which allows RAN control applications messages to be sent in RTC and modify the state of the underlying network infrastructure or collect any network information is requested via RAN Runtime. To this end, the type of the applications can be divided into two main categories: a) Soft real-time applications, e.g., monitoring/statistics to report any BS information implying low-level of criticality and b) Hard real-time applications, e.g., MAC scheduler to change the state of any BS implying high level of criticality. FlexRAN can be characterized as a reference for COHERENT RTC implementation interplaying in real-time with RAN Runtime (via Southbound API) as well as with the corresponding control applications (via Northbound API) for network local control. Such applications can directly interact with COHERENT RTC; especially in the case of hard real-time ones. In other cases, COHERENT C3 can communicate with RTCs for centralized control rendering it a liaison between network services and local RAN control; especially in the case of soft real-time applications via a logical control framework. Complementary to that, FlexRAN can also act as a C3 physical distributed control instance being in charge of centralized network control via direct interaction with the service plane applications.

2.2.2.2 FlexRAN APIs

To control and manage the BS user-plane (UP), FlexRAN API offers a set of functions that stand for the southbound APIs. These functions allow the control-plane (CP) to interplay with the user-plane (UP) as follows (see Table 1). More details can be found: http://mosaic-5g.io/apidocs/flexran/

API#	Target	Direction	Example	Applications
Get config	eNodeB,UE	Controller→RAN Runtime	• Get UL/DL RAN configuration • Get DRB configuration • Get Measurement Configuration	• Monitoring
Get stats	eNodeB,UE	RAN Runtime→Controller	• Channel Quality Indicator • SINR/RSSI measurements / Localization • UL/DL performance at specific layer	• Monitoring • Optimization
Commands	eNodeB,UE	Controller→RAN	• Scheduling policy • Handover policy • Radio Resource Management Policy	• Slicing (creation/update/deletion)
Event trigger	eNodeB,UE	RAN Runtime→Controller	• UE attachment • Scheduling Request • Measurement Events • Cell load	• Monitoring • Control Action
Control Delegation	eNodeB	Controller→RAN runtime	• Delegate radio resource management • Update RAN DL/UL scheduling • Update HO algorithm	• Programmability • Slicing and multi-service

Table 1: FlexRAN API

It is noted that RAN Runtime will be in charge of retrieving the cell and user related information from the underlying RAN and can trigger events when a state changes. It can be observed that different types of network applications, ranging from pure network monitoring to network control and optimization can be deployed using the provided abstractions via control modules. Such control modules have been designed to acquire abstracted information and support control applications being agnostic to the underlying access infrastructure.

2.2.2.3 Integration with CUPS based EPC in the OAI and FlexRAN platforms

Traditionally, the monolithic network entities in the mobile core are designed to handle several millions of users at the same time with high availability on a physical single node. They are implemented as monolithic pieces of software running on dedicated hardware where the control plane (CP) and user plane (UP) functions[3] are collocated. Nevertheless, there are limitations in such design. For example, in a softwarized implementation of PGW at the LTE EPC, a single software instance (hosted on a virtual or bare metal machine) can quickly become a bottleneck under high load or failure. With the introduction of NFV, SDN and cloud computing, mobile core networks open the door to revisit the deployment architecture. In particular, the introduction of Control and User Plane Separation (CUPS) for EPC nodes in Release 14, allows for better (and independent) scaling of the control and user plane functionality as well as its evolution both from a software deployment practice but possibly also from a standards viewpoint.

Although the main focus in COHERENT is to provide agility and programmability for RAN, the architecture of COHERENT could be exploited to operate the core network by integrating OAI/FlexRAN control plane with CUPS based EPC. Figure 4 illustrates how the COHERENT C3 integrates with a CUPS based EPC deployment where one (or more) PGW control plane (CP) elements employ multiple user plane (UP) elements to handle end-user traffic. The associated architecture reuses the existing interfaces in FlexRAN as well as the 3GPP standard interfaces implemented in OAI [6] and Travelping erGW [25]. The Travelping erGW is an open-source project with the goal to deliver a 3GPP GGSN and PGW implemented in Erlang. It strives to eventually support all the functionality as defined by 3GPP TS 23.002 Section 4.1.3.1 for the GGSN and Section 4.1.4.2.2 for the PGW.

[3] *The examples of the CP functions in the PGW are path/tunnel setup and mobility management while UP functions are data packet forwarding/processing.*

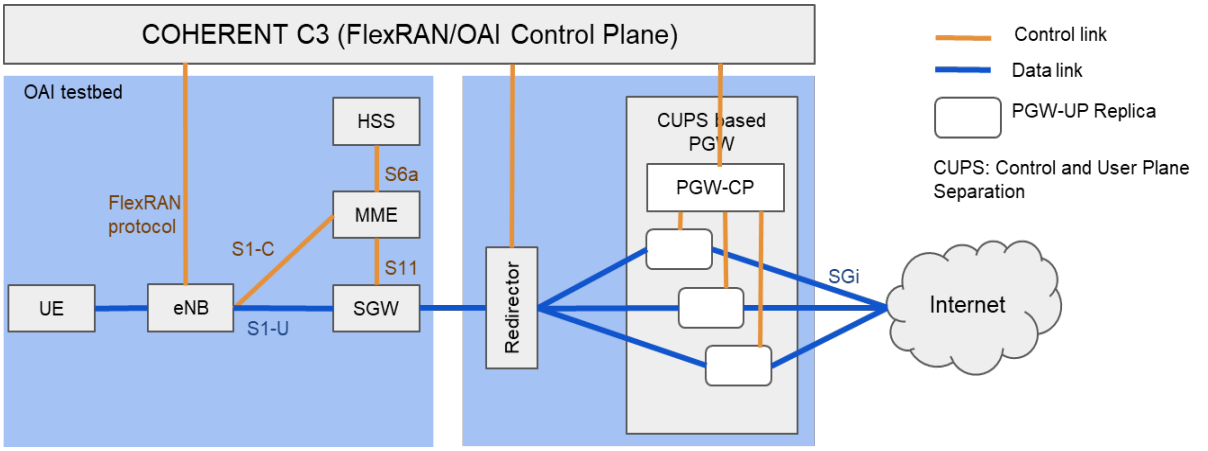


Figure 4: Load balancing with clustered PGW Testbed

The integration with CUPS based EPC allows COHERENT C3 to coordinate traffic between the core network and the radio access network. The PGW CP interwork with OAI control plane for retrieving the network information of RAN or update network information in the core network. Based on the actual traffic load, PGW CP can dynamically scale out/in the replicas of PGW UP. In the integrated architecture, there is a front-end redirector entity acting as a load balancer to distribute incoming traffic across multiple replicas for CUPS based PGW. The main features of the integrated system include:

- Decoupling the C/U-plane in the entity using an SDN-based approach.
- A clustered PGW appears a logical entity the network. However, it is internally implemented one PGW CP and multiple replicas of PGW UP. The replicas of PGW UP can scale out/in according to load or different reliability request.
- One redirector is placed in front of the PGW logical entity to distribute the incoming traffic amongst the multiple replicas of PGW UP.
- The redirector and the PGW CP interwork with C3.

In our implementation, the PGW CP and UP are containerized with Docker and the images are publicly available in [27] and [28]. The CUPS-based PGW is deployed on IBM Cloud with 4 GB of memory and 2 AMD64 CPU. The container runtime version is Docker 17.3.2. The system is managed by the orchestration tool, Kubernetes with version 1.10.1. To provide high flexibility in scaling out/in CUPS based PGW on demand, it is necessary to quickly manage the lifecycle of PGW instances. Following table shows the time required to create and start one container on the same host. The difference between *create* and *start* is that *start* is the time for the network functions (PGW CP or PGW UP) to bootstrap after being created in the cloud and *create* is the time for initializing the network function instance from scratch, including polling the container image, creating the containers in the cloud and network function bootstrapping time. For fast reaction to host failure or increasing traffic demand, it could enable/disable new PGW CP or PGW UP by proactively creating them and starting them when required.

	Create (Seconds)	Start (Seconds)
PGW CP	34.6	18.9
PGW UP	34.3	16.7

Table 2: Measurements of boot time for the Docker containers of PGW CP and PGW UP in the Kubernetes Cluster.

2.2.3 Network Graphs

2.2.3.1 Network Information Function and network graph in the 5G-EmPOWER and OAI/FlexRAN Platform

The network information function (NIF) as defined in Section 5.4 in COHERENT Deliverable D2.2 [2] is a building block of the network graph, which performs: collection, processing, and abstraction

of data sources for appropriate representation of the underlying network state. This network state information in the form of a network graph is accessible to controller applications at both the RTC and the C3 as defined in D2.2 [2].

Our probabilistic attainable throughput based NIF, described in detail in COHERENT Deliverable D3.2 [7], aggregates network state information over multiple RANs to form a RAT agnostic network graph at the RTC and C3. Control and management decisions, e.g., mobility management in the form of handover decisions are made based on the NIF in the network graph.

Generation of network graph

A local monitoring instance on the serving node (eNodeB/AP) aggregates and communicates base metrics to the controller (FlexRAN/5G-EmPOWER) as required. A monitoring instance at the controller collects and aggregates this data into a consolidated network state. The objective of the NIF is to monitor the variability in the network state caused due to dynamic changes in radio propagation environment, dynamic user density, bandwidth requirement etc. An average measurement of network state over these variations is insufficient to capture the impact of variability on user Quality of Service (QoS), e.g., throughput in our PoC. By measuring the probabilistic characteristics of throughput over serving links and estimating probabilistic characteristics of attainable throughput on candidate links we quantify and react to the variability in the network. This allows controller applications to provide better adherence to user QoS by making informed decisions about resource allocation, user associations, etc.

Control and monitoring applications of NIF generated network graph

Proactive control and management applications on the RTC/C3 use this NIF-generated network graph to trigger appropriate network action. The trigger is based on the tolerance levels and throughput thresholds that quantify the accepted variability in the users' QoS. Performance on serving links are compared with estimated performance on candidate links to trigger proactive management and control. A detailed description of the applied estimation models and the approach of using it in a probabilistic framework to make control and management decisions has been previously reported in the COHERENT Deliverable D3.2 [7] and published in [9].

The abstracted representation of the network state in the NIF-generated network graph can be used for various applications that improve network and user performance. Handover of users between serving nodes based on meeting user QoS is one proactive monitoring application that has been evaluated, with results presented in Sections 3.1 and 3.2. It describes how the network graph was used to trigger/perform intra-RAT handover in LTE and Wi-Fi networks on OAI and 5G-EmPOWER testbeds, respectively. The NIF can be used to perform a range of other controller functions such as proactive resource allocation to users such that each user's QoS is met within the previously described probabilistic framework. Users with low tolerance for variability can be assigned more resources in accordance with their strict requirements. User association sets can also be created and updated to maximize the adherence to QoS requirements of each user in the network, wherein users with low tolerance for variability being scheduled on low variance links and users with requirements that are not as strict being scheduled on links with higher variability.

These applications demonstrate the need for a probabilistic representation of the network state as generated by NIF in effectively handling the uncertainty of varying links. It provides the controller applications with the tools to proactively and effectively manage this variability and meet strict user QoS requirements. All the aforementioned control and management actions can be evaluated over heterogeneous multi-RAT networks allowing for these actions to be optimized over serving nodes of different access technologies.

2.2.3.2 Network Graphs realization in FlexRAN

Data collection

COHERENT project targets wireless networks with millions of network elements and billions of end users that are connected through multi-RAT infrastructures. Applications that are running on the top of the underlying network can collect end-users and RAN statistics or network control information of the RAN systems that constitute the so-called network information. Such information can be represented in the form of Network Graphs as it can be further used for offline processing or real-time use cases in order to provide interactive information control in complex environments. Data collection can be triggered by network services applications that are running on the top. Any network information represented as Network Graph is stored in a database (DB), the so-called Network Graph DB and it can be used to extract complex network agnostic information so as to optimize network control decisions.

Network Graph Infrastructure

The main architecture components in Network Graph Infrastructure are presented in the following figure, which depicts the network components and services that are interacting with each other in order to collect the RAN information requested by the higher network layers (see Figure 5).

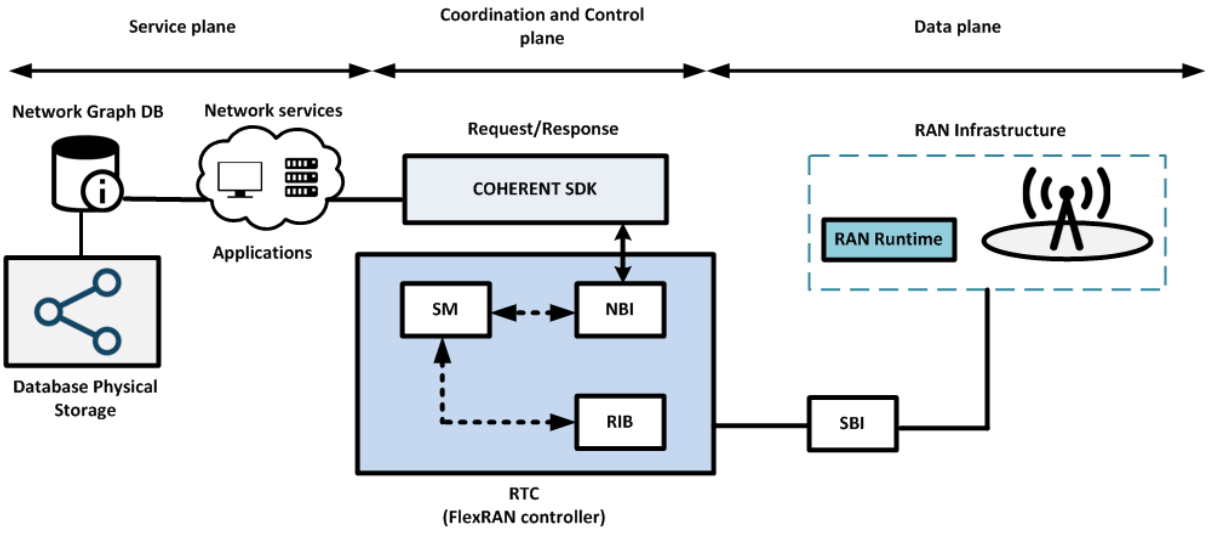


Figure 5: Network Graph Infrastructure in FlexRAN

Such network components and services are listed in the following:

- **Network Graph DB:** A graph database uses graph structures for semantic queries with nodes, edges and properties to represent and store data. Unlike with relational or non-relational databases, this type of structure is optimized to aid operations on graphs, e.g., graph traversals or "joint-like" navigation operations, frequently used in network control.
- **Network services:** A group of network applications that are developed using COHERENT Software Development Kit (SDK) and are running on the top of the RAN interacting with COHERENT RTC via REST commands, i.e., RESTful HTTP request/response examples. These examples can be formatted in JSON/XML/String format and can be further used for storage in the Network Graph DB. The requested information can be edge-user/network statistics, network control information etc.
- **Northbound Interface (NBI):** The NBI facilitates communication between the network services and the COHERENT RTC controller. REST API is the reference implementation for the case of FlexRAN RTC and corresponds to JSON formatted messages interchange.
- **Statistics Manager (SM):** The FlexRAN RTC statistics manager gets the network services requests via the corresponding NBI and collects the requested network information from the lower network layers. Such requests can be launched in one-shot, periodical, continuous, or event-driven. All the statistics and network control information are received over time and are stored in a radio information base (RIB) at the RTC, which can be then reused for further processing and analysis.

- **Radio Information Base (RIB):** The RIB is in charge of storing instantaneous network information collected by any RAN components. The SM is connected with the RIB to manage any information received after the network services request. A millisecond granularity level of network data information is provided by the RTC.
- **Southbound API (SBI):** The SBI facilitates the communication between the FlexRAN RTC controller and the RAN Runtime. Specifically, any network service request is translated to an SBI protocol (the so-called FlexRAN protocol) and is forwarded to the RAN Runtime that acts as a local RAN agent to collect any network information from any RAN components.
- **RAN Runtime:** The local RAN agent will invoke any network specific functions to collect the requested information related to edge users or BSs and send it back to the RTC.

As mentioned before, the requested information in the form of network graphs can be used by a network service that supports a group of network applications. These applications can range from pure network monitoring to network control and optimization and can be implemented using the provided abstractions, i.e., COHERENT SDK.

2.3 Proof-of-Concepts

2.3.1 RAN Monitoring and Proactive Control

This PoC description of RAN monitoring and proactive control done using Network Information Functions (NIF), has two sections that detail the PoC as implemented on an LTE testbed (OAI/FlexRAN) and a WiFi testbed (5G-Empower). The PoC and results presented in this deliverable have been done in testbeds that consist of either WiFi or LTE RATs. However, the approach of using NIFs for RAN monitoring and proactive control is applicable in a heterogeneous network that consists of both Wi-Fi and LTE RANs under a single controller. With additional models for attainable throughput, it can even be extended to any arbitrary RAT. In a heterogeneous network scenario the controller would manage the links over all available RATs (e.g. WiFi and LTE) and choose the link that best meets the requirements of the user. This is enabled by the network state representation provided by the NIF, which is RAT-agnostic. Evaluation of the NIF in a heterogeneous network environment using simulated networks is provided in [7] and [9].

2.3.1.1 Proactive control and management on OAI/FlexRAN LTE testbed using NIF

The NIF, as recapped in Section 2.2.3.1 is used in this PoC for proactive monitoring and control of the OAI-FlexRAN LTE. The proactive monitoring and control PoC scenario considered involves a testbed with 2 eNodeBs running Open Air Interface (OAI). The eNodeBs are monitored and controlled by the FlexRAN controller. This FlexRAN controller has a library of functions implemented as an SDK, on top of the controller for the purpose of easy development of control applications over it.

In the implementation of NIF on the OAI testbed the metrics that are monitored have been listed below table along with additional information. To obtain these metrics and achieve the goal of a central cross layer proactive decision, we design and implement a flat statistics manager application in FlexRAN to obtain the statistics of different layers including Physical layer (PHY), Medium Access Control (MAC), Packet Data Convergence Protocol (PDCP) and Radio Resource Control (RRC). The control modules of each layer are added to the FlexRAN control protocol and hook up to the COHERENT SDK which is then used by the NIF application.

Metric	Measured location	Measured for
Packet arrival rate (Pkts/sec)	PDCP layer at eNodeB	Each UE
Packet length (Bytes)	PDCP layer at eNodeB	Each UE
RSRP (dBm)	PHY layer at UE	Each candidate eNodeB-UE link
Bytes successfully delivered (Bytes)	PHY layer at eNodeB	Each serving eNodeB-UE link

Table 3: Monitored metrics of OAI testbed

Figure 6 shows the general components including FlexRAN SDK, **monitoring_app** and **NIF_app** that uses statistics functions from the SDK to monitor, aggregate and process information from the LTE network. The **monitoring_app** obtains monitoring metrics from the eNodeBs at the granularity of each transmission time interval (TTI). The *Packet arrival rate* and *packet lengths* are measured at the PDCP layer using counters for each serving user equipment (UE). *RSRP* is obtained using measurement reports through RRC reconfiguration message of FlexRAN reported by the UEs and obtained at the RRC layer at each eNodeB. *Successfully delivered bytes* are measured using hybrid automated repeated request (HARQ) control mechanism for transmissions and retransmissions in LTE. It counts the number and size of transport blocks at the PHY layer and corresponds them with the ACK or NACK response to obtain the number of successfully transmitted packets.

Downlink HARQ implementation in OAI is a synchronous HARQ process where the HARQ process ID changes from 0 to 7 cyclically such that the process ID can be obtained from the TTI number. The transmission attempt of a transport block with a particular HARQ ID is provided by the OAI scheduler as a **round number**. This round number updates and changes as the transmission attempt counter increases and is reset when the transport block (TB) is successfully delivered, i.e., a HARQ ACK corresponding to it is received. When this happens the transport block size (TBS) of the TB is recorded and added to the number of successfully delivered downlink bits. Therefore the observation of the HARQ round numbers and their correspondence to HARQ process IDs and TBS allows us to measure downlink throughput.

Figure 6 also depicts the components in the NIF_app and its relation to other components in the FlexRAN SDK. **NIF_app** aggregates these metrics in 100 ms measurement windows and adds it to a sliding window of 20 measurement window samples. An empirical CDF of throughput is obtained from the number of bytes successfully received, as observed at the HARQ process of the PHY layer in the eNodeB, in each measurement window. This is used to measure the probability of throughput satisfying the required threshold. When this probability drops below a specified tolerance, the evaluation of candidate links is triggered. Using the aggregated base metrics, attainable throughput is estimated over candidate links in similar measurement windows. If a candidate link is found that meets the QoS requirement of that UE then a handover trigger is created at the controller for the control application to use. Details of the experiment setup and their results to validate this PoC has been presented in Section 3.

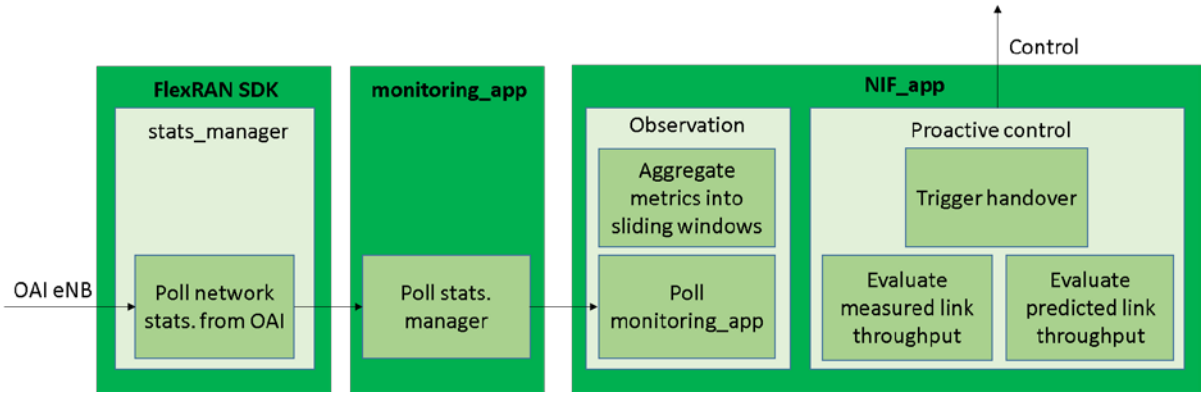


Figure 6: Diagram representing the components of NIF application and its interaction with other components in the FlexRAN SDK

2.3.1.2 Proactive control and management on 5G-EmPOWER WiFi testbed using NIF

The general description of monitoring and control using the NIF has been described in Section 2.2.3.1. In this PoC we perform proactive monitoring and control using NIF on the 5G-EmPOWER Wi-Fi testbed. The testbed consists of multiple thin access points referred to as Wireless Termination Points (WTP). These WTPs are monitored and controlled by the 5G-EmPOWER controller. This controller implements a rich SDK that allows users to easily build controller applications that run over the 5G-EmPOWER controller. We use a subset of the WTPs available to create a simple PoC that

effectively demonstrates the benefits of probabilistic threshold based proactive control using the NIF application.

For the testbed evaluation, we implement NIF on 2 WTPs with 3 clients in the network. Each WTP has 2 Wi-Fi interfaces (interface1 and interface2), set on non-overlapping channels (namely channel 6 and 1 in the 2.4 GHz range). One interface is used to host the LVAPs [3] of clients (virtual AP instance for each client) and the other interface monitors the frames on the other chosen channel. On neighboring WTPs, LVAPS are hosted on different non-overlapping channels, with the other interface being used to monitor the frames from the neighbors. WTP1 hosted LVAPS on interface1 on channel 6 and set interface2 on channel 1. WTP2 hosted LVAPS on channel 1 and set interface2 on channel 6.

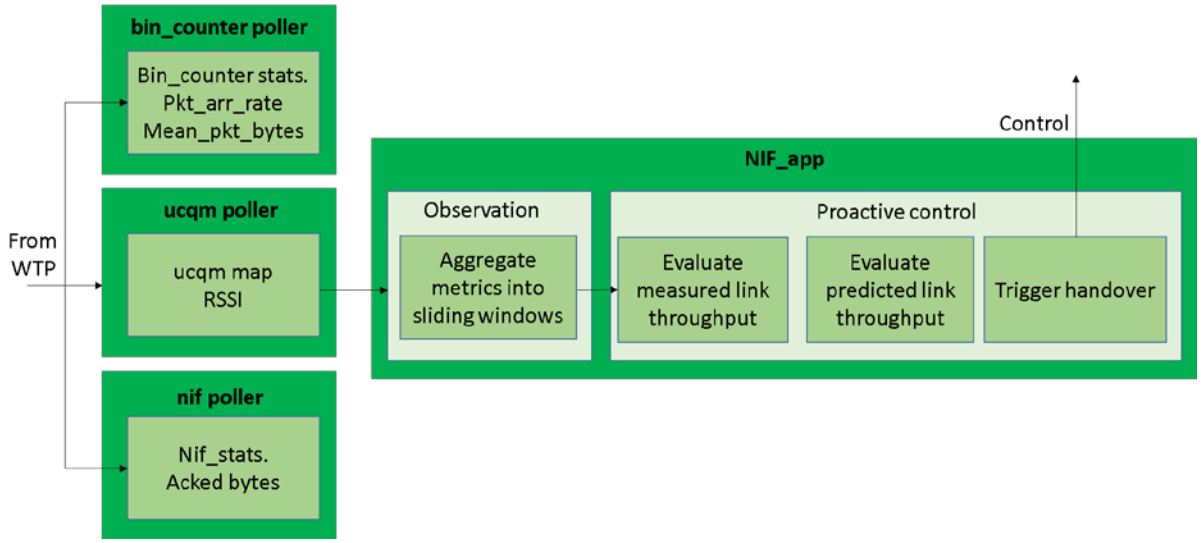


Figure 7: Diagram representing the components of NIF application and its interaction with other components in the 5G-EmPOWER SDK

Figure 7 depicts the components in proactive monitoring and control using NIF on the 5G-EmPOWER testbed for the scenario of mobility management. In the implementation of NIF on the 5G-EmPOWER testbed the metrics that are monitored have been listed below along with additional information.

Metric	**Measured location**	**Measured for**
Packet arrival rate (Pkts/sec)	IP layer at WTP	Each client
Packet length (Bytes)	IP layer at WTP	Each client
RSSI (dBm)	PHY layer at WTP	Each candidate WTP-client link
Packet delivery rate, PDR (fraction)	MAC layer at WTP	Each WTP
Bytes successfully delivered downlink (Bytes)	MAC layer at WTP	Each serving WTP-client link

Table 4: Monitored metrics of 5G-empower testbed

The packet arrival rate and packet length metrics are obtained using the **bin_counter** primitive from the SDK which maintains counters at each WTP for each client associated. These counters have bins for each packet length and increment them for each packet moving downlink. RSSI over links was obtained from sniffing frames on both available channels at the WTPs. The RSSI values are averaged over measurement windows of 500 ms and are obtained using the **ucqm (user channel quality map**) primitive from the SDK. PDR and successfully delivered bytes are obtained from the MAC layer by using counters at the Minstrel rate setting component. The Minstrel component provides information about how many packets were transmitted at which rates and what their measured success rate were. The number of successfully received bytes at the user is also measured in this component using the successfully transmitted bytes in correlation with the packet length. These metrics are obtained using the **nif_stats** primitive which was implemented by us to extend the SDK with important monitoring information required by the NIF.

These base metrics are aggregated in 500 ms measurement windows and added to a sliding window of 20 samples. An empirical CDF of throughput obtained from the number of bytes successfully received (as indicated by the **nif_stats** primitive) in each measurement window is used to observe the probability of throughput satisfying the required threshold of 14 Mbps. When this probability drops below tolerance level 0.7, the evaluation of candidate links is triggered. Using the aggregated base metrics, attainable throughput is estimated over candidate links in similar measurement windows. If a candidate link is found that meets the QoS requirement of that client then a handover trigger is created at the controller for the control application to use. Details of the experiment setup and their results to validate this PoC has been presented in Section 3.

2.3.2 RAN Inter-app Communications

This section describes the PoC on the communication between control applications. The app-to-app communication is used to implement network visualization and remote control of several control applications. The PoC shows programmability and extensibility in implementing COHERENT control applications.

COHERENT introduces the hierarchical control framework in Deliverable D3.2 [7], which separates the data and control plane of RAN, and thus allows control applications implemented on top of high layer control plane as control services. The purpose of introducing RAN control applications is to provide vendor-independent and software-defined control of radio network resources to satisfy requirements of network slicing and different services. COHERENT has implemented a set of control applications based on COHERENT SDK, for spectrum management, radio resource management (RRM), monitoring and network load balancing. The COHERENT SDK has been reported in Deliverable D2.3 [22] and D2.4 [20].

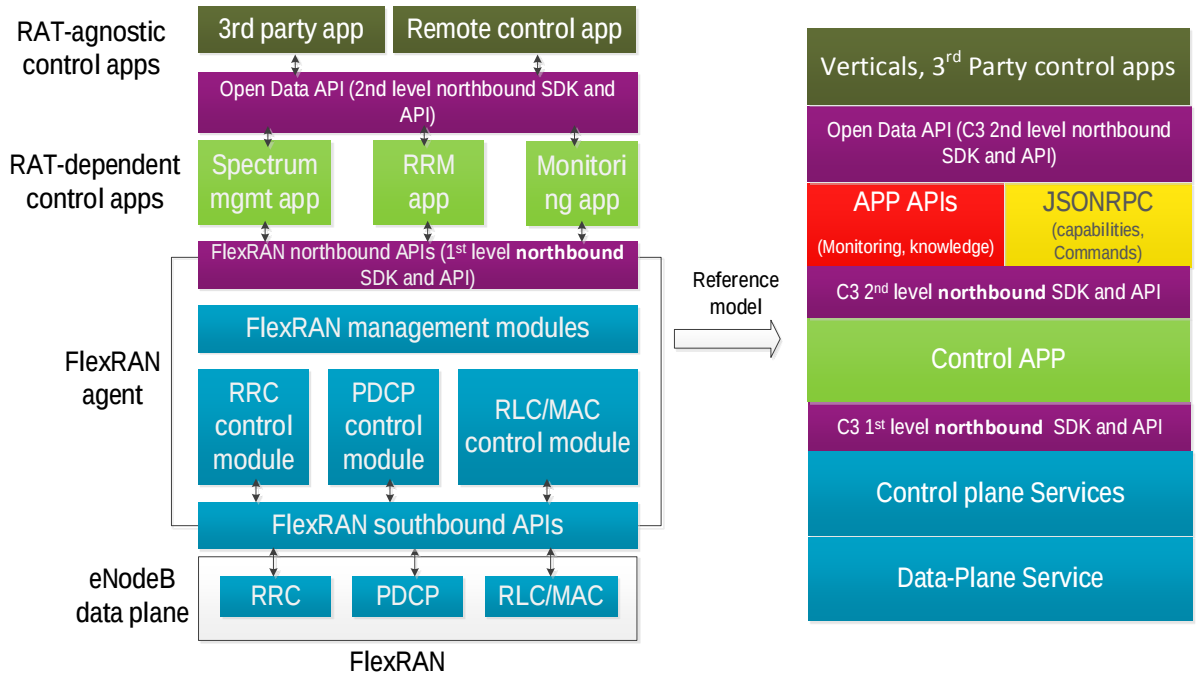


Figure 8: Control application reference framework in COHERENT

To illustrate the COHERENT control framework implementation, we use FlexRAN, which is one of COHERENT demonstration platforms, along with the other 5G-EmPower platform, as the baseline in this section. The PoC described in this section is built based on OAI and FlexRAN. FlexRAN is an open source software-defined RAN implementation following COHERENT architecture design principles. It separates the RAN data and control plane through derived southbound APIs. On the top of the control plane, it offers the northbound APIs to support RAN control applications, which allow flexible coordination of RAN infrastructure entities. The FlexRAN architecture and COHERENT control application reference framework are shown in Figure 8.

This reference framework can be divided into three layers. At the bottom (the blue parts) are data plane and control plane services, which form the functionalities of conventional 4G or new 5G base stations. In the middle layer (the first purple part from the bottom) is the C3 1st level northbound SDK and APIs, which are derived from control plane functions and open the RAN control to the upper layer programmable control applications. The upper layer (the green part and above) accommodates control applications. Note that a control application may further provide APIs to other control applications. Normally control applications are part of operators' network functions. However, through open data API, third parties, for instance, the verticals, can access the limited network status information to optimize their services or manage the network slices dedicated to them.

Each control application has its control purpose, and thus it relies on different granularity of network status information from the platform SDK and has different control realization. There is a need to provide the communication channel between them for direct information exchange. The reasons are three-fold: i) one control application may utilize the outputs of the other application. For instance, the monitoring application can monitor customized network parameters and provide inputs to RRM application; ii) control applications can synchronize their status if certain cooperation is needed between them. For instance the slicing in RRM application may need to synchronize spectrum usage policies with the spectrum management application; iii) the direct communication channel allows implementing common use interface for different specific control applications. This common use interface can be seen as an application on top of other applications to provide network visualization and remote control. In the project demo, spectrum management and RRM application use the common remote control interface.

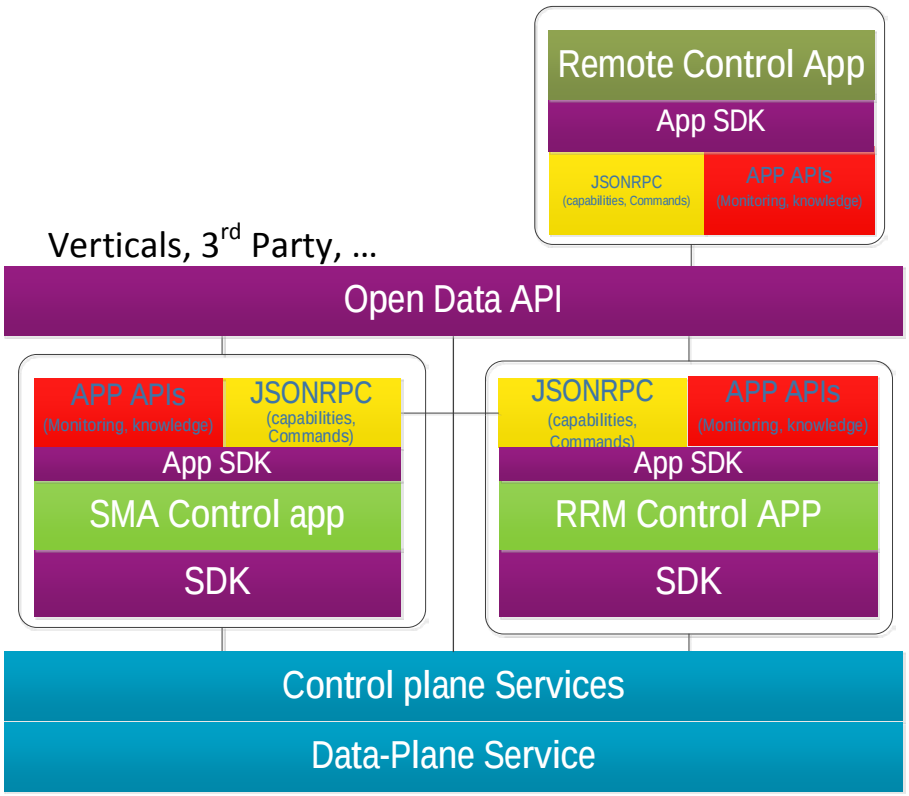


Figure 9: App-to-app communication in COHERENT control framework

The app-to-app communication is depicted in Figure 9, which shows different layers of the proposed approach to enable application service chain and control loop for a modular and coordinated decision-making. There are two types of applications as depicted in Figure 9. The first type is the application that communicates through RAN northbound APIs and thus tightly related to the RAN functions, e.g., the Spectrum Management Application (SMA) and RRM control app. The second type is the application that communicates through open data API, and thus the application on top of other applications, e.g., the remote control app. The app-to-app communication is implemented using JSON-RPC protocol [24]. The detail of the protocol is described as follows.

2.3.2.1 Open data protocol

Open Data API[4] is the interface through which different network components such as control apps and controller, can communicate with each other. The interface implements JSON-RPC v2.0 in support of monitoring, visualization and control among diverse network control applications. This brings second level northbound APIs in addition to the controller northbound, with their respective application and SDKs. With Open Data APIs, app-to-app communication protocol is specified based on JSON-RPC allowing requesting capabilities and status and send commands. The associated workflow between apps is shown in Figure 10.

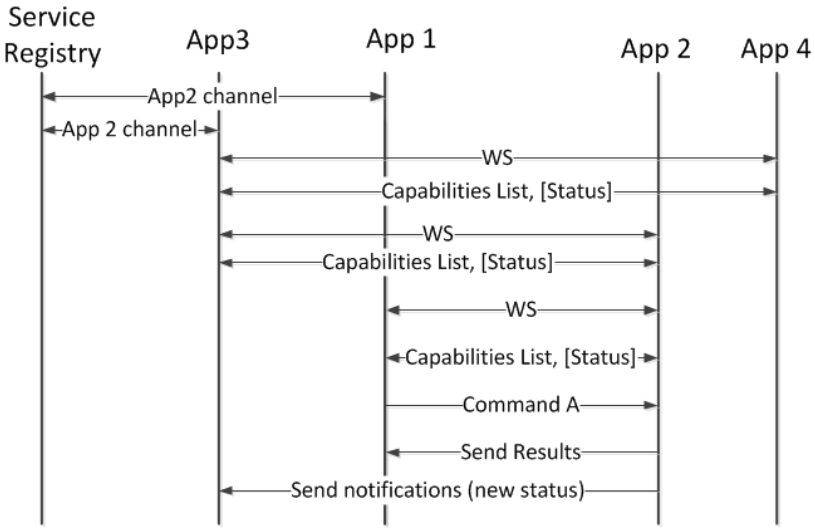


Figure 10: Workflow of app-to-app communications

An application will register its offered services to other applications in the service registry. So the service discovery is done through the service registry. An application will use the web socket (WS) information retrieved from the service registry to communicate with other applications. The first information exchanged between two applications is the Capabilities List. The Capabilities List is a JSON-RPC notification message giving the list of commands the application is willing to accept from the opened connection. Although the Capabilities List notification should be sent as first message on new connection, the list of accepted commands does not need to be fixed. An application can send updated Capabilities List whenever it is needed.

The general format of the Capabilities List notification:

```
{"method": "capabilities",
 "params": {
     "cap1": {<description of capability 1>},
     "cap2": {<description of capability 2>},
     ...
 }
}
```

The following general fields have been defined for *capability description* (all OPTIONAL):

- **help**: The value must be a text string representing a short description of the capability that could be used as a "tooltip" in graphical user interface.
- **group**: The value should be a simple string token. This is the information for the user interface about grouping of capabilities. If not specified, the capability belongs to an unnamed default group.

[4]*Visit* https://gitlab.eurecom.fr/mosaic5g/store/blob/develop/sdk/PROTOCOL.md *for the detailed information*

- **label**: The value is a string containing a HTML fragment. This can be used by the user interface in place of the method name. This allows the service to style the presented command in the user interface, e.g., replace text with some predefined icon image.
- **schema**: If present, it indicates the request accepts parameters. The value must be an array of *parameter definitions* for parameters in the 'params' of the JSON-RPC request object. In user interfaces, the input elements for the parameters are presented in the order they appear in the schema array. Each *parameter definition* is either the **name** of the parameter as a plain string, or dictionary describing the parameter in more detail:
  - **name**: The name of the parameter (REQUIRED)
  - **type**: Type of the parameter (OPTIONAL). The "number" value indicates that value should be serialized as integer in the JSON.
  - **help**: Description of the parameter (OPTIONAL). This can be used as a "tooltip" for the parameter.

  Moreover, optionally, **only one** of the following:

  - **choice**: The value must be an array of choices (strings). The corresponding message parameter value will be one of the listed choices. Currently, the following special keywords exist:
    - **#ENBID**: If present, it represents a list of known eNodeB identifiers to be substituted into the choice list.
    - **None**: (represented as null in JSON). If chosen, the resulting parameter value is an empty string, unless also the **type** "number" is present, in which case the choice omits the parameter totally.
  - **range**: Must be an array of at most 4 integer values: [<default>,<min>,<max>,<step>], any of which can be set to None (null in JSON). In GUI these are used for number input element. For example, "range": [20] requests an integer with default 20, but no other limitations.
  - **schema**: Defines a parameter as an object, the value must be an array of *parameter definitions*.

The command (method) names used in requests and notifications MUST be unique and consistent across applications, e.g., if two applications report support of “sensordata”, then the format of the "result" of reply (or "params" in notification) message MUST be same from both applications (though, actual data values can be different representing values from different “sensors”).

2.3.3 RAN Sharing

In this PoC, RAN sharing refers to a RAN slicing concept, which is defined per one RAT (i.e., intra-RAT radio resource management). In the real world, RAN sharing also applies to the concept of sharing a single RAT infrastructure among multiple operators, without slicing RAN resources as proposed in COHERENT, for instance, the Mobile Virtual Network Operator (MVNO) can be implemented only by the Core Network (CN) side.

As already presented in D2.3 and D2.4, the concept of slicing is an enabler for managing the end-to-end services and can be applied to both the RAN and CN parts. This more general definition presents the network slice as a collection of specific network functions/applications and RAT configurations, which are aggregated together for some specific use cases or business applications. Therefore, a network slice can span all domains of the network: software programs running on cloud nodes, specific configurations of the transport network, a dedicated radio access configuration, as well as settings of the underlying infrastructures. However, for some applications such as Public Safety (PS), it is imperative to have slicing defined at RAN level which is a main requirement expressed by public safety actors. The main advantage of RAN slicing compared to CN slicing techniques is a better control of the RAN and radio resource in general. The RAN slicing concept is sufficiently flexible to result, if needed, in a real physical separation of the resources, an isolation which allows better interference control and coexistence among different services. As we will see later in this deliverable, the advantage of RAN slicing results in a guarantee of service for public safety users.

Moreover, on top of the PS slice resources, we may run different network applications or configuration settings that are different from normal network setting. For example, in some cases, we may need define the slice as latency-critical or with high-reliability (intra-slice specific PS management), or in other cases we may want to define the slice with higher priority compared with the others defined slices (inter-slice specific PS management). The main idea behind the PS slice management is that the PS slice(s) should be defined with higher priority with respect to other slices, or if the same priority is defined, the slice(s) should be completely isolated from other slices providing different (legacy) services.

2.3.3.1 Network Slicing – Known Advantages

Network slicing represents a key 5G feature which may unlock a huge revenue potential and therefore the work on this topic is encouraged by network operators. Moreover, there are huge advantages from the end customer point of view (which will be able to have its own slice or a shared slice with other tenants for a given type of service with the same QoS), or for the vertical application point of view (which could acquire slices with guaranteed QoS). One network slicing advantage is the separation of application service flows to different slices. Therefore, a UE may run different service applications at the same time, thus increasing service differentiation and guaranteed QoS per slice. On the other hand, from the network point of view, multiple UEs sharing the same type of service or application could eventually connect to a single type of slice allowing simplified control (for example a single MAC scheduler entity per slice guaranteeing a fair QoS for every user). The idea of regrouping similar service types on the same slice may simplify the complexity of dealing with complex scheduling where URLLC has to be combined with eMBB or MTC-type traffic at the same time. A physical separation between time and frequency domain resources would therefore make sense.

2.3.3.2 Network Slicing – Known Issues

The first known issue is related to the slice dimension and the optimal number of slices. If the slice control is not adaptive enough, the slicing may result in a wastage of resources, for instance in cases where the allocated slice resources are higher than required by the service. Similarly, if the allocated slice resources are inadequate, the QoS may not reach the required level. If the network is flexible enough and the slices are allocated in a dynamic way, the number of slices and their capacities may decrease or increase in a real-time manner; therefore alleviating the fluctuation of traffic/services/users (users may come and leave in an unpredictable manner). However, given the dynamism and scalability involved in slicing, we believe that the management and orchestration of slicing in a multi-service multi-user multi-application flow scenario is not straightforward. Slices need resources to be assigned “on the fly” in very short time intervals, with i) just a few sporadic (and scattered) resource demands or ii) more frequent (and continuous) resource demands - depending on the number of users and service type, with service demands which can be low or high in terms of volume and requested access – and which can vary from low to high or from high to low, both within and across slices. Therefore, resource demands may not be constant or periodic and may vary dramatically.

Another issue is the slice isolation. In order to optimize the revenue, an operator may decide to re-use some of the resources for other services/functionalities. And since each slice can be customized for different services and businesses/vertical markets, we may have a problem of slice coexistence. As a matter of fact, slices are not always completely isolated from each other if they use the same resources. For example, if the same UE is connected to multiple slices or if a slice is designed to support machine-type communications (e.g., applications with low battery consumption), it may not necessarily optimally coexist with a best effort slice or a slice requiring enhanced broadband. A similar reasoning is applicable to the network side e.g., for the coexistence between a slice requiring low latency and another one required enhanced broadband. If the slices coexist on the same physical resources, none of the slices can be completely optimized for specific (different) services at the same time while using RAN in the most efficient way. These examples give sufficient good reasons for which Public Safety operators and/or Public Safety users may need completely isolated slices, especially for highly critical services such as alarms messages which they may configure separately with improved QoS characteristics, improved security, and/or different communication modes (such as D2D, broadcast or others) which may not be applicable to all slices.

Finally, the system must be adaptive and react in real-time in order to deal with 1) slice consistency for intra-slice operation and 2) slice coexistence for inter-slice operation, i.e., which is basically a verification if the slice definition and configuration are valid, and if there is no contradiction in terms of requirements. This kind of verification/validation can be done for all slices in a single (common) controller or application or can be done in a hierarchical manner, e.g., using two layers of applications: 1 lower-layer application per slice (which will verify/validate slices independently) and 1 higher-layer application to control all slices applications from the lower-layer. Moreover, the verification is performed by both end-processes in application and controller by using the SDK framework, i.e.,

- Before sending a slice configuration command towards the controller, the application verifies if the configuration required by the service provider/vertical market or operator is feasible;
- Before applying a configuration received from the application, the controller verifies if the slice configuration in terms of percentage, throughput, and other parameters is possible and applicable. If true, the controller applies the strategy and informs the slice application.

As detailed above, the controller will have the final word on the slice configuration (at least from the system point of view), because the controller is responsible of the real-time management of slice resources. However, an (initial) configuration command (or request) comes from the application as the Application represents a tool for the service provider (or the vertical market) to create and to map a given service on a slice.

2.3.3.3 Demonstrator Setup

The PoC has been configured for the use case of a single RAT (LTE in this case) with one eNodeB. This is sufficient as a proof of concept in the case of RAN slicing applicable to PMR radio. However, the demonstrator could be extended for usage with other RATs (e.g., Wi-Fi), in a scenario with multiple Access Points (APs) and multiple eNodeBs.

In Figure 11, we depict the envisaged demonstrator for RAN slicing. Note however that, in terms of hardware, the testbed could also be deployed in many other ways. The RRM slicing application sends commands (policy information) using a WebClient to the WebServer mounted in the FlexRAN RTC (FlexRAN Controller), through a Yaml interface (over HTTP protocol). In reply, FlexRAN RTC (FlexRAN Controller) sends values (BSR, MCS, etc.) collected from RAN to the RRM slicing application through JSON interface (again, over HTTP protocol). RTC FlexRAN is connected to the Agent on the eNodeB stack using ProtoBuf and allows controlling L2, and more precisely, the MAC scheduler on the eNodeB. This mechanism allows, to some extent, to allocate different schedulers to different slices which can use different scheduling mechanisms. In other words, the slice separation is done at MAC level by using multiple intra-slice MAC schedulers (per each slice) and a common inter-slice MAC scheduler, but on the same PHY substrate. Then, through a Wi-Fi-based access and an internal routing it is possible to connect the users directly to an external IP network or Internet.

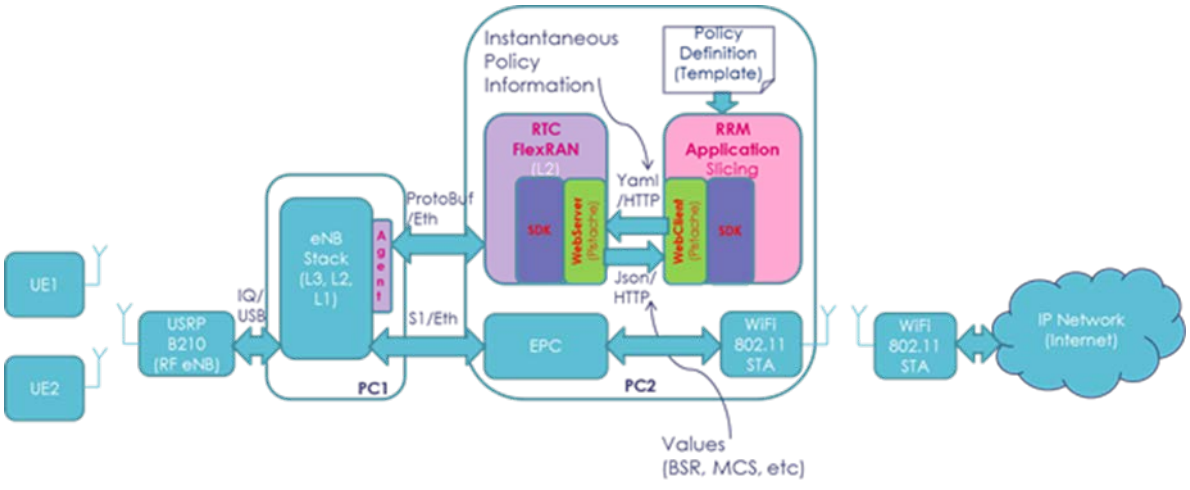


Figure 11: Envisaged Demonstrator for RAN Slicing

For demonstration purposes, we define different types of slices following the slice descriptor in Section 3.2.2 in Deliverable D2.4. For this demonstration the slice service policies are applied through the Yaml file descriptor. Several parameters are relevant to identify the slice: the slice name, the cell identity and the service policy (which can be different depending on the DL or UL direction of the flow). The service policy is characterized by multiple parameters: requested resource blocks, requested rate, requested latency, requested priority, requested resource isolation, which are further detailed in Section 5.1. The previous slice descriptor is processed by the RRM application, which also takes into account information coming from the controller and other RRM policy parameters. The RRM application produces an RRM policy which is applied though the controller. The RRM policy parameters can be in general different from the service policy parameters and are specific to the direction of the flow ('UL' or 'DL'). RRM policy parameters includes: slice id, slice service label, percentage of the slice resources, priority of the slice (the slice scheduling order), maximum MCS of the slice, and so on, as described further in Section 5.1. Moreover, required material and procedures to run the PoC are also described in the Annex.

2.3.3.4 Demonstration Scenarios

To begin with, please note the throughput values that can be achieved by an LTE system for a SISO case in UL or in DL for different system bandwidths in Table 5. The MCS and TBS indexes have been taken from Table 7.1.7.1-1 in 3GPP TS 36.213 V9.2.0 and the throughput is computed accordingly excluding signaling and control messages (in order to compute the maximum LTE system capacity).

Mod. Order	MCS/TBS Index	1.25 MHz (6 RBs)	2.5 MHz (12 RBs)	5.0 MHz (25 RBs)	10.0 MHz (50 RBs)	15.0 MHz (75 RBs)	20.0 MHz (100 RBs)
2 (QPSK)	0/0	0,152	0,328	0,68	1,384	2,088	2,792
	4/4	0,408	0,84	1,8	3,624	5,352	7,224
	9/9	0,936	1,864	4,008	7,992	11,832	15,84
4 (16QAM)	10/9	0,936	1,864	4,008	7,992	11,832	15,84
	13/12	1,352	2,728	5,736	11,448	16,992	22,92
	16/15	1,8	3,624	7,736	15,264	22,92	30,576
6 (64QAM)	17/15	1,8	3,624	7,736	15,264	22,92	30,576
	23/21	2,984	5,992	12,576	25,456	37,888	51,024
	28/26	4,392	8,76	18,336	36,696	55,056	75,376

Table 5: LTE achievable throughput [Mbits/sec.] (FDD, 1 RF chain, SISO)

Normally, under a severe SINR decrease, the user may have to use QPSK modulation instead of 16QAM or 64QAM which may decrease throughput by up to 30 times the normal maximum achievable data rate. On the other hand, good SINR allows better MCS. In the slicing definition, the notion of robustness is included by using different mapping between channel qualities and the corresponding applied MCS. For instance, a robust configuration would use a more conservative link-quality-to-MCS mapping, than a normal configuration used for standard best effort or broadband services. In Table 6**,** one may find examples of maximum expected slice throughput with a 10 MHz LTE configuration for different MCS/TBS indexes for different slice percentage (in terms of maximum allowed physical resource percentages). Similar result can be further obtained in a 5 MHz band in Table 7.

Mod. Order	MCS/TBS Index	10.0 MHz	Slice using 25% of resources	Slice using 35% of resources	Slice using 55% of resources	Slice using 65% of resources	Slice using 75% of resources
2 (QPSK)	0/0	1,384	**0.346 Mb/s**	**0.4844 Mb/s**	**0.7612 Mb/s**	**0.8996 Mb/s**	**1.038 Mb/s**
	4/4	3,624	**0.906 Mb/s**	**1.2684 Mb/s**	**1.9932 Mb/s**	**2.3556 Mb/s**	**2.718 Mb/s**
	9/9	7,992	**1.998 Mb/s**	**2.7972 Mb/s**	**4.3956 Mb/s**	**5.1948 Mb/s**	**5.994 Mb/s**
4 (16QAM)	10/9	7,992	**1.998 Mb/s**	**2.7972 Mb/s**	**4.3956 Mb/s**	**5.1948 Mb/s**	**5.994 Mb/s**
	13/12	11,448	**2.862 Mb/s**	**4.0068 Mb/s**	**6.2964 Mb/s**	**7.4412 Mb/s**	**8.586 Mb/s**
	16/15	15,264	**3.816 Mb/s**	**5.3424 Mb/s**	**8.3952 Mb/s**	**9.9216 Mb/s**	**11.448 Mb/s**

Table 6: Expected LTE achievable throughputs in terms of slicing in 10 MHz

Mod. Order	MCS/TBS Index	5.0 MHz	Slice using 25% of resources	Slice using 35% of resources	Slice using 55% of resources	Slice using 65% of resources	Slice using 75% of resources
2 (QPSK)	0/0	0,68	**0.17 Mb/s**	**0.238 Mb/s**	**0.374 Mb/s**	**0.442 Mb/s**	**0.51 Mb/s**
	4/4	1,8	**0.45 Mb/s**	**0.63 Mb/s**	**0.99 Mb/s**	**1.17 Mb/s**	**1.35 Mb/s**
	9/9	4,008	**1.002 Mb/s**	**1.4028 Mb/s**	**2.2044 Mb/s**	**2.6052 Mb/s**	**3.006 Mb/s**
4 (16QAM)	10/9	4,008	**1.002 Mb/s**	**1.4028 Mb/s**	**2.2044 Mb/s**	**2.6052 Mb/s**	**3.006 Mb/s**
	13/12	5,736	**1.434 Mb/s**	**2.0076 Mb/s**	**3.1548 Mb/s**	**3.7284 Mb/s**	**4.302 Mb/s**
	16/15	7,736	**1.934 Mb/s**	**2.7076 Mb/s**	**4.2548 Mb/s**	**5.0284 Mb/s**	**5.802 Mb/s**

Table 7: Expected LTE achievable throughputs in terms of slicing in 5 MHz

Further, for the demonstration purposes we have imagined a few scenarios which could be tested with their respective behaviors:

- **Scenario 1**: PMR user requires more throughput in a PMR eMBB slice, which will directly increase the PMR slice percentage of total resources and decrease the normal user slice (eMBB, best effort or URLLC) percentage of resources;
- **Scenario 2**: Normal user requires more throughput (in eMBB, best effort or URLLC), which will directly increase the normal slice percentage of resources till it reaches maximum accepted value (without PMR slice percentage) and without affecting PMR slice percentage of resources or PMR slice performance, depending on the type of implemented isolation;
- **Scenario 3**: PMR user connected in eMBB mode requires URLLC at the same throughput, which will indirectly increase the PMR slice percentage, decrease the normal user slice (eMBB, best effort or URLLC);
- **Scenario 4**: Normal user moves from eMBB or Best Effort to URLLC and requires more throughput, which will indirectly increase the normal slice till it reaches maximum accepted value (without PMR slice percentage) and without affecting PMR slice percentage;

All these scenarios could be separately tested for Uplink or Downlink (please see Yaml description for slicing purposes). Moreover, in Figure 12, one can find a graphical representation (with some examples) of the system behavior before and after the action (i.e., the configuration command), for

each of the above scenarios. In Section 5 only scenarios 1 and 3 have been tested and validated even if the scenario 2 and 4 are supported by the testbed. Afterwards, the description and analysis of scenarios 2 and 4 can be found too in order to illustrate the expected results in those situations, which are actually stressing the isolation property of the PMR slices.

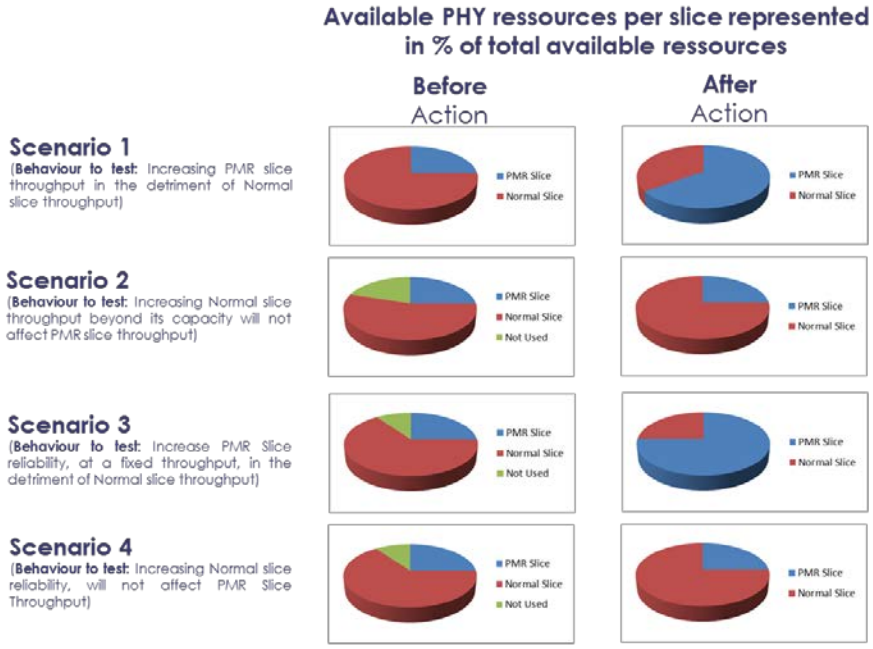


Figure 12: Graphical Examples of Envisaged Scenarios

The purpose of scenario 1 is to show that in case of emergency, the PMR slice can take resources from the Normal slice. Similarly, the purpose of scenario 2 is to show that the Normal slice can use non-allocated resources (not already used) but cannot use or take resources from already defined/existent PMR slice. In other words, PMR slice has priority against the Normal slice. The behavior of the PMR slice should not be affected except for when there is an action from PMR authority indicating that the PMR slice behavior can be changed.

The purpose of scenario 3 is to show that in case of emergency, PMR slice can be configured with higher robustness for the same expected throughput. Finally, the purpose of scenario 4 is to show that increasing reliability or robustness in the Normal slice will not affect PMR slice QoS. Normal slice can take unused resources but cannot use resources from the PMR slice. Normal slice can also decrease total expected throughput or use a configuration in terms of robustness that allows having same previous throughput (before the action) but without interfering with PMR slice. As indicated in Table 8, since there may not be any optimal configuration for this last strategy, this action may result in un-used resources, which is not optimal but is respecting the given constraints. The main problem here is that LTE TBS configuration does not allow sufficient flexibility: the granularity of the resource scheduling is quite high and therefore for a target expected UE throughput some configurations may not be possible, resulting in a situation where resources are not actually entirely used (in order to respect the user requirement, the system may require to choose a TBS size that allows to provide sufficient available throughput, but which is above what is really required by the user).

Taking the previous described scenarios into account (some more others can be defined), in Table 8 we give a few examples of actions, initial system state (before the action), expected system behavior (after action) and a general observation with respect to each of the scenarios. Please note that in scenarios 2 and 3, “capacity” means the maximum throughput that a slice can support using a certain MCS set. Similar results can be obtained in Table 9, for a 5 MHz band.

Scenario	Action	Initial System State (before action)	Expected System Behavior (after action)	Observation
1	Increase PMR slice throughput in the detriment of Normal slice throughput	UE 1 (PMR UE), **3.816 Mb/s in 16QAM 16/15** UE 2 (Normal UE), **11.448 Mb/s in 16QAM 16/15**	UE 1 (PMR UE), **9.9216 Mb/s in 16QAM 16/15** UE 2 (Normal UE), **5.3424 Mb/s in 16QAM 16/15**	Increasing PMR slice throughput may decrease Normal slice throughput
2	Increase Normal slice throughput beyond the Normal slice capacity	UE 1 (PMR UE), **3.816 Mb/s in 16QAM 16/15** UE 2 (Normal UE), **8.3952 Mb/s in 16QAM 16/15**	UE 1 (PMR UE), **3.816 Mb/s in 16QAM 16/15** UE 2 (Normal UE), **11.448 Mb/s in 16QAM 16/15**	Increasing Normal slice throughput not affecting PMR slice throughput
2	Increase Normal slice throughput beyond the Normal slice capacity	UE 1 (PMR UE), **3.816 Mb/s in 16QAM 16/15** UE 2 (Normal UE), **0.7612 Mb/s in QPSK 0/0**	UE 1 (PMR UE), **3.816 Mb/s in 16QAM 16/15** UE 2 (Normal UE), **11.448 Mb/s in 16QAM 16/15**	Increasing Normal slice throughput not affecting PMR slice throughput
2	Increase Normal slice throughput beyond the Normal slice capacity	UE 1 (PMR UE), **3.816 Mb/s 16QAM 16/15** UE 2 (Normal UE), **0.7612 Mb/s in QPSK 0/0**	UE 1 (PMR UE), **3.816 Mb/s in 16QAM 16/15** UE 2 (Normal UE), **1.038 Mb/s in QPSK 0/0**	Increasing Normal slice throughput not affecting PMR slice throughput
3	Increase PMR slice reliability, at a fixed throughput, in the detriment of Normal slice throughput	UE 1 (PMR UE), **2.862 Mb/s in 16QAM 13/12** UE 2 (Normal UE), **9.9216 Mb/s in 16QAM 16/15**	UE 1 (PMR UE), **2.862 Mb/s in QPSK 4/4** UE 2 (Normal UE), **3.2094 Mb/s in 16QAM 16/15**	Increasing PMR slice reliability may decrease normal slice throughput, at an expected PMR QoS defined by PMR service parameters
4	Increase Normal slice reliability (with throughput decrease for Normal UE)	UE 1 (PMR UE), **3.816 Mb/s in 16QAM 16/15** UE 2 (Normal UE), **7.4412 Mb/s in 16QAM 13/12**	UE 1 (PMR UE), **3.816 Mb/s in 16QAM 16/15** UE 2 (Normal UE), **5.994 Mb/s in QPSK 9/9**	Increasing normal slice robustness not affecting PMR slice throughput or robustness
4	Increase Normal slice reliability (without throughput decrease for Normal UE)	UE 1 (PMR UE), **3.816 Mb/s in 16QAM 16/15** UE 2 (Normal UE), **3.816 Mb/s in 16QAM 16/15**	UE 1 (PMR UE), **3.816 Mb/s in 16QAM 16/15** UE 2 (Normal UE), **3.816 Mb/s in QPSK 9/9**	Increasing normal slice robustness not affecting PMR slice throughput or robustness

Table 8: Expected System Behavior for Slicing Use Case (indicative values for different TBS/MCS in 10 MHz)

As we can see from the previous table, and highlighted in D2.4, the concept of priority is essential when used for enforcing slices supporting PS applications (i.e., PMR slice(s)), e.g., in the situation of a crisis/emergency. The resource and performance isolations are guaranteed for slices with equal priority (e.g., two Normal slices). However, in case of different priorities i.e., a PMR slice and a Normal slice, isolation may be broken in the sense that slices with lower priority i.e., Normal slice(s) can see their guaranteed resources be allocated (totally or in part, according to a special repartition agreement) to slices with higher priority i.e., PMR slice(s).

2.3.3.1 Conclusions

Many descriptions of network slicing assume that the scope of a slice must go across the entire network from a user application to the RAN and then CN to a server in the heart of a fixed network or to the other user on the other end. However, as a matter of fact, some services may stay within the RAN while others may go directly between data centers in the CN. In the case of Public Safety scenarios and Professional Mobile Radio systems it is imperative to completely isolate the PMR/PS slice from the normal legacy use at the RAN level and therefore, network slicing application can be implemented at the RAN level. Firstly, the demonstrator showed that it is possible to increase the level of priority of the PMR slice compared with the normal usage without any noticeable impact on the PMR QoS when other slices are used. Secondly, the demonstrator showed that the application can dynamically configure normal slice(s) when PMR slice application demands more throughput (or less throughput), or when the PMR slice application demands for an increase of the robustness of the communication (or a decrease, because e.g., all PMR users are in proximity).

Scenario	Action	Initial System State (before action)	Expected System Behavior (after action)	Observation
1	Increase PMR slice throughput in the detriment of Normal slice throughput	UE 1 (PMR UE), **1.934 Mb/s in 16QAM 16/15** UE 2 (Normal UE), **5.802 Mb/s in 16QAM 16/15**	UE 1 (PMR UE), **5.0284 Mb/s in 16QAM 16/15** UE 2 (Normal UE), **2.7076 Mb/s in 16QAM 16/15**	Increasing PMR slice throughput may decrease Normal slice throughput
2	Increase Normal slice throughput beyond the Normal slice capacity	UE 1 (PMR UE), **1.934 Mb/s in 16QAM 16/15** UE 2 (Normal UE), **4.2548 Mb/s in 16QAM 16/15**	UE 1 (PMR UE), **1.934 Mb/s in 16QAM 16/15** UE 2 (Normal UE), **5.802 Mb/s in 16QAM 16/15**	Increasing Normal slice throughput not affecting PMR slice throughput
2	Increase Normal slice throughput beyond the Normal slice capacity	UE 1 (PMR UE), **1.934 Mb/s in 16QAM 16/15** UE 2 (Normal UE), **0.374 Mb/s in QPSK 0/0**	UE 1 (PMR UE), **1.934 Mb/s in 16QAM 16/15** UE 2 (Normal UE), **5.802 Mb/s in 16QAM 16/15**	Increasing Normal slice throughput not affecting PMR slice throughput
2	Increase Normal slice throughput beyond the Normal slice capacity	UE 1 (PMR UE), **1.934 Mb/s 16QAM 16/15** UE 2 (Normal UE), **0.374 Mb/s in QPSK 0/0**	UE 1 (PMR UE), **1.934 Mb/s in 16QAM 16/15** UE 2 (Normal UE), **0.51 Mb/s in QPSK 0/0**	Increasing Normal slice throughput not affecting PMR slice throughput
3	Increase PMR slice reliability, at a fixed throughput, in the detriment of Normal slice throughput	UE 1 (PMR UE), **1.434 Mb/s in 16QAM 13/12** UE 2 (Normal UE), **5.0284 Mb/s in 16QAM 16/15**	UE 1 (PMR UE), **1.434 Mb/s in QPSK 4/4** UE 2 (Normal UE), **1.6266 Mb/s in 16QAM 16/15**	Increasing PMR slice reliability may decrease normal slice throughput, at an expected PMR QoS defined by PMR service parameters
4	Increase Normal slice reliability (with throughput decrease for Normal UE)	UE 1 (PMR UE), **1.934 Mb/s in 16QAM 16/15** UE 2 (Normal UE), **3.7284 Mb/s in 16QAM 13/12**	UE 1 (PMR UE), **1.934 Mb/s in 16QAM 16/15** UE 2 (Normal UE), **3.006 Mb/s in QPSK 9/9**	Increasing normal slice robustness not affecting PMR slice throughput or robustness
4	Increase Normal slice reliability (without throughput decrease for Normal UE)	UE 1 (PMR UE), **1.934 Mb/s in 16QAM 16/15** UE 2 (Normal UE), **1.934 Mb/s in 16QAM 16/15**	UE 1 (PMR UE), **1.934 Mb/s in 16QAM 16/15** UE 2 (Normal UE), **1.934 Mb/s in QPSK 9/9**	Increasing normal slice robustness not affecting PMR slice throughput or robustness

Table 9: Expected System Behavior for Slicing Use Case (indicative values for different TBS/MCS in 5 MHz)

2.3.4 Load Balancing

2.3.4.1 Inter-RAT LTE/WiFi Load Balancing Testbed

For Inter-RAT LTE/WiFi Load Balancing Testbed, we refer to the testbed in D5.2 [21] consisting of one srsLTE eNodeB, one Wi–Fi access point and two mobile clients. This PoC is a SDN platform supporting mobility management applications over heterogeneous Wi–Fi/LTE RANs and aims at optimizing resource utilization in a heterogeneous Wi-Fi/LTE RAN. To reach its goal, the traffic-aware user association solution was implemented in the COHERENT C3 (5G–EmPOWER).

2.3.4.2 Intra-RAT LTE Load Balancing Testbed

The LTE radio testbed consists of two CommAgility eNodeBs, EPC, and two commercial UEs. As UEs, commercial LTE Huawei dongles or smartphone Samsung Galaxy S4 are used. In order to use fully LTE-compliant base station and UE, an EPC emulator is used to emulate the EPC's core network entities MME, S-GW and terminate protocols like S1AP, NAS.

The network setup is composed of two LTE eNodeBs and the C3. The C3 is implemented by capitalizing on the 5G-EmPOWER coordinator and agents running on the LTE nodes. Below is an

example of the topology considered for load balancing in LTE, composed of serving nodes (eNodeBs) and UEs (smartphones) and the COHERENT C3 (Figure 13).

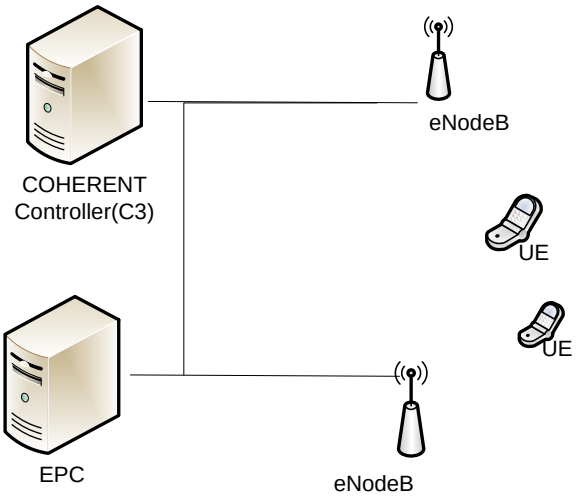


Figure 13: Topological overview of the load balancing use case infrastructure.

The CommAgility eNodeB is based on the high-performance ARM and DSP based processing card AMC-K2L-RF2 that represents a complete baseband and RF small cell solution. The AMC-K2L-RF2 is designed to support wireless baseband processing and a 2x2 MIMO air interface in eNodeBs, for standard or specialized LTE and LTE-Advanced systems up to and beyond Release 10.

Figure 14: AMC-K2L-RF2 card

The main processor is the TCI6630K2L, part of Texas Instruments' new KeyStone II generation of DSP/ARM SoCs. It includes four C66x DSP cores and two ARM® Cortex®-A15 MPCore™ processors, all operating at up to 1.2 GHz. The TCI6630K2L also contains wireless base station coprocessors which offload layer 1 and layer 2 processing demands to keep the cores free for receiver algorithms and other functions. A range of I/O is provided to the TCI6630K2L, including Gigabit Ethernet, PCIe and CPRI connectivity, giving connectivity to networks, host processors and additional RF. Using external cavity filters, the card can be directly deployed as an LTE small cell.

AMC-K2L-RF2 provides a comprehensive DSP Board Support Library along with a full SMP Linux implementation for the ARM cores. The physical layer software is highly integrated with the Texas Instruments TCI6630K2L SoC for maximum performance. Above the physical layer software, LTE protocol stack compliant with Small Cell Forum MAC-PHY API is fully integrated. For the testbed the eNodeB software architecture depicted in Figure 15 is used.

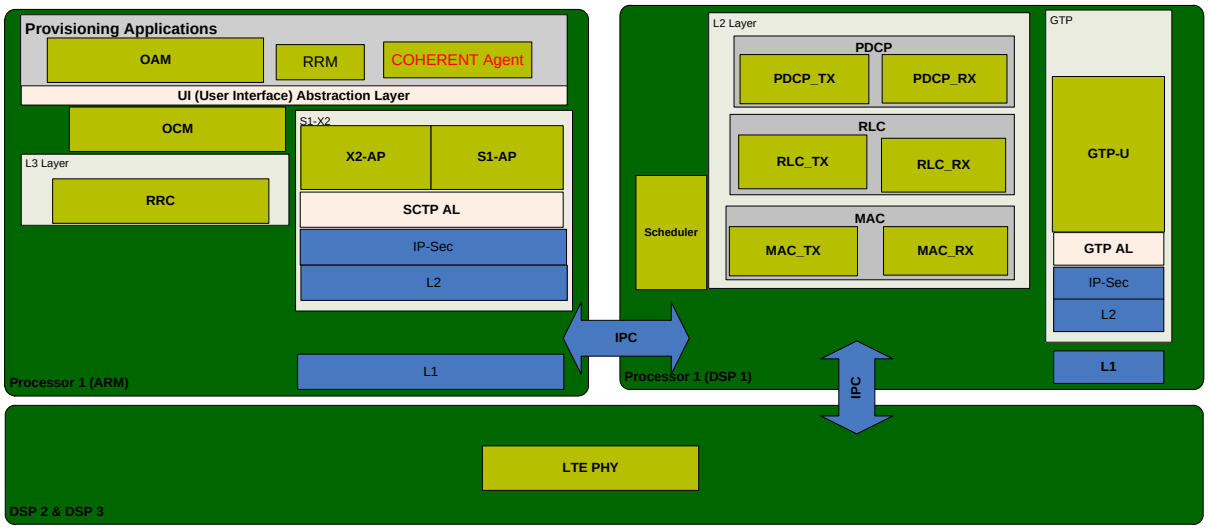


Figure 15: CommAgility eNodeB software architecture

The protocol stack and the physical layers run on different processors. LTE control plane is on ARM ARM/Linux processor and LTE data plane in on DSP core 1 (SYSBIOS). Running LTE control plane on ARM allowed to free DSP core 1 from control mechanisms of LTE stack (like measurement, handover procedure and RRM algorithms). Other DSP cores run physical layer software.

Inter-process communication (IPC) between processors is achieved through use of TI Msgcom API in this case. The communication framework facilitates exchange by the inter component messages (ICMs) between the eNodeB stack components. Each ICM is a structured C type having a fixed header as the first field. COHERENT control services on the eNodeB will be implemented to run on the ARM processor.

In the project two different IPC mechanisms are used. If source component X and destination component Y run on the same processor but on different threads, message posting (see Figure 16) to destination component (non-blocking for source component) is used.

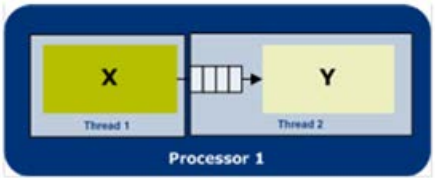


Figure 16: Message posting

If source and destination components run on different processors (implies different threads), then the inter-processor message passing mechanism is used (see Figure 17). This mechanism would be dependent on the inter-processor communication method (DMA, Message queue with pull/push mechanism, etc.). Then the communication framework at the destination processor posts a message to destination component (non-blocking for source component). This communication is done using the board support package (BSP) for AMC-K2L-RF2.

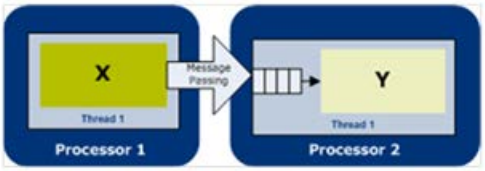


Figure 17: Inter-processor message passing mechanism

The integration with COHERENT C3 has been done by implementing the COHERENT Agent that runs on ARM processor at the CommAgility eNodeB and interacts with an EMPOWER controller using a set of protocols. The COHERENT Agent responds to the action issued by a controller.

The PoC, which uses the previously described testbed, aims to demonstrate the traffic offloading decision designed to operate in technology agnostic manner. Particularly, it is used to demonstrate the virtualization of handover (mobility) function by moving it from eNodeB to the cloud network. The LTE standard was customized to provide the possibility to trigger handover not only from eNodeBs (as specified in the current LTE specification), but also from the coordinator through a new interface between the eNodeB and COHERENT C3.

2.3.5 Spectrum management

The spectrum management PoC includes two experiments. The first experiment is built on the OAI/FlexRAN platform to show the dynamic spectrum allocation among phantom cells. The second experience is a spectrum sharing trial using Licensed Shared Access on 2.6 GHz band.

2.3.5.1 Spectrum management on OAI/FlexRAN platform

In the context of 5G networks, where various abstraction and virtualization techniques will be applied, effective spectrum management is still highly challenging and plays, in many cases, very critical role. It is widely known that limited amount of spectrum resources must be utilized in a proper way among different verticals and stakeholders in order to fulfil their expectations and needs. Although the static licensing for exclusive spectrum use among operators on one hand and the license-exempt solutions on the other will still be the most popular solutions, the technological trends indicate precisely the need for more advanced and flexible approaches for spectrum management. This observation is of particular importance for networks where the NFV concept have been implemented, meaning that the control functions are (may be) fully separated from the underlying hardware. From this perspective, access to the spectrum resources (assigned to various operators and accessible in a flexible way) may be treated as the new function (service) in the virtualized world. COHERENT proposed a fresh network architecture with logically centralized controller and coordinator, where all aspects related to the spectrum management are subject to processing in dedicated application called SMA. Its role is to manage and process in an intelligent way various requests, messages, guidelines, rules and limitations originated and defined by various stakeholders (such as National Regulatory Authority – NRA, operators, spectrum license owners, etc.). For example, the NRA could define some specific rules for spectrum usage for a certain geographical area to facilitate decent service delivery to the end users during a scheduled mass event. Another possibility is that one license owner (operator) may define its own rules for spectrum sharing with other interested players, or two collocated operators may define their own mutual agreements for spectrum sharing. In all cases, SMA will read and interpret such rules and derive recommends on the possible ways of spectrum use; these recommends are then sent, either automatically or by request, to the logically centralized coordinator and controller. More details about the implementation of SMA in various use cases and contexts can be found in [18].

Although theoretically interesting and tractable, huge effort has been put to identify and implement a suitable scenario which will illustrate simultaneously the flexibility of spectrum allocations and measurable benefits resulting from it. As the final testbed considers the OAI platform, the key problem in this context was to add efficiently the functionality of flexible frequency change to the platform. The stringent constraints in this case originated from the fact that OAI implements the true LTE-A protocol stack and in even the simple change of assigned channel bandwidth for a certain base-station results in almost full software reconfiguration of this base-station in order to adapt it to the new needs.

In the final COHERENT testbed, a dedicated SMA has been designed, developed, and integrated with the OAI/FlexRAN platform with the purpose of dynamically switching on or off the so-called *phantom cell use-case*. Following [19], one may understand the phantom cell concept as the dedicated small cell which is not configured with all cell-specific signals or channel (such as primary and secondary synchronization signals, or master/system information blocks) and is intended to carry only user data traffic. All radio resource control procedures between the UE and the network are managed in this concept by the collocated macro cell. In our case, we consider the presence of the flexible phantom-cell, which will be activated in specific situations (e.g., more capacity has to be offered to the end users in the network).

The SMA considers various entries stored in dedicated databases, mainly includes:

a. general rules defined by the NRA for a given region and time,
b. mutual agreements between any two (or more) operators regarding the spectrum sharing,
c. open (for public) spectrum sharing rules, such as those following the Licensed Shared Access (LSA) or Citizen Broadband Radio Service (CBRS) approach (These spectrum sharing rules define the ways how any allowed stakeholder may bid or request from the spectrum licensee a certain amount of spectrum resources.), and
d. priorities which have to be applied while allocating the spectrum among the interested base-stations (for example, one of the priorities may be to maximize the operator incomes due to the spectrum sharing).

This scheme can have various practical applications. First of all, one may observe that by adaptive identification of prospective spectrum opportunity the overall network sum-rate may be increased, i.e., a new phantom cell will be switched on and the identified spectrum resources will be associated with it delivering new capacity to the system. Moreover, by deployment of the phantom cell, some of the traffic may be "offloaded" (or better moved) from the contested cells to the new phantom cell leading to overall improvement of network performance (reduced congestion rate). When the phantom cell is no longer needed, it may be switched off (reducing overall energy consumption) and releasing the spectrum resources.

In our experiment, we have concentrated on the first aspect where the SMA is applied jointly with the OAI/FlexRAN platform in order to manage the usage of spectrum in a dynamic way for increasing the overall network throughput. Once the SMA identifies new spectrum opportunities, it decides regarding the new frequency (channel) allocation for controlled base stations and informs the FlexRAN application about this decision. In the next step, the affected base stations are reconfigured and new phantom-cell may be switched on.

2.3.5.2 Spectrum sharing using Licensed Shared Access on 2.6 GHz band

So far, almost all radio licenses for mobile broadband have been allocated with nationwide licenses. Mobile network operators have shared a specific frequency band as block licenses. The high number of IMT frequency bands and higher frequencies of the bands have initiated the ideas of sharing mobile operator bands also geographically. The ideas include sub-leasing, neutral host operators, overlapping macro and small-cell networks, and local 5G licenses for industry verticals. In these proposals, two or more mobile operators potentially have adjacent geographical operating areas.

LSA was developed by ETSI to allow licensed and protected secondary use of spectrum for mobile network operators. The ETSI RRS technical report from February 2018 extends the concept by applying LSA for local and temporary licenses. In this demonstration, we show how LSA could be applied to spectrum sharing between a national mobile operator and vertical sector having a local spectrum license. The chosen approach in this demonstration is based on different priorities, which could for example be applied, when the mobile operator has a nation-wide license and the operator allows secondary use of the band by vertical sectors. The vertical sectors are private LTE operators in this demonstration.

System description

A sharing scenario is implemented where one eNodeB acts as a primary mobile operator (U1) and has the highest priority while the other eNodeB is a Private LTE operator (U2) and has a lower priority. The 2.6 GHz FDD band 7 is considered with downlink frequencies 2620-2690 MHz and 10 MHz channel bandwidth. In the testbed at OTE premises, there are two 2.6 GHz FDD eNodeBs. Fairspectrum LSA Controllers can access the eNodeBs through remote access from virtual machine in OTE premises. Although there are only two operators in the demonstration, the scheme can be generalized to account for multiple core networks and eNodeBs due to the distributed eNodeB Controllers.

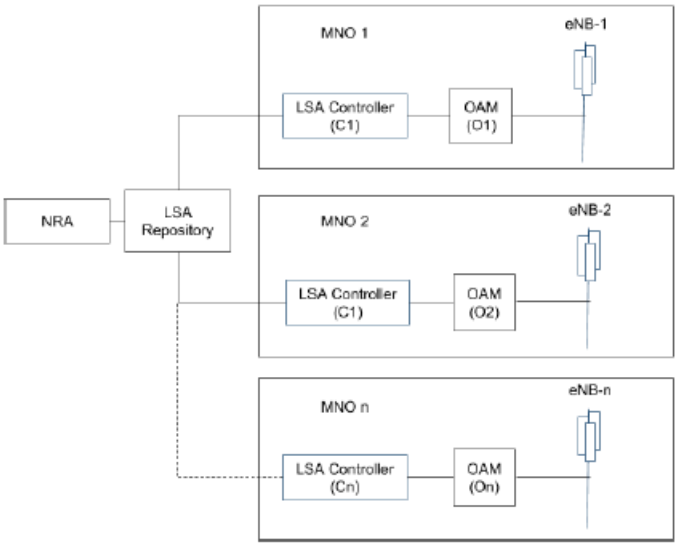


Figure 18: Prototype setup of the spectrum sharing using LSA

2.3.6 Distributed antenna system

The current work mainly focuses on the details of implementation of LTE PHY on a Texas Instruments (TI) DSP and the corresponding processing blocks that needs to be adapted in LTE PHY to support the functionality of DAS. Effort has been made to implement a commercial LTE system on DSP and ARM boards. The LTE PHY layer is ported on a commercial TI DSP which gives flexibility to modify or enhance the functionalities towards DAS. The description of our work is organized in the sections below: starting with implementation details for DAS functionality on LTE testbed, description of a set of adaptations of LTE procedures for DAS deployment, Description of coordinated pairing in DAS, Network graphs generation and concludes with testbed description.

2.3.6.1 Implementation details

Figure 19 shows the block diagram view of architectural components involved in Distributed Antenna System use case. The PHY layer processing of DAS LTE based eNodeB is implemented on the DAS Testbed and network graphs (NG) formulation is simulated in MATLAB. The algorithm for coordinated pairing, which is implementable on RTC/C3 has been only simulated in MATLAB because of non-feasibility of MAC and higher layers.

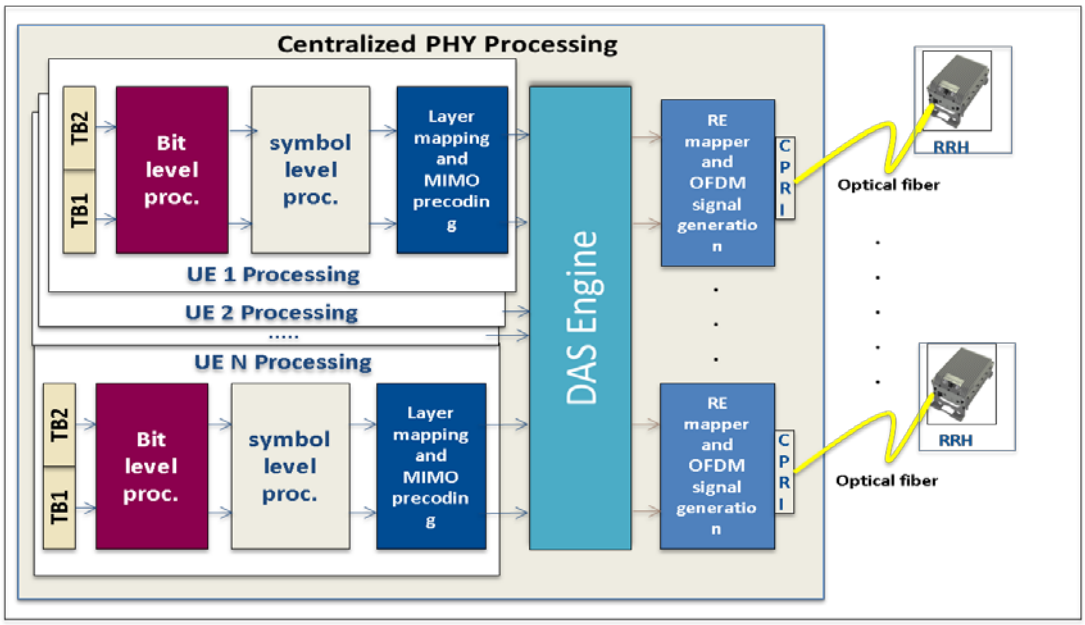


Figure 19 DL PHY processing modules layout for DAS+LTE

Figure 19 represents a PHY level, DAS centralized processing unit, where the PHY level processing for all the UEs are done, that replaces the ubiquitous base stations or cellular architecture, connected to many remote radio heads (RRHs). It is to be noted that, in the COHERENT architecture [2] R-TP and RRH mean the same, so the terms RRH and R-TP are used interchangeably in this description. Without loss of generality, LTE downlink shared channel processing for transmission has been taken as the channel of interest, to analyze realistically the necessary adaptations required for downlink shared channel (DLSCH) processing blocks to support DAS functionality. This analysis can easily be extended for all downlink and uplink channels.

For the LTE downlink shared channel processing chain [12], a case with one Transport block (TB) per user equipment (UE) per subframe is assumed. In our LTEPHY implementation, PHY starts to process user by user TB after receiving all the TBs from MAC. At first bit level like CRC attachment, code block segmentation, Turbo coding, rate matching and code block concatenation is perfomed on the TB bits. After that,symbol level processing which includes scrambling with a pseudo random sequence and modulation mapper is performed. The generated modulation symbols for each user, i.e., the symbols from all the TBs pertaining to that UE are further passed through layer mapping and MIMO precoding blocks. According to LTE specification, after the MIMO precoding the symbols are mapped into theallocated resource elements, depending on the resource allocation type and number of RBs, the OFDM signal is generated. But in the DAS, the outputs of the MIMO precoder for all the UEs allocated in that subframe could be further processed by DAS Engine block.

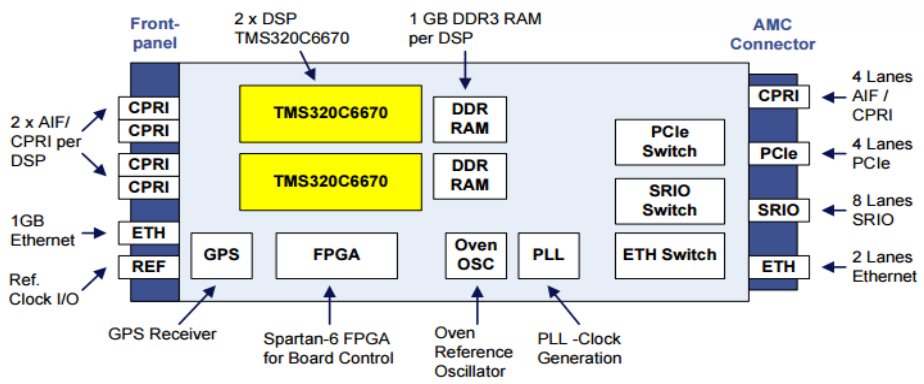


Figure 20: Dual TI DSP board

In the current work, the DAS Engine block has been implemented, which performs the selection transmission technique of DAS. Figure 20 represents the block diagram of Dual TI DSPs board, which is employed for the testbed. The driver to Antenna Interface module (AIF2) on TI DSP, which receives the OFDM time domain signal and packs then into common public radio interface (CPRI) packets, has been extended to support the second optical link in order to support the second RRH. The memory layout of the whole LTE PHY implementation has been readjusted to support the additional memory requirements of DAS Engine. The control framework of the LTE PHY implementation is also modified to support this functionality.

Figure 21 describes the software partitioning of the PHY layer on the DSP, besides showing the hardware processing blocks: AIF2, which is a hardware accelerator on TI DSP, used for baseband IQ signals transmission to the different RRHs, ARM where MAC and protocol stack (PS) can run and the Linux PC where evolved packet core (EPC) can run. The DSP can communicate with protocol stack via Gigabit Ethernet. Protocol stack communicates with evolved packet core (EPC), running on a Linux PC via Gigabit Ethernet. However in the current work only PHY processing of DAS in conjunction with a simple MAC Emulator has been implemented and tested due to non-feasibility of integration with MAC and higher layers. Although DSP is capable of processing for many users, without loss of generality, we have selected a use case with two UEs per subframe and two RRHs...

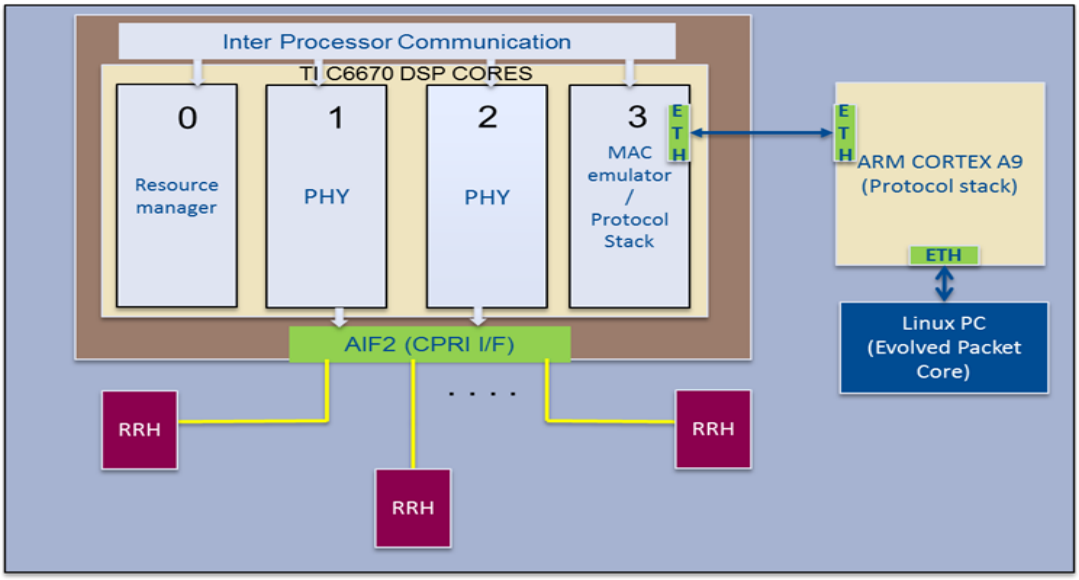


Figure 21: Hardware Implementation blocks

2.3.6.2 LTE procedures

According to the 3GPP LTE standard [13], the UE must perform certain procedures, like cell search, cell selection, decoding of system information, random access and UE attach procedure, before it can transmit and receive data. We will summarize here our investigations about deployment of DAS in a

LTE System, in the view of modification of LTE procedures such as: cell search, cell selection, random access and multi-user scheduling of resources.

Cell selection and attach procedure

In a typical LTE system, each cell is uniquely identified based on physical layer cell identity (cellID) number, that is broadcasted using primary synchronization sequence (PSS) and secondary synchronization sequence (SSS) channels. After turning on, the UE sets the radio frequency (RF) carrier frequency to the different center frequencies of the operator allowed bands according to E-UTRA absolute radio frequency channel number (EARFCN) defined in [15] [16]. The UE then tries to find the PSS, which is a part of the synchronization signals transmitted by the eNodeB. Zadoff-Chu sequences, which have good auto-correlation properties in frequency domain, are used to generate the PSS. After detection of the PSS signals, UE measures the RSSI of that cell. After completion of the cell search, the cell with the greatest RSSI value will be selected by the UE and decides to camp on the cell, which is defined in LTE as cell selection procedure [13]. In DAS, we assume different CellIDs can be transmitted from different RRHs, so that the cell selection procedure in a typical LTE cell can be applied for RRH selection. The pairing can be achieved w.r.t. the RSSI values, which is a good metric of the observed SINR at the UE location w.r.t. different RRHs. After the UE gets attached to a RRH, the eNodeB can analyze and exploit the opportunities to reuse the resource blocks (RBs) to different UEs based on CQI and the observed SINR.

RSSI based RRH pairing

Received signal strength indicator (RSSI), is defined in the standard [14] as the total signal power of each resource element (RE) including interference and noise. RE, which is one subcarrier times one symbol, is the smallest discrete part of the LTE frame, defined as 15 KHz in frequency. In a typical LTE system, each cell is uniquely identified based on physical layer cell identity (cellID) number, that is broadcasted using primary synchronization sequence (PSS) and secondary synchronization sequence (SSS) channels. After turning on, the UE sets the radio frequency (RF) carrier frequency to the different center frequencies of the operator allowed bands according to E-UTRA absolute radio frequency channel number (EARFCN) defined in [15]. The UE then tries to find the PSS, which is a part of the synchronization signals transmitted by the eNodeB. Zadoff-Chu sequences, which have good auto-correlation properties in frequency domain, are used to generate the PSS. After detection of the PSS signals, UE measures the RSSI of that cell. After completion of the cell search, the cell with the greatest RSSI value will be selected by the UE and decides to camp on the cell, which is defined in LTE as cell selection procedure [13]. In DAS system, we assume the RRHs will be transmitting different cellID, so we consider applying the cell selection procedure in a typical LTE cell to the DAS for RRH selection.

LTE based multi-user scheduling of resources in DAS

Efficient scheduling of the time frequency resources in a pairing based DAS system is the key to achieve the improvement in sum-rate of multiple users. We consider a LTE frequency division duplexing (FDD) system, where the uplink (UL) and downlink (DL) transmissions are in different frequencies separated by a duplexing distance. The effects of decreasing the duplexing distance on the DAS systems will be evaluated in the next section. The multi-user scheduling of resources in LTE is done based on the resource block (RBs), which is a 180 KHz based frequency block. A typical LTE sub-frame with normal cyclic prefix (CP) consists of 14 OFDM symbols in time domain and 50 RBs for 10 MHz bandwidth. In DL, these 50 RBs are allocated to multiple UEs based on many factors like channel quality indicator (CQI), modulation and coding scheme (MCS), quality of service (QOS) etc. The ST based pairing primarily requires the capability of overlapping time-frequency resource scheduling for the UEs that are enough distant from each other that they can be paired to different RRHs. This modification has been implemented in the testbed.

2.3.6.3 Coordinated pairing in DAS

The current work proposes a set of exploitation methods and modifications of LTE PHY processing blocks and the relevant LTE procedures at the eNodeB for supporting the functionalities of DAS in a

LTE system. Selection transmission (ST) is a simple DAS technique, which is based on selecting a single RRH out of all the RRHs for transmission, has gained considerable interest in DAS, as it reduces the other cell interference (OCI) and retains the benefits of spatial diversity. ST based pairing scheme, which pairs the UEs to the RRHs based on their distances or path losses is considered for our work. The ST based pairing offers opportunities to explore the re-use of time-frequency resources in the same cell. Probability of coverage metric is introduced to analyze different DAS scenarios.

Probability of coverage metric

Probability of coverage based on SINR is given by

$$p_c = P\left(SINR_{i,j} > \Theta\right),$$

where Θ is the minimum SINR expected at the UE to achieve the targeted throughput. As the interference is independent to the received power and the interferers are independent as well, the probability of coverage p_c can be derived [17] as

$$p_c = exp\left(\frac{-\Theta\eta {d_{ij}}^{\alpha}}{P_i}\right)\prod_{j=1,j\neq i}^{K}\frac{1}{1+\Theta\frac{P_j}{P_i}\frac{{d_{ii}}^{\alpha}}{{d_{ij}}^{\alpha}}}$$

2.3.6.4 Network graphs

One of the important metrics in selection transmission based UE and RRH pairing is monitoring the probability of coverage P_c of the UE w.r.t. all the RRHs based on the $SINR_{i,j}$ values at the UE locations. This derived probability of coverage metric is implemented in MATLAB. The results provide a reliable statistical estimation of the achieved throughput based on the targeted SINR given the channel to be Rayleigh fading. In this view, we have implemented a graphical user interface GUI in MATLAB, which provides the accessibility of modifying the relevant parameters like number of RRHs, separation distance between RRHs, targeted SINR at UE locations, UE position and attenuation. After selecting the parameters, a network graph is generated as shown in the example in Figure 22. The network graph in this example as shown in figure has 2 RRHs, i.e., RRH0 and RRH1. The values in the vertices of the network graphs are the probability of coverage of the UE with RRH0 and RRH1. Based on this network graph, the RTC can make a decision about the pairing of UE with RRH and send it back to the centralized processing of DAS. These necessary parameters and abstraction metrics are provided by the MAC and PHY layers towards the C3/RTC.

In our measurements, we have made use of different placements of RRHs and UE positions in order to understand the impact on pairing of RRH and UE. This MATLAB GUI provides a theoretical estimate of probability of coverage, which is then used as a benchmark to qualify our error vector magnitude (EVM) based measurement results. The scenarios and results are detailed in Section 7.2.

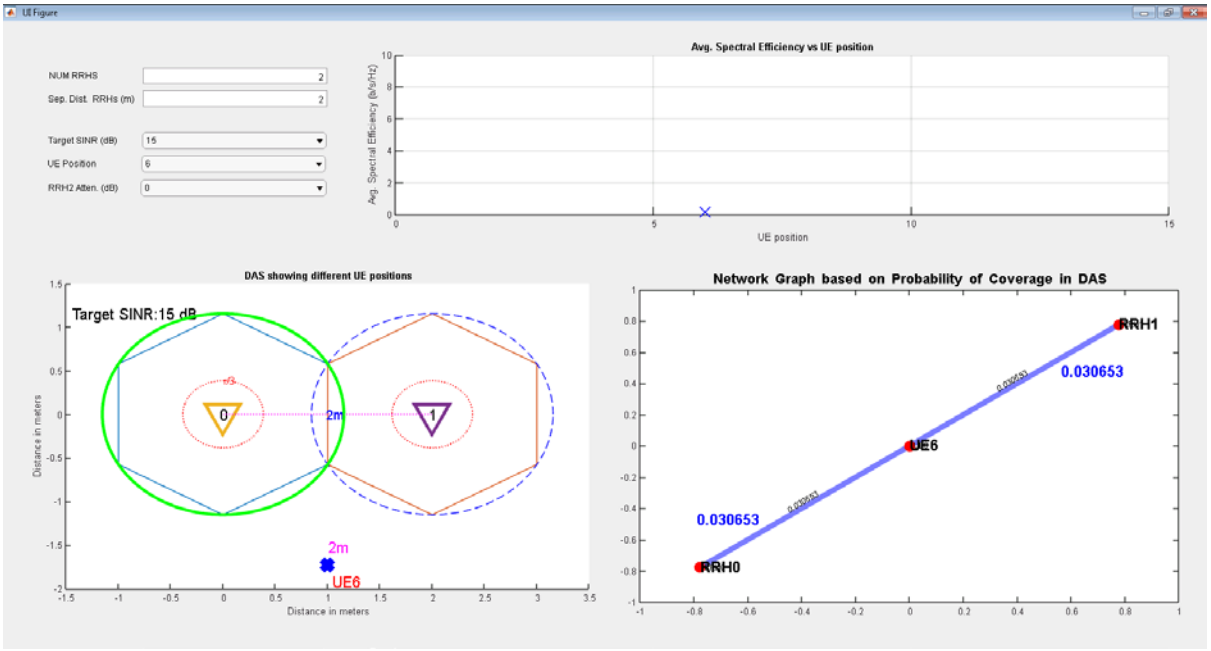


Figure 22: MATLAB based GUI for generation of network graphs in DAS

2.3.6.5 Testbed description

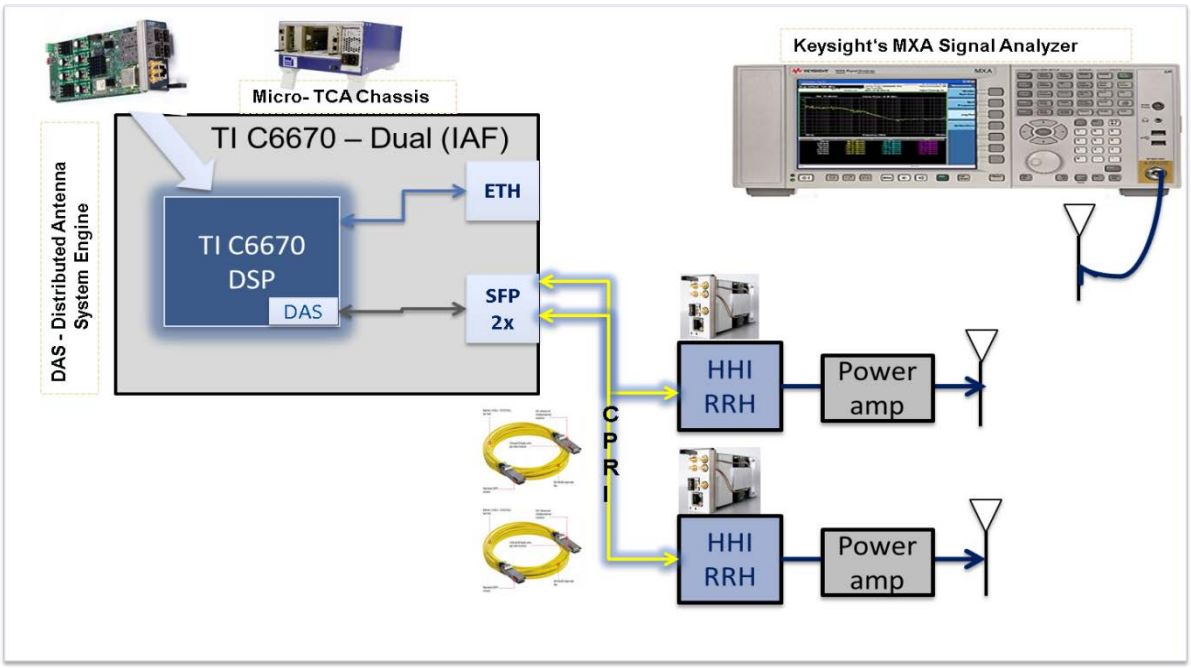


Figure 23: UDE - LTE DAS PHY testbed

Figure 23 represents the block diagram of DAS PHY testbed implemented for measuring and validating the investigations we have made in WP3 regarding coordinated pairing and scheduling of resources using Network graphs in COHERENT framework. As described in detailed in Section 2.3.6, the PHY processing based on 3GPP LTE PHY is implemented on Texas instruments (TI) C6670 DSP. IAF's microTCA chassis provides the power and cooling of different boards with AMC form factor. The Texas Instruments (TI) C6670 DSP board is one of the AMC boards and the RF board from Heinrich Heine Institute, Berlin (HHI) is the another AMC board inside the chassis. Figure 24 shows the picture of the microTCA chassis with DSP, where the PHY level processing of DAS is implemented. The RRH board is connected to the DSP board via CPRI cable through the small form-factor plugged (SFP) interface provided on the Dual 6670 AMC board provided by IAF GmbH.

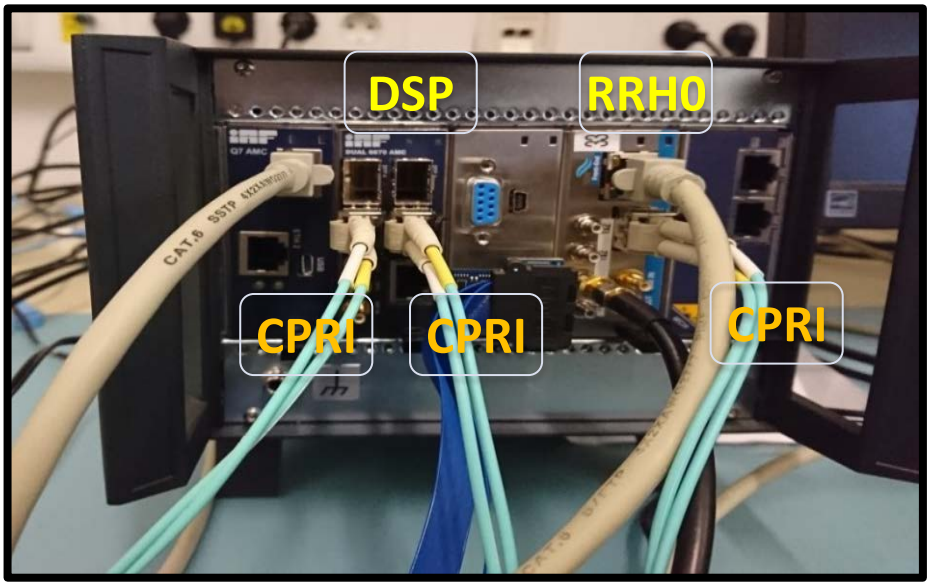


Figure 24: eNodeB Central Unit – Hardware

3. Proactive control and management using NIF based RAN Monitoring - Evaluation Results

3.1 OAI/FlexRAN LTE testbed NIF evaluations results

For the testbed evaluation of LTE RAN monitoring and proactive control decision making, we evaluate NIFs with 2 eNodeBs and 2 UEs in the network. We develop the NIF application within COHERENT SDK to work in 2 modes, online and offline modes. In the online mode the data is processed from a live system and triggers are generated in real-time. The offline mode was implemented to facilitate offline processing of the data for faster control application development. The data was collected in a JSON format such that it can be used to run applications in a "test mode" or offline mode with SDK terminology. In *test* mode, the JSON data file is played offline to process and analyze the network conditions which are used as input to the NIF application. Implementation of the stand-alone NIF is available as open-source code at [8] and the integrated code base of the implementation of NIF on the OAI testbed with a FlexRAN controller using FlexRAN SDK is also available at [10].

We evaluate the generation of handover trigger for one UE which we define as the *tagged* UE. The tagged UE was being served by eNodeB1 and had a QoS requirement that the throughput at the UE should be above 12000 Kbps over 70% of the time. This requirement is based on time-based QoS provisioning for different users in the network which requires throughput to be guaranteed for at least a specific fraction of time in the network. The second user, UE2 with a throughput requirement matching that of the tagged UE (12000 Kbps) is added to eNodeB1 with a delay of 100 seconds, requiring eNodeB1 to allocate a portion of the radio resources to this newly active UE2.

Figure 25 plots the RSRP over the serving link and shows no change correlated with the throughput drop. Figure 26 shows how the measured throughput at UE1 dropped due to the demands of resource from UE2, falling below its required threshold. OAI scheduler implements a proportional fair algorithm where equal throughput will be received by the UEs with the same channel conditions. In our setup configuration, the overall cell throughput is around 17Mbps at the best channel condition.

The NIF_app's trigger for evaluating candidate links is initiated by this degradation of throughput below the threshold more than 30 % of the time (tolerance level = 0.7) as indicated in Figure 27. Figure 28 shows the empirical CDF of the measured throughput indicating that the probabilistic tolerance level has been crossed. On evaluating the probabilistic thresholds on the candidate link between UE1 and eNodeB2 (described in [7]) as shown in **Figure 29** and comparing it to what was obtained on the link between UE1 and eNodeB1 as shown in Figure 28, a handover trigger was created requesting a handover of UE1 from eNodeB1 to eNodeB2, and passed on the controller application that chooses to act on this information.

With this evaluation we demonstrate the benefits of proactive control actions based on probabilistic models that allow us to react to network state variability based on how it affects the user QoS. We also demonstrate how predicted attainable throughput, a probabilistic RAT-agnostic metric allows us to make mobility management decisions (among others) in a centralized manner within a RAT-agnostic scenario.

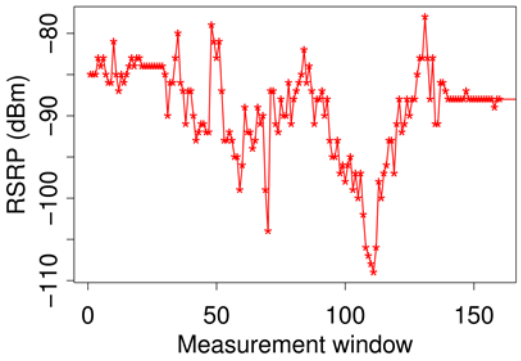


Figure 25: RSRP variations are not always correlated with measured throughput variations

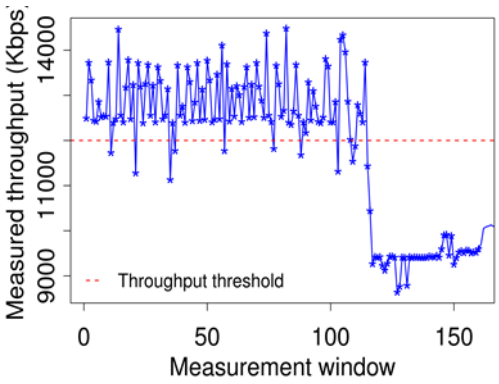


Figure 26: Measured throughput variation for tagged UE

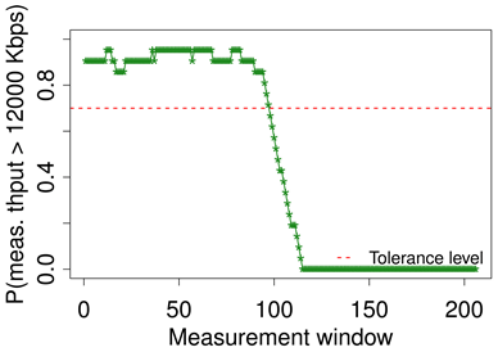


Figure 27: Probabilistic throughput, QoS, variation for tagged UE

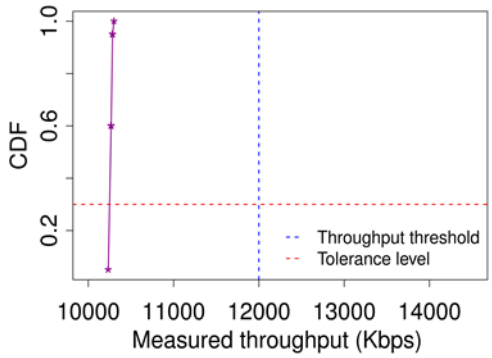


Figure 28: Empirical CDF of measured throughput for tagged UE at eNodeB1

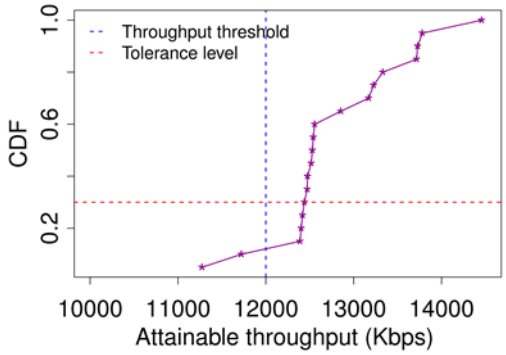


Figure 29: Empirical CDF of attainable throughput for tagged UE at eNodeB2

3.2 5G-EmPOWER Wi-Fi testbed NIF evaluation results

The evaluation of proactive control and management in Wi-Fi RANs was done through mobility management using NIF to perform intra-RAT Wi-Fi handover. The setup included 2 WTPs in the network and 2 clients. Channels 1 and 6 were used as described previously in Section 2. We observe handover trigger for the client under observation which we call the tagged client. The tagged client was being served by WTP1 on channel 6. It had a QoS requirement on throughput that the obtained throughput should be above 14000 Kbps with a probability of 0.7. Client 2 was added to WTP1 with a certain delay and a throughput requirement of 5000 Kbps causing a portion of the radio resources to be consumed by this client. This resulted in the degradation of throughput at client1 or the tagged client as shown in Figure 30. As the throughput at the tagged client dropped, the trigger for evaluating candidate links was initiated by this degradation of throughput distribution observed in the form of an empirical CDF.

Variation in measured throughput at the tagged client in reference to the required threshold of 14000 Kbps is shown is Figure 31. The probability that throughput stays above this threshold is given by Figure 32. As this value drops below the acceptable 0.7 tolerance level to (1-0.4) 0.6 as shown in the

CDF plot in Figure 33, a trigger to evaluate candidate link is generated. Figure 34 shows the CDF of the estimated attainable throughput on the candidate link from the tagged client to WTP2. We see that the probability of attainable throughput being higher than the threshold of 14000 Kbps is (1 – 0.1) which is 0.9, which is higher than the required tolerance level of 0.7. On evaluating this candidate link between the tagged client and WTP2, and comparing it to what was obtained on the link to WTP1 a handover trigger was created if the candidate link offers a stricter adherence to the require QoS or throughput threshold. This handover trigger moves the client's information stored in the LVAP at WTP1to the target WTP2 and sends out a CSA (channel switch announcement) message to the tagged client indicating to it that it must change the channel it now functions on after the predefined interval of 3 beacon time periods. This CSA message completes the handover of client1 from WTP1 to WTP2.

Figure 30 presents a view into how the system would have responded if the handover was purely based on RSSI. We see that even though the measured throughput distribution at the client varies due to varied channel availability due to increasing clients, RSSI is not correlated with this change. We should emphasize that a purely load based handover mechanism would also have a downside of not accounting for the MCS at which traffic can potentially be sent (based on SNR and hence RSSI). Therefore a throughput variability based handover captures both the variations in sending MCS (using RSSI), and traffic load [7].

This implementation and evaluation of handover between 2 WTPs demonstrates the applicability and benefits of probabilistic throughput based handover trigger and the probabilistic attainable throughput based selection of candidate links so as to meet the QoS requirements of users in a dynamic network. Implementation of the stand-alone NIF is available as open-source code at [8] and the integrated code base of the implementation of NIF on the 5G-EmPOWER testbed with a 5G-EmPOWER controller using its SDK is available at [11].

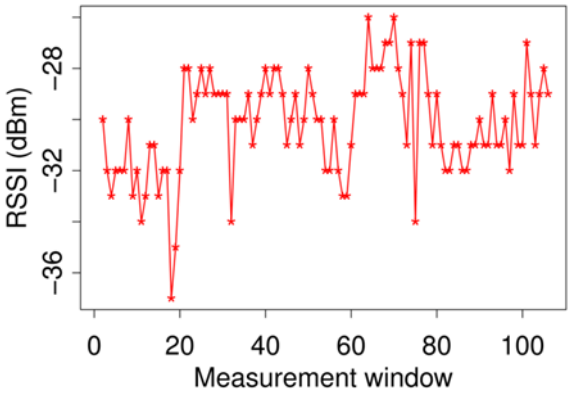


Figure 30: RSSI variations are not always correlated with measured throughput variations.

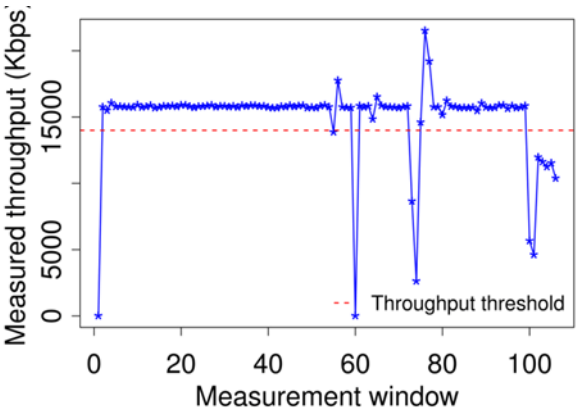


Figure 31: Measured throughput variation for tagged client.

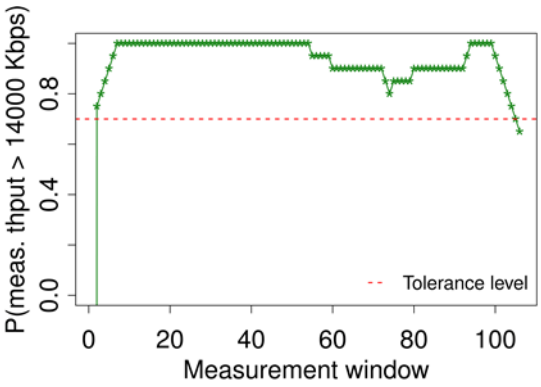


Figure 32: Probabilistic throughput, QoS, variation for tagged client.

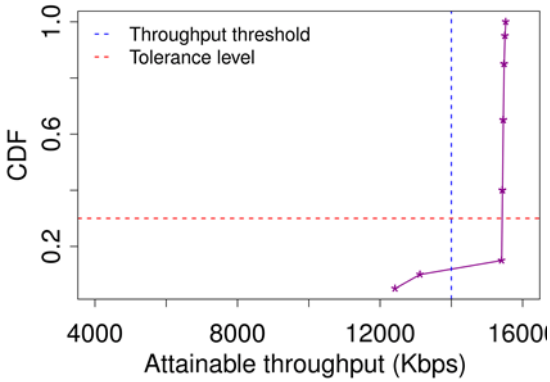


Figure 33: Empirical CDF of measured throughput for tagged client on WTP1.

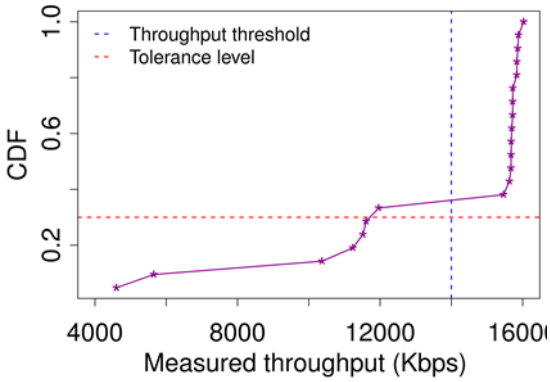


Figure 34: Empirical CDF of attainable throughput for tagged client on WTP2.

4. RAN Inter-App Communications - Evaluation Results

We implemented a remote control application to show the network topology status and the control for SMA application (cf. Section 7.1). The PoC Remote control application is a HTML5 JavaScript application running on any client browser. Before showing or controlling anything, the information sources must be configured beforehand. Figure 35 shows a simple sources configuration for only one FlexRAN-rtc (for eNodeB/UE information) and one example application (SMA).

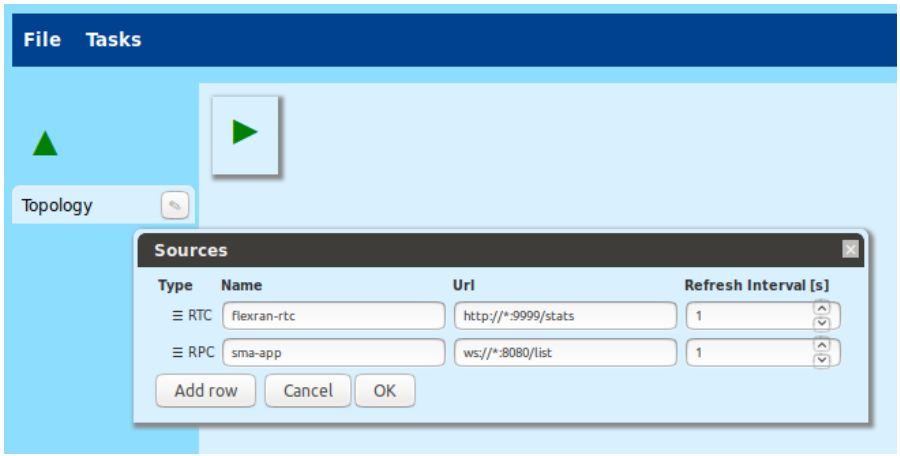


Figure 35: The information sources for the remote control

The control application has a built-in support for FlexRAN-rtc "stats" output. From this built-in implementation, it can construct a network graph of known eNodeB's and their connected UE's for different remote control applications. It also recognizes the SMA's "get-list" notification format, which indicates current desired frequency parameters for each eNodeB. From this information, the application constructs and shows the system components and their connections graphically in real time.

Figure 36 shows the graphical interface to monitor the network status. It shows two active UE devices, connected to different eNodeBs, and both pulling a video stream over the Internet. At the same time, the SMA is trying to change the frequency of an eNodeB. Both eNodeBs are connected to Internet via same EPC node (not shown).

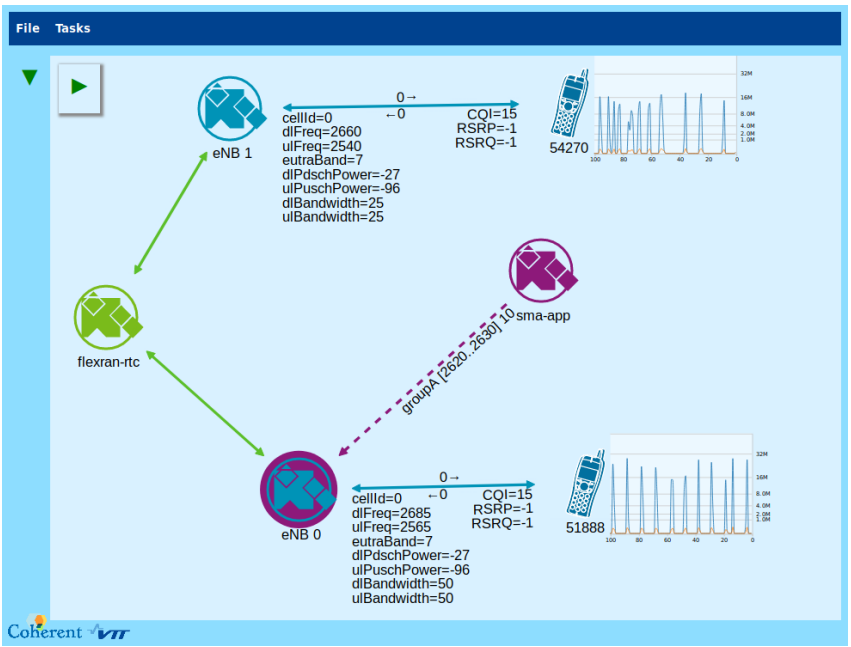


Figure 36: The example of GUI to show system state

For the remote control, if the remote control application receives a "Capabilities List" notification from a configured source, it will construct a control panel section for that application. For example, assuming the SMA app sends the following capabilities notification:

```
{"method": "capabilities",
 "params": {
   "get-list": {"help": "Get the current list"},
   "set_rule_group_A": {"help": "Prefer lower cost", "group": "rule"},
   "set_rule_group_B": {"help": "Prefer higher bandwidth", "group": "rule"},
   "set_rule": {
       "help": "Change rules for 'enb_id' and/or 'group'",
       "schema": [
           {"name": "enb_id","type": "number","choice": [None,"#ENBID"]},
           {"name": "group", "choice": ["A", "B", "C"]}]},
   "enable_graph": {"help": "Turn on graph.", "group":"graph"},
   "disable_graph": {"help": "Turn off graph.", "group":"graph"},
   "save_status": {"help": "Calls method to save current app status"},
   "load_status": {"help": "Calls method to load current app status"}
 }
}
```

Figure 37 displays the resulting control panel section for the SMA app and the parameter filling panel, which pops up when the "set_rule" has been selected. As an example, if, in the parameter specification, enb_id is selected as 234881024 and group as A, accepting the input panel, generates the following request message to the sma_app:

```
{"method": "set_rule",
 "params": {"enb_id": 234881024, "group": "A"},
 "id": "set_rule"}}
```

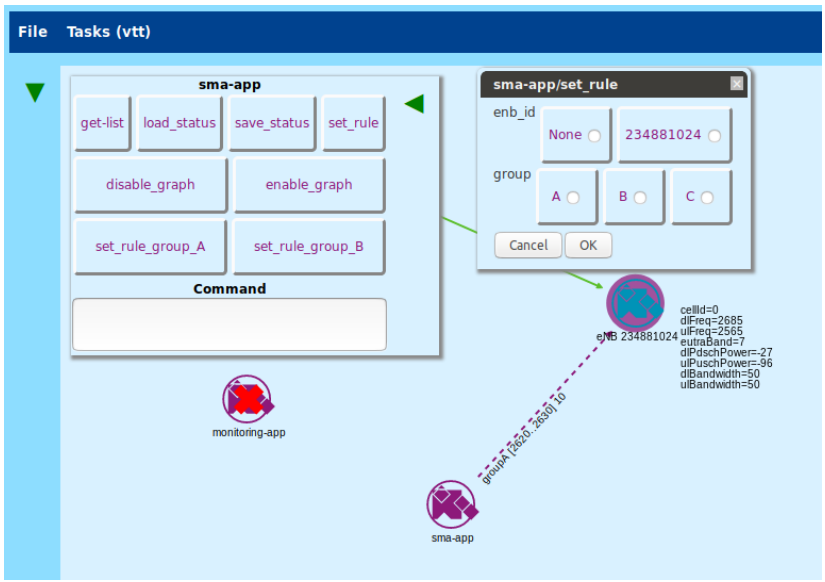


Figure 37: Control Panel with SMA app capabilities activated

We show how the remote control app controls the spectrum band of eNodeB through SMA app from Figure 38 to Figure 42. We show only one eNodeB and one UE in the network. When the system is initialized, as shown in Figure 38, the eNodeB selects the Spectrum Group A, where the downlink and uplink carrier frequency are 2625MHz and 2505MHz, respectively. From the control panel interface for SMA app, we change the spectrum of eNodeB to Group B, where the downlink and uplink carrier frequency are 2650MHz and 2530MHz respectively. This operation is shown in Figure 39. The SMA receives the command from the remote control application and applies the change to eNodeB through FlexRAN-RTC. The change will take effect at eNodeB after several seconds. We reconnected the UE to eNodeB and now the UE works on new carrier frequency. The result after the change is shown in Figure 40. We can also change the carrier frequency to Group C, where the downlink and uplink

carrier frequency are 2633MHz and 2513MHz respectively. The control operation and the result after the change are shown in Figure 41 and Figure 42 respectively. Consequently, the remote control application controls the other control applications based on the COHERENT SDK and provide the network visualization to show the network statistics. The remote control app can be extended to show network graph at the different abstract layer as the future work.

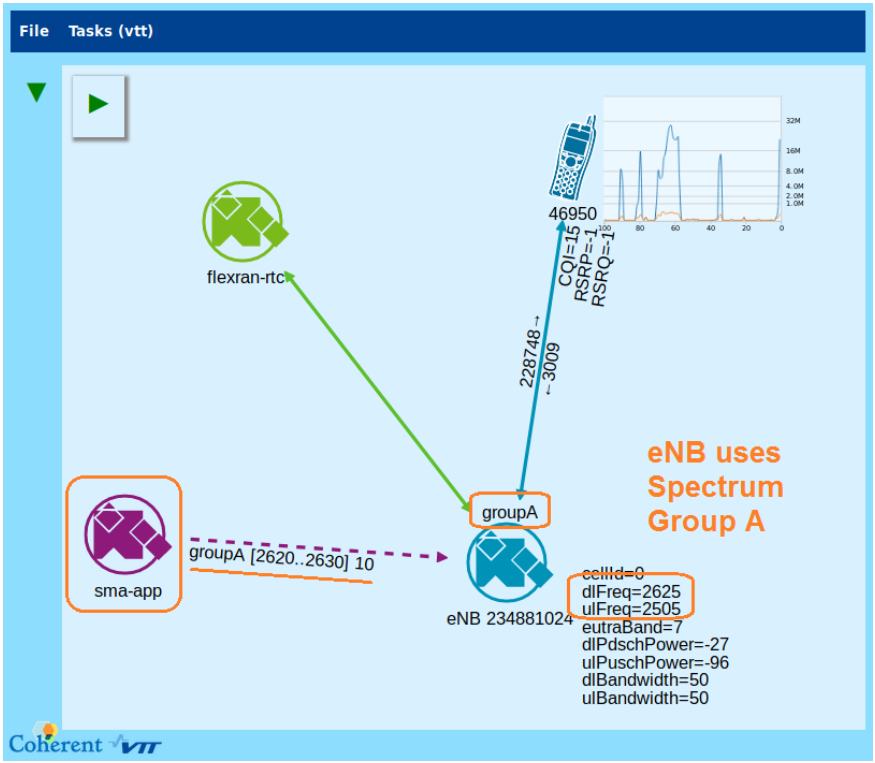


Figure 38: Step 1 - Initialize eNodeB works on Spectrum Group A

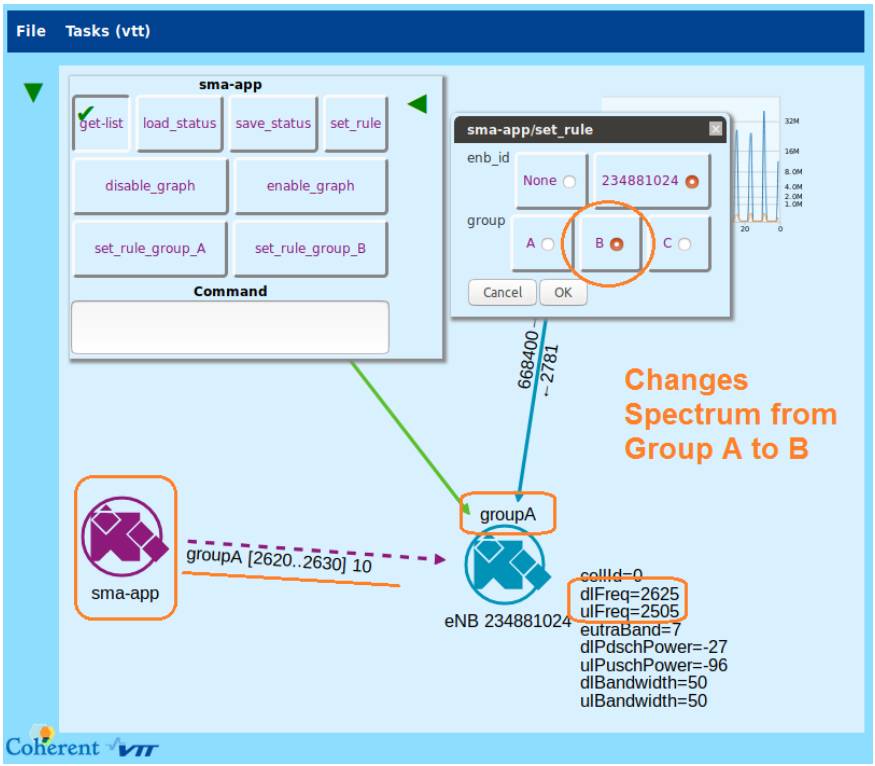


Figure 39: Step 2 - Change eNodeB Spectrum Group from A to B

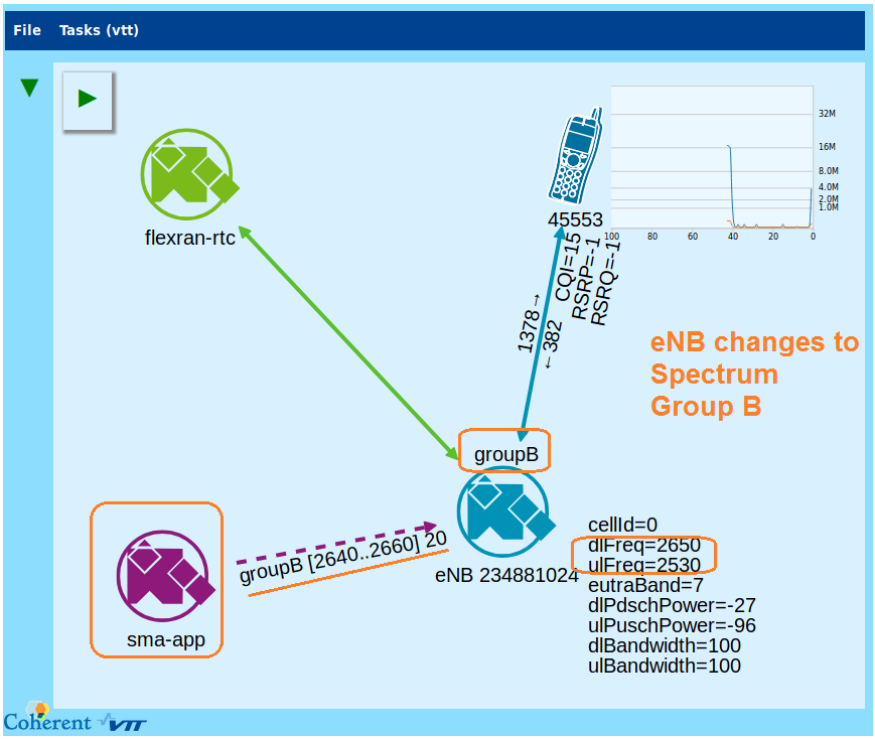


Figure 40: Step 3 - eNodeB works on Spectrum Group B

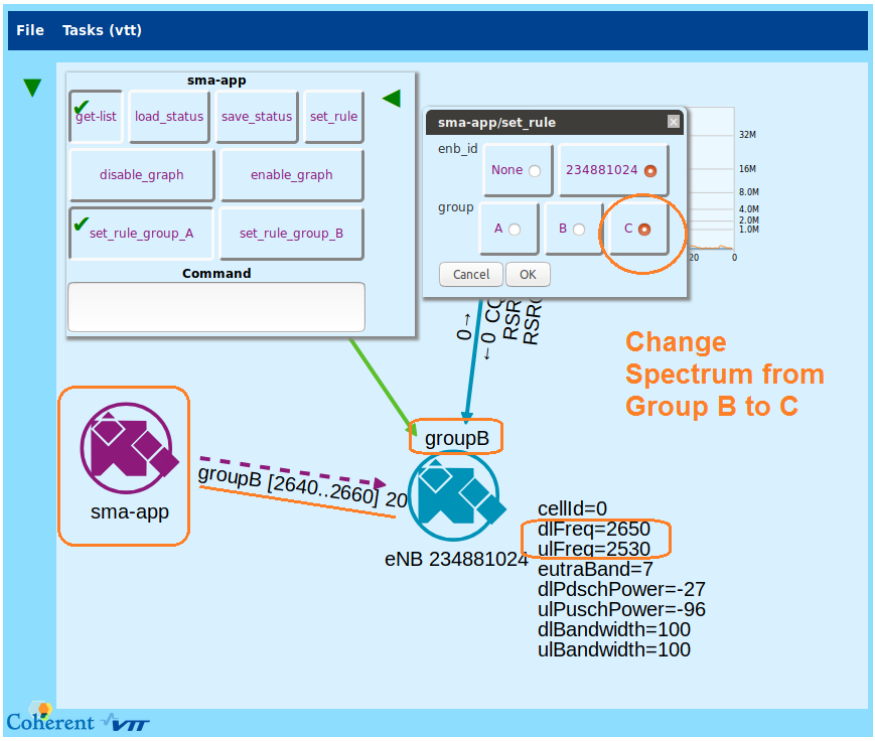


Figure 41: Step 4 - Change eNodeB Spectrum Group from B to C

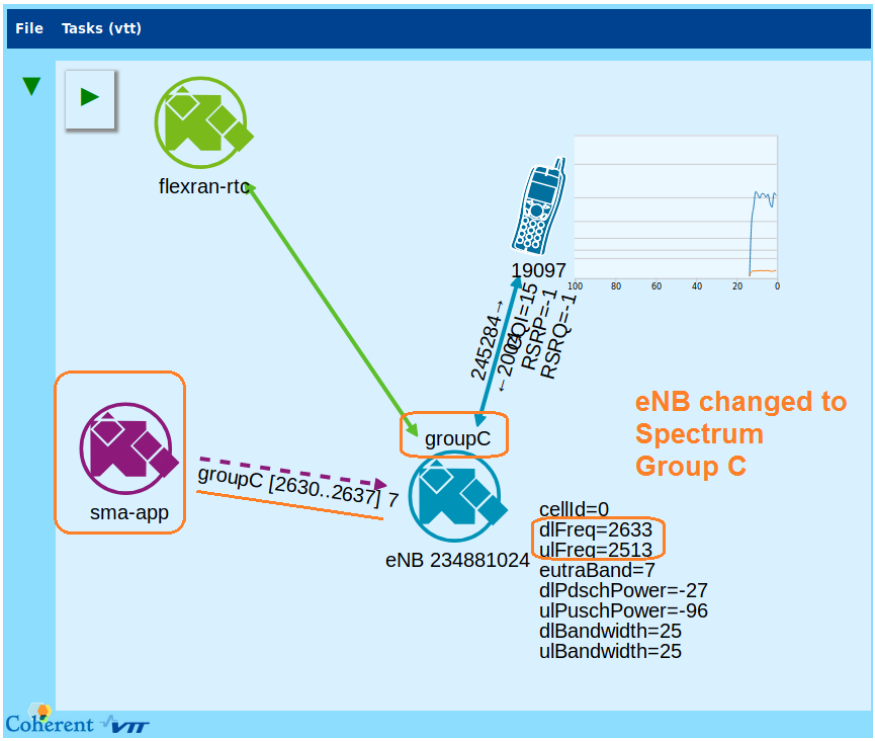


Figure 42: Step 5 - eNodeB works on Spectrum Group C

5. RAN Sharing – Evaluation Results

Stemmed from the 3G/4G era, the notion of RAN sharing aims to utilize available resources at the RAN segment and to reduce the capital expenditures, while still satisfy a certain level of quality of service. In concept, there are two types of RAN sharing model. The passive RAN sharing case aims to share the passive network elements such as cell sites, antenna masts and power supply. However, the active RAN sharing can share the active network elements (e.g., transport infrastructure), a portion or all radio spectrum, and physical/virtual network functions. Further, 3GPP defines two levels for active RAN sharing. The Multi-Operator RAN (MORAN) approach shares the same RAN infrastructure but with dedicated frequency bands for different operator, while Multi-Operator Core Network (MOCN) allows to also share the spectrum among operators as standardized in LTE Release 8.

Based on the RAN sharing concept, the RAN slicing issue is correspondingly raised in 5G to provide different levels of dedication and sharing to allow a slice owner to customize its service across user-plane, control-plane, and control logic while increasing the resource utilization of RAN infrastructure. In that sense, our focus is on the provision of the RRM application that can program the control logic to be enforced at the level of physical resource blocks (PRBs) through providing the virtualized resource blocks (vRBs) by a novel resource visor toward each network slice to provide the isolation between slices. Further, the RRM application will rely on the resource abstraction/virtualization scheme to have a slice-specific virtualized view of the underlying resources (i.e., not one-to-one mapping to the physical resources) and the performance isolation between different slices can be guaranteed, i.e., the performance of one slice will not be impacted by others. Such application development is based on the FlexRAN platform aforementioned in 2.2.2 with the exposed SDK.

5.1 Processing flow

The processing flow of RRM application is provided in Figure 43. Before enabling the RRM application, the service template shall be updated in which the requested resource, data rate, latency are described for each deployed service (detailed later). Afterwards, the RTC shall register the status update to the agent from the underlying BSs. Upon receiving the latest status from the SDK in every *T* seconds, the RRM application will generate the RRM policy to be validated by RTC, enforced by agent at the MAC scheduler to reconfigure the downlink/uplink scheduler at corresponding BSs. In the following paragraph, we will elaborate more details on the input (i.e., service template) and output of the RRM application (i.e., RRM policy).

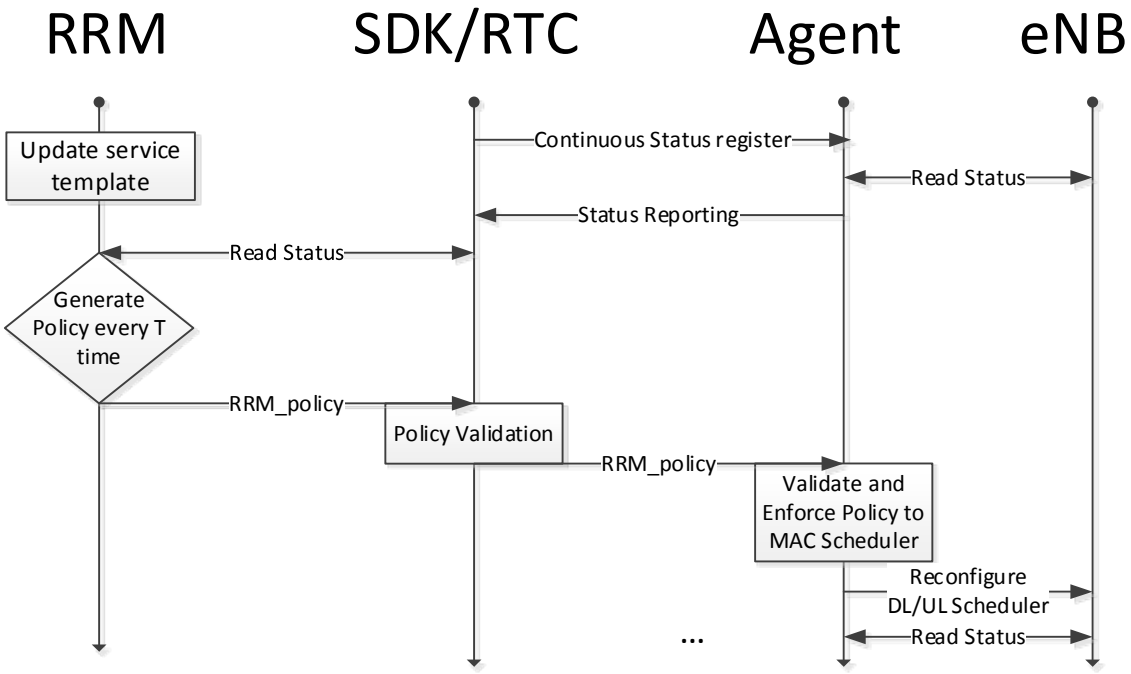


Figure 43: Processing flow of RRM application

The input and output of RRM application are shown in Figure 44, and the service template include following information:

- **enb_slices**: It defines the cell service context in terms of number of instantiated slices and their service type
    - 'name': It is the name of the target BS that incorporates the cell types (phantom, small, macro) with the supported features.
    - 'cell_id': It represents a unique BS identifier that is a concatenation of the BS module ID (first 16 bits), eNodeB ID (unique in a PLMN, 32 bits) and the physical cell ID (16 bits). Note that aforementioned numbers of bits are the ones used in the FlexRAN platform rather than the ones defined in the 3GPP standard.
    - 'service_type': It represents the number of slices in the BS and the applied policy for each slice in DL and UL directions. Note that the association of a user to these slices can be asymmetric in DL and UL directions.
- **Service policy**: It includes the specific service policy to be applied in terms of the radio resources
    - 'requested_vrbg': The number of requested virtual resource block group (vRBG). Note that the vRBG can be mapped from the physical resource block group (PRBG) but is not necessary follow the 1-to-1 mapping (i.e., resource virtualization).
    - 'requested_rate': The requested data rate (i.e., in the level of kbps) for the service to be provided in the long run. Such requirement does not explicitly state the number of resource block which shall be translated by the RRM based on latest status.
    - 'requested_latency': The requested latency (i.e., in the level of second) for the service to be provided.
    - 'requested_priority': The service priority to be utilized by the RRM to accommodate the unused resources and preempt resources of other slices.
    - 'requested_resource_isolation': It indicates that the requested resource shall be dedicated or shared. Note that such property can allow a slice to reserve its resources (i.e., without multiplexing).

Some additional parameter may also be included in the template such as the positioning of vRBG in the frequency domain.

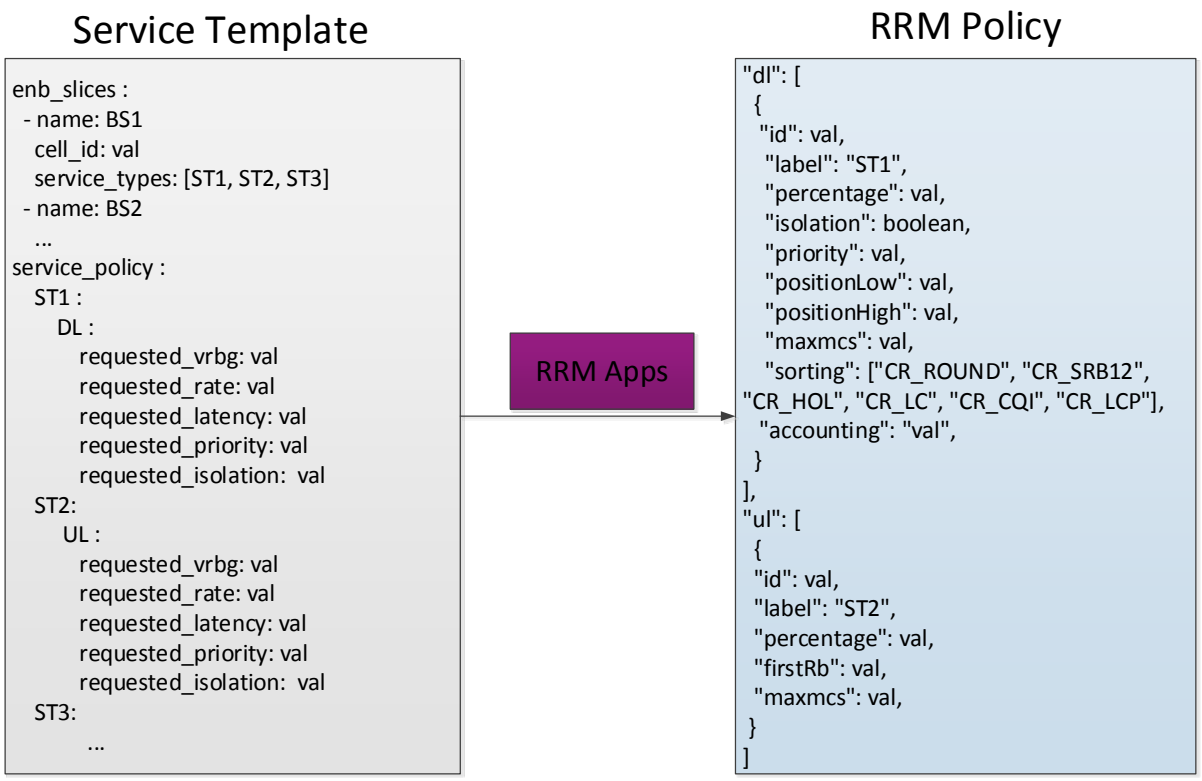


Figure 44: Input and output of RRM application.

The RRM application gets the RAN-agnostic service template as a high-level input and generates a RAN-dependent RRM policy (cf. right part of Figure 44) that will be translated into low-level primitives in the RAN agent when the policy is executed. The RRM policy includes the following information:

- 'dl' or 'ul': It describes the link direction, uplink and downlink, to which the RRM policy is applied.
- 'id': It is the unique identifier of a slice that requires to add/modify/delete a RRM policy in a given direction.
- 'label': It represents the service type of this slice, which is used by the agent to apply the proper slice QoS based on the service level agreement (SLA).
- 'percentage': It indicates the amount of resources that this slice requires for a given direction. This actual amount will be decided by the agent and MAC scheduler based on the slice priority, required level of isolation, and the actual aggregated BS workload (i.e., PRB utilization ratio).
- 'isolation': It describes whether the downlink radio resources must be isolated for a given slice. Specifically, it means that the amount of resources allocated to this slice will be reserved and dedicated, and thus such dedicated resources will not be shared.
- 'priority': It defines the order in which different slices must be scheduled. If a slice has the maximum priority, it can also pre-empt the radio resource of other slices if they are not isolated.
- 'positionLow' and 'positionHigh': They define the portion of the spectrum in terms of its low and high position in frequency domain subject to the admission control.

- 'maxmcs': It defines the maximum modulation and coding scheme such slice can apply for, which will be adapted by the MAC scheduler based on the actual channel quality of particular users.

- 'sorting': It instructs the scheduler on how the users must be sorted within a slice (intra-slice), for instance, CR_ROUND suggests to use the round-robin manner (i.e., fairness) between users.

- 'accounting': It determines how the slice-specific resources (defined by its percentage) must be allocated to the slice users.

- 'firstRb': It represents the location of the first uplink radio resources for a given slice, which is used to position the slice user (similar to 'positionLow' and 'positionHigh' in the downlink direction).

5.2 Evaluation Results

In this section, we evaluate how the RRM application can impact the perceived UE performance and the quality of experience (QoE). The first two experiments show the impact on the user-plane perceived performance according to different RRM policies with specific focuses on resource partitioning and slice priority respectively. The last experiment compares the impact of different user profiles on the QoE through the RRM application.

5.2.1 Resource partitioning impact on user-plane performance

In this experiment, we instantiate two different slices at the RAN domain and there is one UE for each slice, respectively. In the beginning of the experiment, the RRM applies the 50/50 policy between these two slices, i.e., 50 percent of radio resource are dedicated for slice 1 and another 50 percent of radio resource are for the slice 2 as shown in Figure 45. Then, the UDP traffic is transmitted in the downlink direction toward both UEs simultaneously to utilize the overall 5MHz radio bandwidth. We can observe that both UEs experience the same level of good-put (at the top) and delay jitter (at the bottom) from 0 s to 45 s time duration. However, the RRM policy is changed into 20/80 policy from 45 s to 50 s, in which only 20 percent of radio resources are dedicated to slice 1, while 80 percent of dedicated radio resources are reserved to slice 2. We can observe that, from 50s to 100s time interval, the perceived good-put is significantly increased and delay jitter is largely reduced for the UE that attaches to slice 2. In contrast, the UE that is attached to slice 1 suffers from a lower good-put and higher delay jitter. Note that such change of RRM policy may be resulted from the change of input service template of slice 2 which requests more dedicated vRBG (i.e., 'requested_vrbg') and it is a case which illustrates Scenario 1 in Section 2.3.3.4, where the slice 2 corresponds to a PMR service which, at a certain moment in time, uploads a new RRM policy with more demanding requirements.

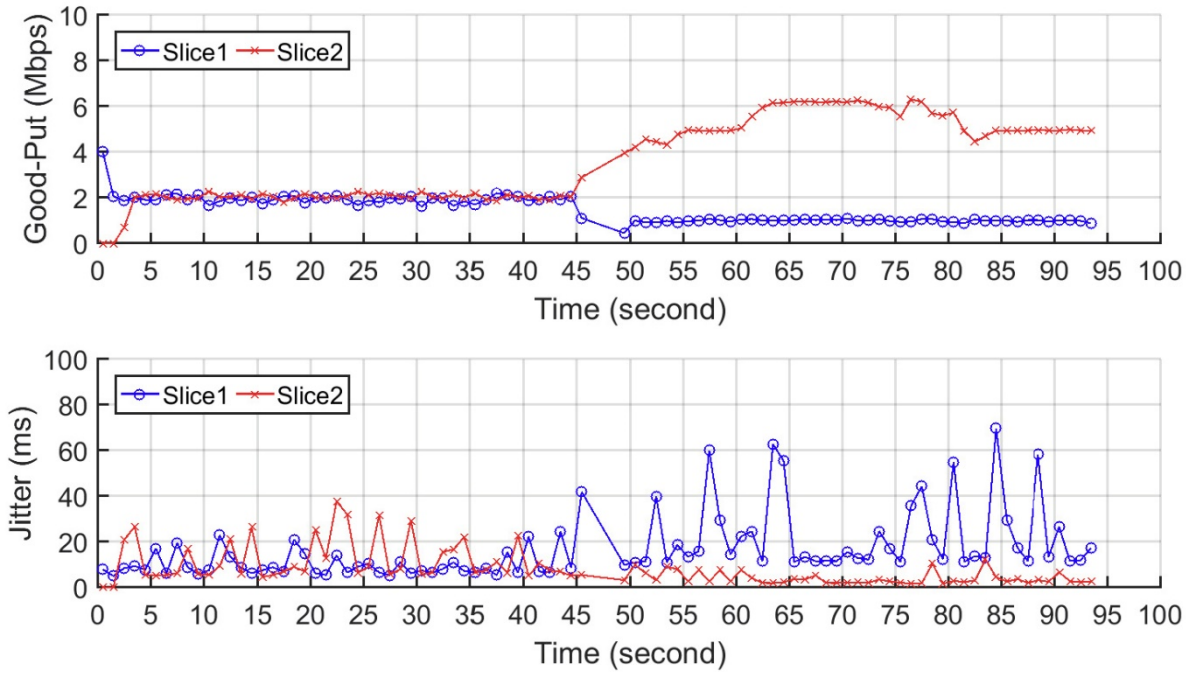


Figure 45: Impact of slice resource partitioning on good-put (top) and delay jitter (bottom).

5.2.2 Slice priority impact on user-plane performance

In the next experiment, we inspect the impact of the slice priority on the perceived user-plane performance. In Figure 46, there are three slices with different priority levels: (1) slice 1 can pre-empt resources of all other slices when the actual (aggregated) rate exceeds the requested data rate, (b) slice 2 can only increase its multiplexing gain by utilizing the unallocated resources, and (c) slice 3 may sustain its requested data rate as it can neither pre-empt nor multiplex resources but is subject to the pre-emption from high priority slice (i.e., slice 1). Furthermore, there are three UEs attached to three different slices respectively. When examining the good-put and delay jitter, it can be seen that slice 1 can flexibly adapt its data rate as a function of workload via pre-empting the resources from other slices, i.e., from 3 Mbps to 6 Mbps, while slice 2 experiences a data rate drop from its desired 10 Mbps to 8 Mbps. The same trend is observed in the delay jitter measurement (at the right), where slice 1 experiences minimum jitter as it has the highest priority (even its workload is increased from 3 Mbps to 6 Mbps). However, slice 3 suffers from the highest delay jitter since it has the lowest priority. To sum up, the RRM application can enable some high priority slices (e.g., Scenario 3 for public safety, as being already discussed in Section 2.3.3.4) to adapt the date rate to the dynamic workload, while other low priority slice will be served in the best effort manner.

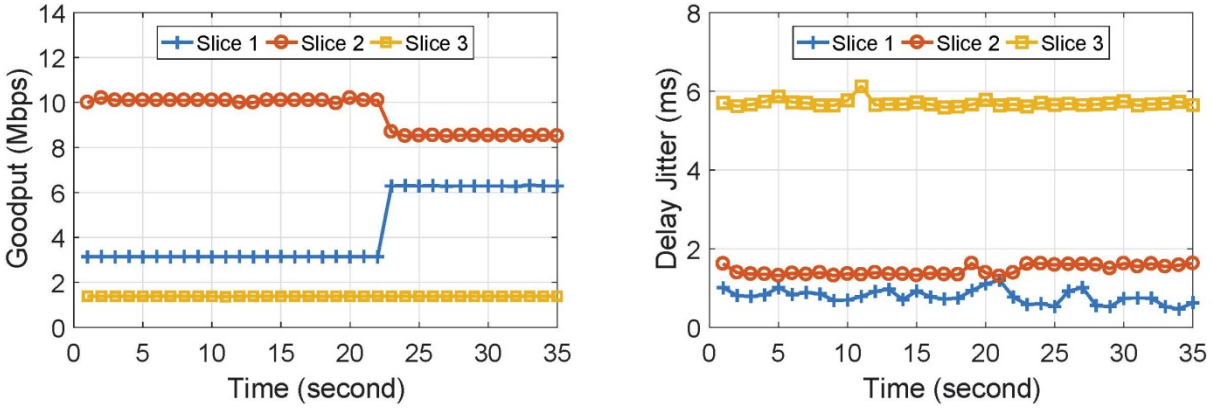


Figure 46: Impact of slice priority on good-put (left) and delay jitter (right).

5.2.3 User classification impact on QoE

The last experiment aims to compare the experienced QoE when different user profiles are used by RRM application. Here, we classify all users into 5 different classes: Super, Platinum, Gold, Silver and Bronze. These five user classes will restrict the maximum amount of radio resources a user can consume in downlink and uplink directions. Specifically, only 100%, 80%, 50%, 30%, and 10% of radio resource can be consumed in correspondence to super, platinum, gold, silver and bronze user class respectively. These definitions utilize aforementioned 'percentage' in the RRM policy as shown in the right part of Figure 44. Hence, these five user classes can be seen also as five different slices each one containing users of the same class.

To quantify the overall QoE of different user classes, we use the nPerf application installed on the commercial-off-the-shelf (COTS) UE that can measure the ping round-trip-delay (RTT), uplink/downlink bitrate, browsing rate and video streaming rate as shown in Figure 47. The super user gets the best score that considering bitrates, ping RTT, web browsing and video streaming performance rate. Moreover, platinum users have a close score to the super user, while there are performance drops for gold and silver users in web browsing and bitrate. We can see that even though the average data rate is dropped almost at the ratio of 2 (8.55 Mbps/3.03 Mbps and 4.96 Mbps/1.51 Mbps for gold and silver user in downlink/uplink respectively) from gold to silver user class, the user experience score is only reduced by approximately 12%. The reason behind is that the exact QoE is not only related to the network characteristic but also to the content to be delivered. We can see that the scores of web browsing and video streaming drop a little as the quality of these contents can still be acceptable (e.g., both user classes can still stream 360p video and provide an acceptable web browsing experience). Whereas bronze users suffer from significant drops in the bitrate, web browsing and video streaming, e.g., only 240p or lower video bitrate are supportable. Summing up, such experiment results reveal that the RRM application can distinguish user experience via modifying the RRM policies.

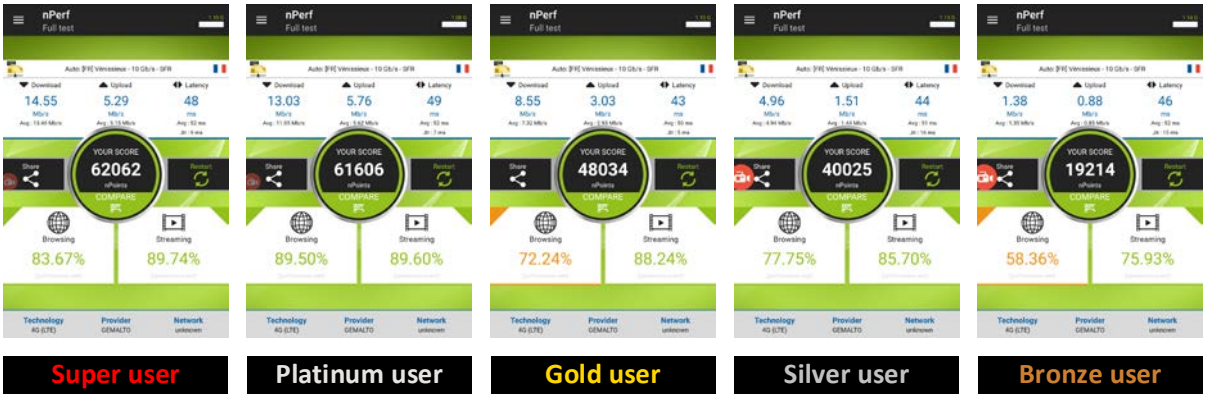


Figure 47: Impact of user classes on the perceived QoE.

5.3 Summary

In this chapter, we provide the experiment results for the RRM application via applying different RRM policies for different slices in terms of resource partitioning, slice priority and user class. These different policies can also be the high-priority user/slice to adapt the data rate based on their current workload. Note that more sophisticated control logic can be provided via chaining of SMA (cf. Section 7) and RRM application, e.g., adapt the RRM policies based on the SMA control decisions in terms of changing the carrier frequency or radio bandwidth.

6. Load Balancing – Evaluation Results

6.1 Inter-RAT Scenario: LTE and Wi-Fi

The results for Inter-RAT Load balancing are given in Section 6.5.3 of D5.2 [21] where an algorithm for traffic–Aware User Association in Heterogeneous RANs have been designed and verified with simulations and a proof of concept. Offloading decisions are done at the C3 controller by offloading low-priority traffic to the Wi-Fi access point in case of resource utilization at eNodeB becomes above 80%. Through offloading to Wi-Fi, more resources at eNodeB become available for high-priority flows and for new clients.

6.2 Intra-RAT Scenario: LTE

The sequence diagram of load balancing for LTE with C3 assistance is shown in **Figure 48**. The resource utilization of RAN Node and UE reports are passed to COHERENT coordinator where they will be abstracted for technology-agnostic graph construction. The C3 detects that the other neighboring RAN Nodes are not loaded and from the channel point of view are acceptable for some UEs. Then, COHERENT coordinator will request the serving eNodeBs to do handover of UEs to the other RAN Nodes and then UEs can extend their throughputs. Thus, by distributing UEs among RAN Nodes, COHERENT coordinator decision takes into account not only radio conditions, but also cell load for efficient spectrum management and sharing. The algorithm for load balancing is specified formally in Deliverable D3.2 [7].

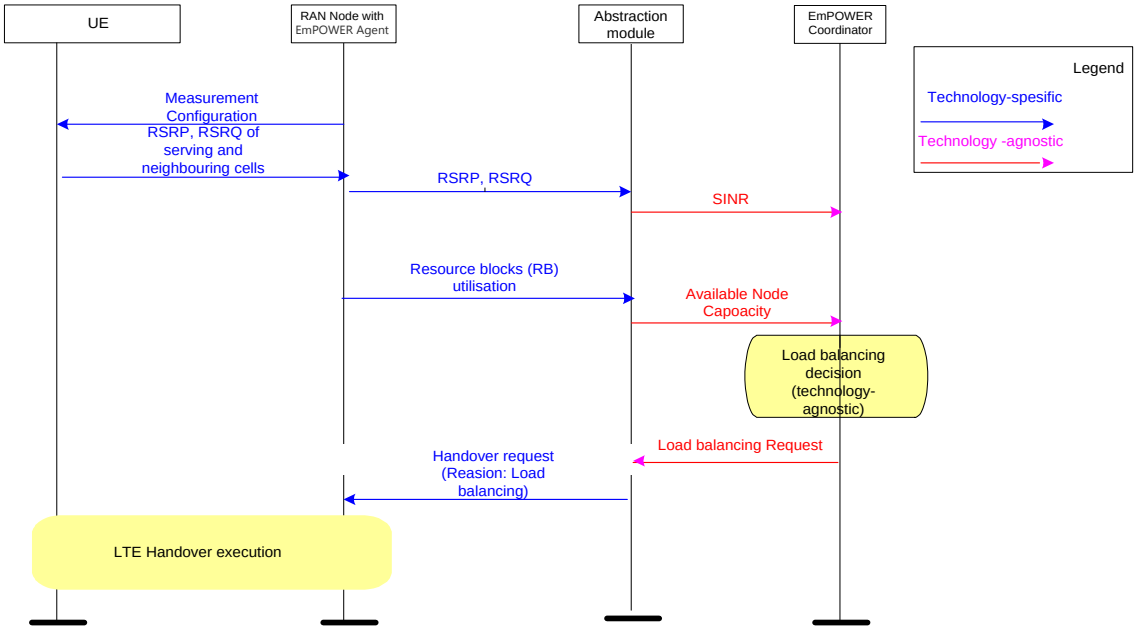


Figure 48: Sequence diagram for load balancing in LTE

The verification of the algorithm correctness was done with 2 real eNodeBs acting as a RAN Nodes connected 5G-EmPOWER acting as C3 and 1-2 commercial UEs.

At the beginning of the test, 2 UEs are connected to one eNodeB based on radio condition. The second eNodeB will have no UEs attached (the UEs have a weaker channel to it comparing to the first eNodeB). The 5G-EmPOWER Agent at eNodeBs provides the 'slave' counterpart in the controller-agent communication of 5G-EmPOWER coordinator that acts as a part of COHERENT coordinator. The resource block usage of two eNodeBs and UE reports with RSRP/RSRQ measurements of neighboring cells are passed to 5G-EmPOWER where they can be abstracted for technology-agnostic graph construction. The 5G-EmPOWER detects that the second eNodeB is not loaded at all and from the channel point of view is acceptable for UEs. Then, 5G-EmPOWER requests the first eNodeB to do handover of one UE to the second eNodeB and then both UEs can extend their throughputs.

The part of the log demonstrating connection of eNodeBs, reporting and handover command is listed below:

```
#2 eNodeB are connected to 5G-EmPOWER
INFO:core.service:Incoming connection from ('172.25.128.148', 55570)
INFO:core.service:Incoming connection from ('172.25.128.149', 36779)

#capability requests sent to 5G-EmPOWER to eNodeBs
INFO:vbsp.vbspconnection:Sending caps_request to 00:00:00:00:00:00 at
172.25.128.148 last_seen 0
INFO:vbsp.vbspconnection:Sending caps_request to 00:00:00:00:00:01 at
172.25.128.149 last_seen 0

#5G-EmPOWER receives capability requests sent to 5G-EmPOWER to eNodeBs
INFO:vbsp.vbspconnection:Got message type 2 (caps_response)
INFO:vbsp.vbspconnection:caps_response from 172.25.128.148 VBS
00:00:00:00:00:00 seq 1
INFO:core.pnfdev:PNFDev 00:00:00:00:00:00 mode connected->online


INFO:vbsp.vbspconnection:Got message type 2 (caps_response)
INFO:vbsp.vbspconnection:caps_response from 172.25.128.149 VBS
00:00:00:00:00:01 seq 1
INFO:core.pnfdev:PNFDev 00:00:00:00:00:01 mode connected->online

#eNodeBs send periodically hello messages to keep connection
INFO:vbsp.vbspconnection:hello from 172.25.128.148 VBS 00:00:00:00:00:00
seq 441
INFO:vbsp.vbspconnection:Got message type 1 (hello)
INFO:vbsp.vbspconnection:hello from 172.25.128.149 VBS 00:00:00:00:00:01
seq 1025
INFO:vbsp.vbspconnection:Got message type 1 (hello)

#5G-EmPOWER receives updates information about connected UEs (at this
moment one UE has been attached)
INFO:vbsp.vbspconnection:ue_report_response from 172.25.128.148 VBS
00:00:00:00:00:00 seq 457
INFO:core.ue:UE c47fa5d8-360f-494a-92e1-299bb6ea65a5 transition None-
>active
INFO:core.service:UE JOIN c47fa5d8-360f-494a-92e1-299bb6ea65a5 (001f01)


#5G-EmPOWER receives reports about Resource Block utilization from EnodeB
1
INFO:vbsp.vbspconnection:Got message type 6 (mac_reports_resp)
INFO:core.service:Received mac_reports response (id=2, op=1)
INFO:core.app:New mac report received from vbs 00:00:00:00:00:00 eNodeB id
0 cell 0 DL_earfcn 1500 UL_earfcn 19500

#5G-EmPOWER receives reports about Resource Block utilization from EnodeB
2
INFO:vbsp.vbspconnection:Got message type 6 (mac_reports_resp)
INFO:core.service:Received mac_reports response (id=1, op=1)
INFO:core.app:New mac report received from vbs 00:00:00:00:00:01 eNodeB id
1 cell 1 DL_earfcn 1500 UL_earfcn 19500

#UE's RRC measurements of neighboring cell are received
INFO:vbsp.vbspconnection:Got message type 5 (rrc_response)
INFO:core.service:Received rrc_measurements response (id=1, op=1)
INFO:core.app:New rrc measurements received from 5db03458-dc63-43af-a4cd-
a2776e22e077

#Start load balancing module
INFO:root:Importing module: empower.apps.handovermanager.handovermanager
INFO:core:Registering 'empower.apps.handovermanager.handovermanager'
INFO:core.app:Setting control loop interval to 5000ms

#Trigger handover
INFO:vbsp.vbspconnection:Sending ue_ho_request to 00:00:00:00:00:00 at
```

```
172.25.128.148 last_seen 264

#Receive handover response
INFO:vbsp.vbspconnection:Got message type 7 (ue_ho_response)
INFO:vbsp.vbspconnection:ue_ho_response from 172.25.128.149 VBS
00:00:00:00:00:01 seq 257
INFO:core.ue:UE 5db03458-dc63-43af-a4cd-a2776e22e077 transition
ho_in_progress->active
```

For large scale verification of load balancing with the indicated approach, the second experiment has been executed where 15 UEs and 4 eNodeBs are simulated, while a real instance of C3 (5G-EmPOWER) is used. At the starting point, all UEs are connected to eNodeB 1. After other 3 eNodeBs are added in the neighborhood, C3 starts to handover some UEs to the neighboring eNodeBs that finally results into even distribution of UEs between 4 eNodeBs as shown in Figure 49 .

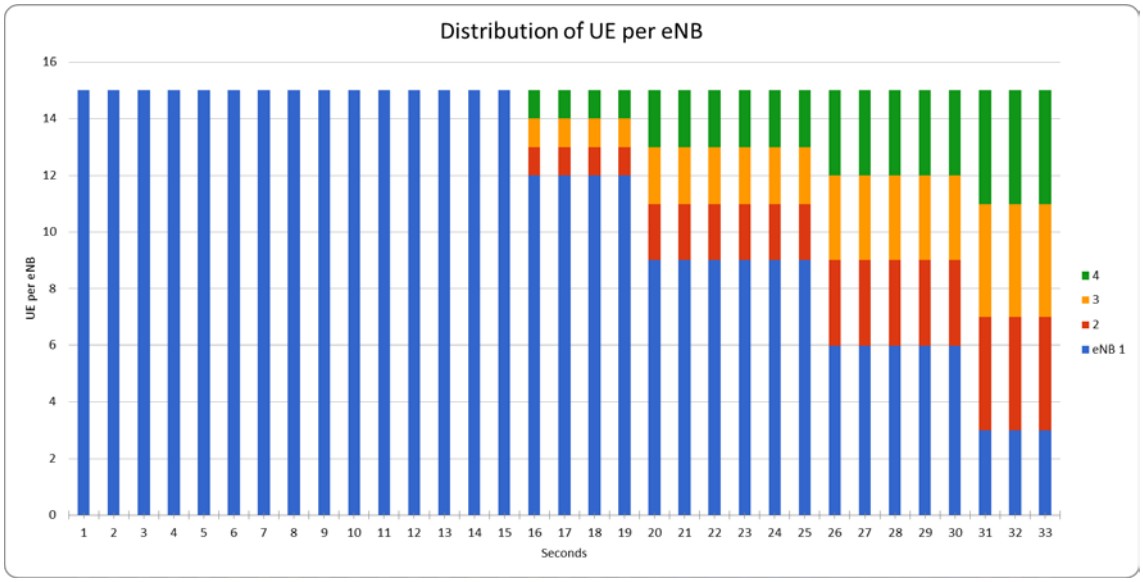


Figure 49: Distribution of UEs between eNodeB over time

In terms of throughput improvement through load balancing, Figure 50 shows an average number of RBs that each UE received during the experiment. It can be seen that at the starting time, all UEs are connected to one eNodeB and evenly share its bandwidth of 25 RBs, such that each UE received about 1.6 RBs in average. After load balancing, between neighboring eNodeBs each UE received 5RBs in average, which was sufficient to fully cover data traffic demand of UEs.

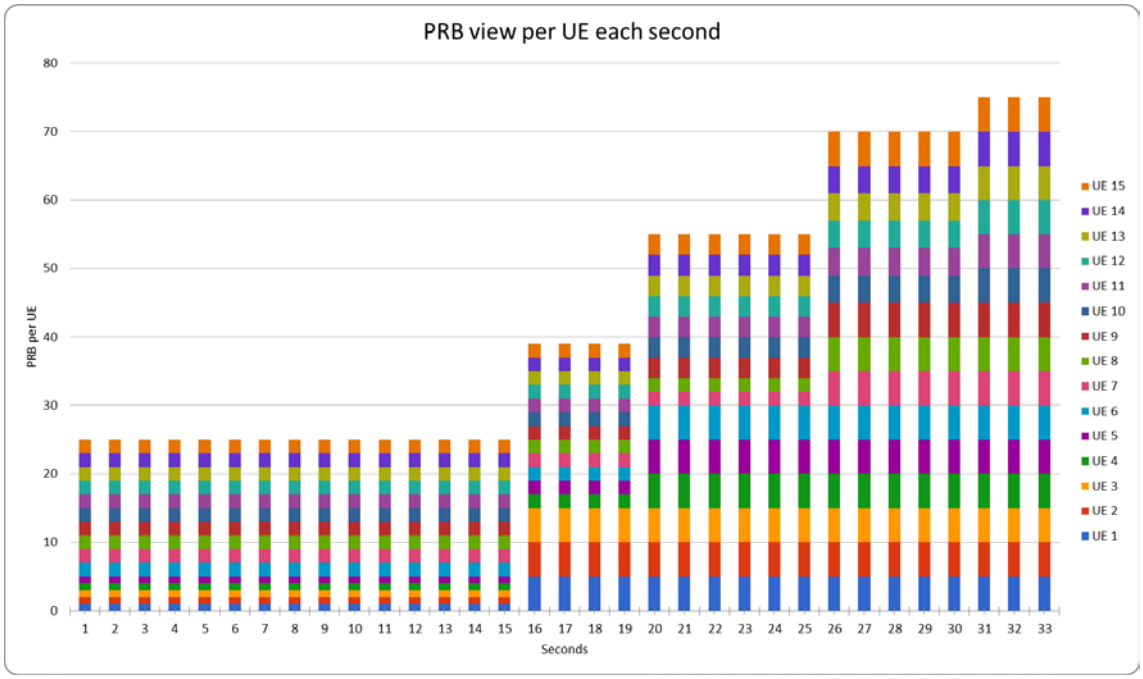


Figure 50: Average number of RBs per UE over time

During experiments, it was observed that the C3 coordinator shall consider not only the load criteria for handover decisions. When selecting UEs for handover, it shall necessary consider link quality to the serving and the target cell to avoid handover failures or radio link failures after a short time after successful handover.

7. Spectrum Management – Evaluation Results

7.1 SMA application

In the upcoming 5G era, one key challenge is on the efficient utilization of the available cross-domain resources to provide and fulfill versatile service requirements such eMBB, uRLLC and mMTC. Among the resources across multiple domains, the radio spectrum seems to be the most challenging since there are limited number of available radio bandwidth and carrier frequency and thus an efficient utilization is necessary to enable service coverage and scalability. Furthermore, the allocated radio spectrum will highly impact on the underlying RAN performance. For instance, a larger bandwidth (e.g., E-band) millimeter wave (mmWave) spectrum is more suitable for the scenario with small cells and low user mobility but it can exploit massive number of antennas (i.e., Massive MIMO) to form a directional beam toward a single user to boost the link budget and provide better mobile broadband quality of experience. In comparison, the low carrier frequency cases (e.g., lower than 2GHz) brings a larger and deeper coverage area but with limited radio bandwidth. Finally, the sub-6GHz case (e.g., 3GHz to 4GHz) can compromise from aforementioned two options.

Beyond aforementioned characteristics of different bands, the network operator needs to pool its available radio bandwidths and dynamically provide the pricing model to better utilize the available radio resources and to maximize its revenue. For instance, the price of the same radio spectrum in an office building area can be different during weekdays and weekends based on the expected traffic dynamics. Also, the network operator can dynamically deploy the available radio bandwidth based on the moving of crowd during some events (e.g., festivals). These examples show that the spectrum shall be managed flexibly and efficiently to satisfy the goals of both service provider (e.g., service availability) and network operator (e.g., resource utilization and revenue) in a dynamic manner (i.e., unlike the static planning in the 3G/4G era). In this sense, we evaluate the prototype of SMA of LTE system in this section based on the FlexRAN SDK exposed by the FlexRAN platform mentioned in Section 2.2.2. Note that the following evaluation results are based on the LTE state-of-the-art; however, the extension to support various 5G radio spectrum such as aforementioned mmWave and sub-6GHz can be naturally done via some proper modifications.

7.1.1 Processing flow

In Figure 51, the processing flow of the SMA application is shown. First of all, the database (DB) located at the SDK can collect and populate the spectrum related information, and the SMA application can utilize the SDK to load the policy and rules that shall be followed when generating its control logic. Then, the RTC can register to the agent in order to get the latest status that can be loaded from the underlying RANs. These statuses include the spectrum information such as available bandwidth and carrier frequency to be exposed to SMA to make the control decisions. After that, the SMA policy will be provided by SMA, validated by SDK/RTC, and enforced by agent to the underlying RANs. Last but not the least, the BSs can provide some feedback information (e.g., unfeasible policy, conflict decision) to be utilized by SMA in order to adapt its decisions.

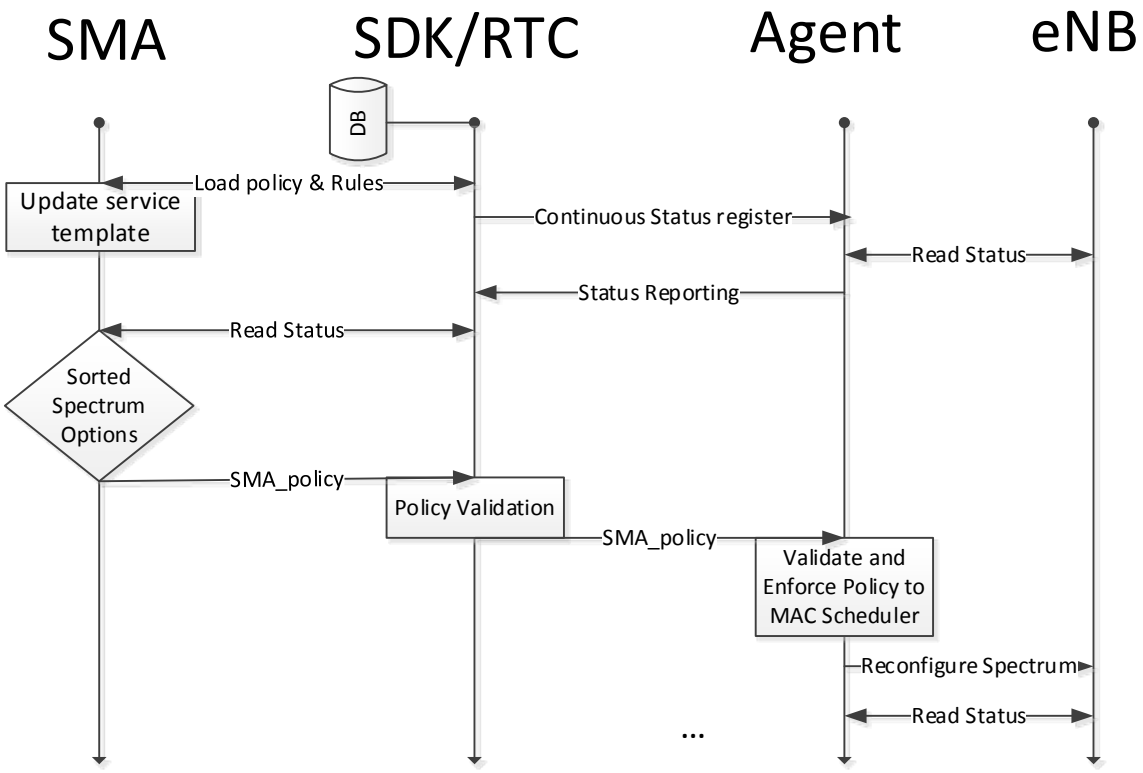


Figure 51: Processing flow of SMA application

Then, Figure 52 presents the high-level architecture of the SMA application. First, all the policies are processed and coupled (i.e., general policy, operator policy and LSA policy) with the sensing data from the region of interest. Then, the spectrum management rules are applied at the level of BS, which associate a weight to each available spectrum based on a number of criteria (e.g., price, bandwidth, BS group membership, etc.), and finally the decision making algorithm will provide a list of available spectrum sorted based on the criteria.

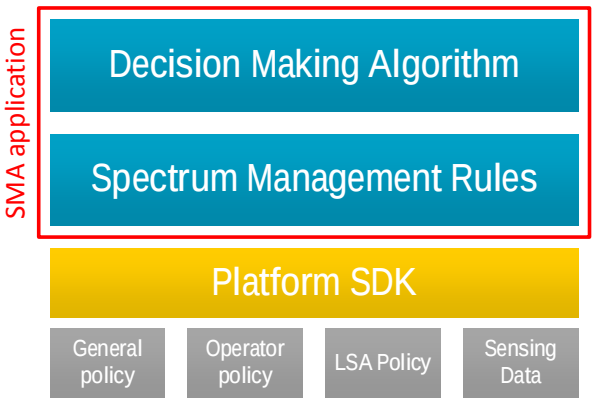


Figure 52: High-level architecture of SMA application

To be more specific, the policy and rules that are stored in the DB is shown in the Figure 53 (left part) as the input to the SMA application, and the SMA policy as the output generated by the SMA application is also shown in the right part of Figure 53. We will give more detail descriptions over such input/output in the following paragraph.

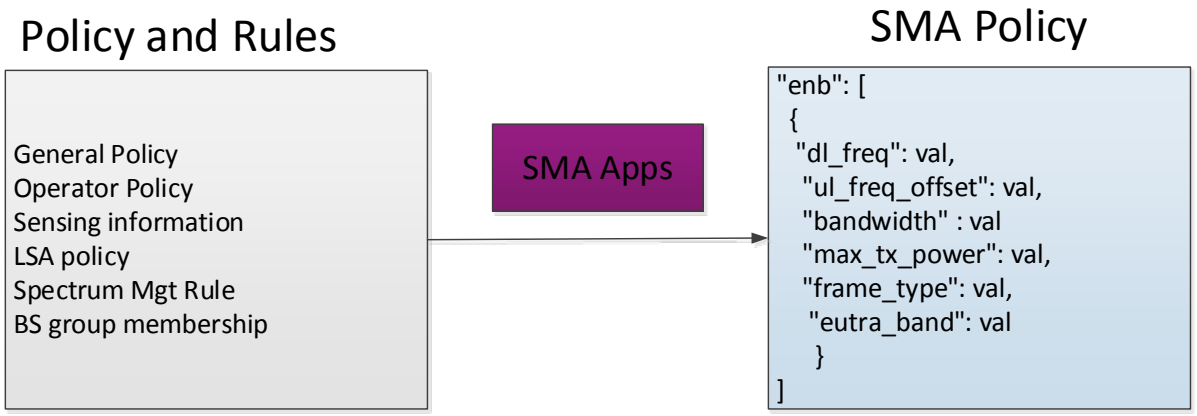


Figure 53: Exposed policy and rules and SMA policy.

Please note that the data retrieved from the databases create various opportunities for spectrum use, thus any interests stakeholder (e.g., mobile network operator) may define its own preferences in order to select the best adjusted option. The details on exemplary policies applied for spectrum opportunity selection are discussed in details in the next section.

Now, the pattern used to calculate the weight of each spectrum offer is defined for each group based on the input policy and rule files that are described below in more details:

- **General policy:** It contains a list of general spectrum usage policies of specific region or country (provided by, e.g., NRA such as Federal Communications Commission (FCC), for US, Office of Communications (OfCom) for UK, or other regulatory bodies like Electronic Communications Committee (ECC) for Europe), for instance, 'freq_min', 'freq_max', 'frame_type' and 'max_tx_power' correspond to the minimum carrier frequency, maximum carrier frequency, frame type (e.g., FDD/TDD) and maximum transmission power, respectively.

- **Operator policy:** It contains the parameter for a list of spectrum offers. Such spectrum can be shared by several operators with some utilization rules, which is defined by the same group of operators. Followings are the specific parameters for each offer:

    - 'operator' - Name of operator that defines this policy
    - 'freq_min' – A list of lower boundaries of possible bands
    - 'freq_max' – A list of higher boundaries of possible bands
    - 'busy' – Such option will be used when the activity of other BSs of this operator will be detected between 'freq_min' and 'freq_max'
    - 'idle' – Option that will be used otherwise
    - 'sub_freq_min' and 'sub_freq_max' – The limits defining a band (possibly sub-band) which can be used when in "busy" state
    - 'power_mask' – Define the power transmission power mask (interference to the narrowest channels/bands)
    - 'min_lease_time' – Minimal time interval (for example, in the level of ms). After this time interval, new spectrum request (query) has to be sent. Such mechanism can provide the periodic update scheme like heartbeat in CBRS model.
    - 'max_lease_count' – Maximum count of intervals (for example, 200*100ms) that the band can be used
    - 'max_time_to_leave' – Maximum time to leave the band when other BSs using this band are detected
    - 'price' – The price (for example in euros) to use such band for 'min_lease_time' time duration

- 'sensing_sensitivity' – Sensing sensitivity in dBm

- **Sensing data:** It includes the detected eNodeB information (i.e., not controlled by the same RTC) with specific frequency range and the corresponding operator name

- **LSA policy:** It includes the same parameters as in 'operator policy', but without breakdown for option "busy" and "idle"; these sets of policies are used to identify the ways to activate licensed spectrum sharing

- **Spectrum management rules**: Include following parameters:

  - 'mvno_group' – Name of set of rules
  - 'pattern' – A custom pattern to calculate weights of spectrum offers other than the default one, for instance, one may use custom parameters including 'freq_min', 'freq_max', 'price', 'bandwidth', 'min_lease_time', 'max_lease_count', and 'max_time_to_leave'
  - 'use_pattern' – Such yes/no option indicates whether switch to aforementioned custom pattern; otherwise, the default pattern will be used
  - 'cost' - A table in which one may set a weight of each parameter of offers in line to the default pattern. Note that such value will be ignored when using the customized pattern.
  - 'operator_preference' - Preference for the spectrum owner (i.e., we can prefer some operators). Note that this value can be set to a 'do-not-care' value to be fair between operators.
  - 'criteria' - A list of criteria that can pre-exclude some offers beforehand based on their parameters. For instance, the maximum 'price' value can be set and some offers will be excluded for consideration when their price is larger than this value. Note that the same mechanism can be set for other parameters.
  - 'Freq preference' – It can be used to indicate the preference of band in term of the absolute normalized distance (in frequency domain, in Hz) between available, offered frequency bands and the preferred band. For instance, if there are two feasible offers over 3.5GHz and 2.6GHz and we prefer 2.8GHz frequency, then the 2.6GHz offer is more preferable.

- **BS group membership:** Determine to which MVNO group a given eNodeB (with a particular cell identity) belongs to

In the output of SMA policy (cf. right part of Figure 53), it includes several parameter values: 'freq_min', 'freq_max', 'max_tx_power', 'fdd_spacing', and 'band' that correspond to the minimum frequency, maximum frequency, maximum transmission power, spacing between uplink and downlink direction in FDD mode, and the band number, respectively. Such policy will then be applied through RTC and an Agent toward the target base station.

7.1.2 Evaluation Results

In following paragraphs, we provide the evaluation results of the SMA application. In the first part, we show how the spectrum is selected. Then, we examine the user perceived performance when applying different spectrum at the phantom cell.

7.1.3 Frequency Selection

In this experiment, we change the association of a given eNodeB to different MVNO groups via applying different policies to select the best spectrum for the phantom cell deployment. Note that such eNodeB will have different BS group membership when associated to different MVNO groups. In particular, there are 4 considered selection policies as follows:

- Policy A: Selects the cheapest spectrum no matter on the bandwidth
- Policy B: Selects the spectrum with the larger bandwidth no matter on the cost

- Policy C: Selects the band with the lowest minimum lease time
- Policy D: Selects the band with the highest lease time

In Figure 54, we can see at the top that there are 5 available offers with their respective randomly-generated price in terms of euros per second between 2.6GHz to 2.7GHz. These 5 offers will be selected from time-to-time based on the aforementioned 4 policies in a specific order: Policy A → Policy B → Policy A → Policy B → Policy C → Policy D. While the considered policies are purely spectrum-oriented and user-agnostics, more complex policies can also be build that incorporate the perceived user performance when a given spectrum is allocated to a particular BS.

When applying policy A (refers to 5s to 75s and 100s to 155s in Figure 58), the one with the lowest price (refers to the highlighted band in Figure 54) will be selected when comparing with other policies. In contrast, when policy B is applied (refers to the result from 80s to 95s and 160s to 265s in Figure 58), the one with the largest bandwidth (refers to the selected band in Figure 55) will be chosen and the price will be much higher than that of applying policy A. While the policy C (refers to 270s to 350s in Figure 58) aims to select the offer with the lowest minimum lease time to match its desired utilization duration (e.g., a service may only need a short peak throughput to carry a single big file) as highlighted band sown in Figure 56. Finally, we apply policy D (refer to 365s to 480s in Figure 58) and the resulted selection (highlighted band in Figure 57) will be the one that can last longer in terms of the least time. To sum up, the most suitable offer which can be selected is dynamically based on the time-varying applied policies and shown in Figure 58.

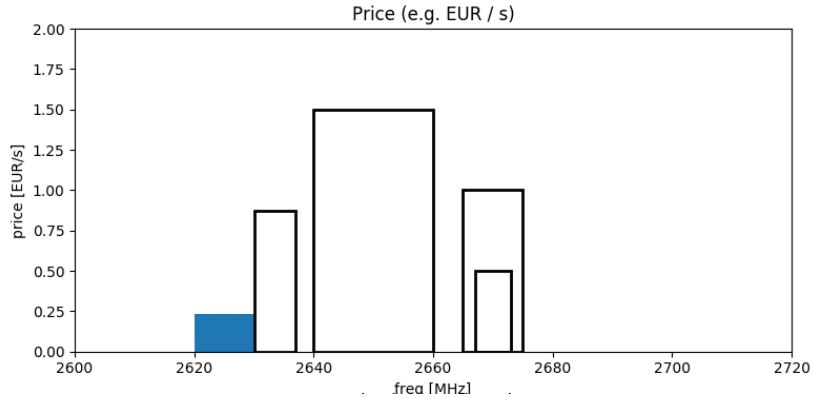


Figure 54: SMA application selection result for policy A.

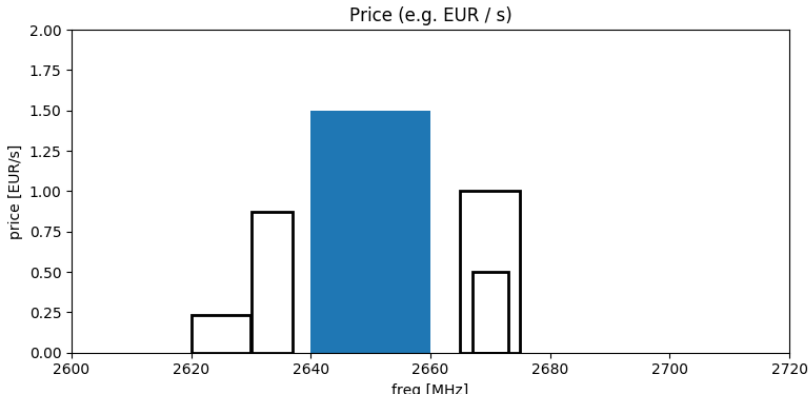


Figure 55: SMA application selection result for policy B.

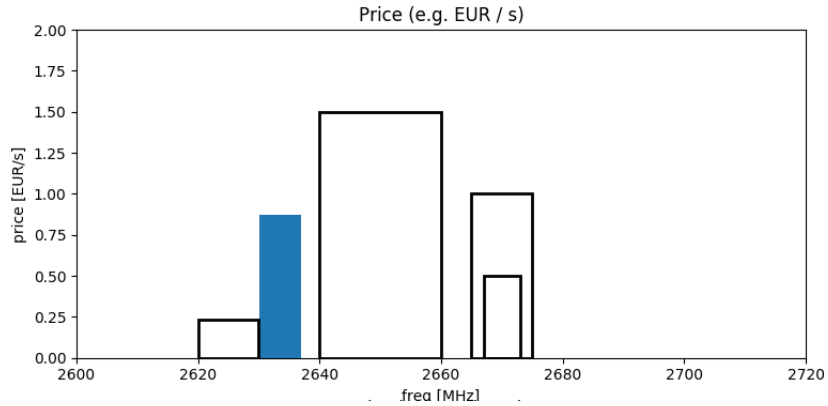


Figure 56: SMA application selection result for policy C.

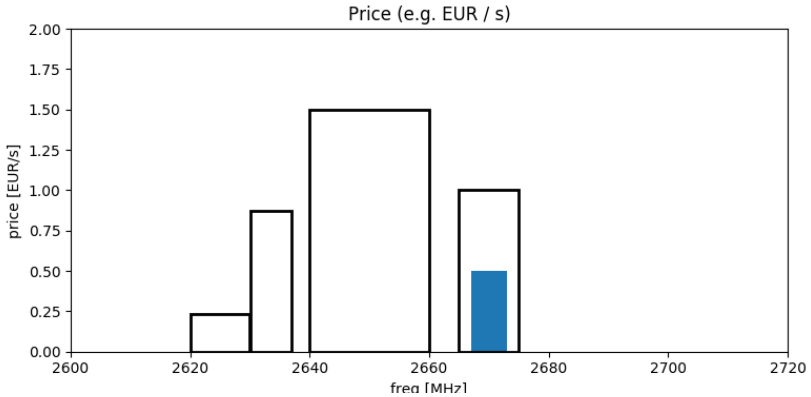


Figure 57: SMA application selection result for policy D.

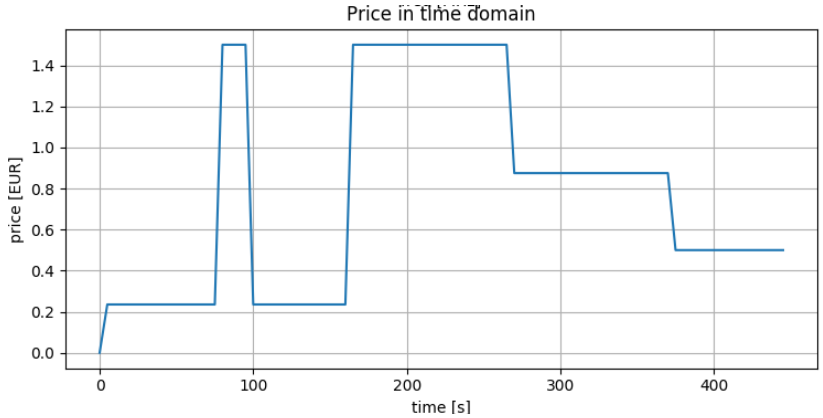


Figure 58: Normalized spectrum prices variability based on the dynamic policies.

7.1.4 User reattachment delay

In the next experiment, we investigate the impact on the UE side when restarting the BS for changing the applied operating frequency. Specifically, we change the applied spectrum from 2.68GHz (band 7) with 10MHz bandwidth to 2.62GHz (band 7) with 5MHz bandwidth. In Figure 59, we provide the snapshots of these two cases in terms of the spectrum measurement and the UE screenshot information. These snapshots clearly differentiate these two cases.

Dl_freq: 2680 Ul_freq_offset: -120 Bandwidth : 10 max_tx_power: 0 frame_type: FDD	Dl_freq: 2620 Ul_freq_offset: -120 Bandwidth : 5 max_tx_power: 0 frame_type: FDD

Figure 59: Spectrum measurement results and screenshot of the two applied spectrums (left: 10MHz bandwidth; right: 5MHz bandwidth).

Then, we measure the overall user attachment delay to the phantom cell due to the change of radio spectrum. Such delay includes several components: (a) BS restart time, (b) the cell re-selection procedure performed at the UE side to connect to the new BS and (c) cell camping delay initiated by the UE to the restarted BS. The 1st and 3rd component delay are platform dependent and is less relevant to the UE side, whereas the 2nd one will highly depend on the number of supported band (cf. 3GPP TS 36.101) and the UE category (cf. 3GPP TS 36.306) of the commercial-of-the-shelf (COTS) UE. In this regard, we conduct the experiment on two different COTS UEs, one being the Samsung Galaxy S6 (UE category 6, supports 13 LTE bands) and the other is Samsung Galaxy S5 (UE category 4, supports 6 LTE bands). The results are provided in Figure 60 and we can see that the Samsung Galaxy S6 has a higher attachment delay when compared to Samsung Galaxy S5 in all three trials. Such phenomenon may mainly due to the number of supported band, i.e., Galaxy S6 spends more time to measure all supported bands to decide the best one for following cell camping procedure.

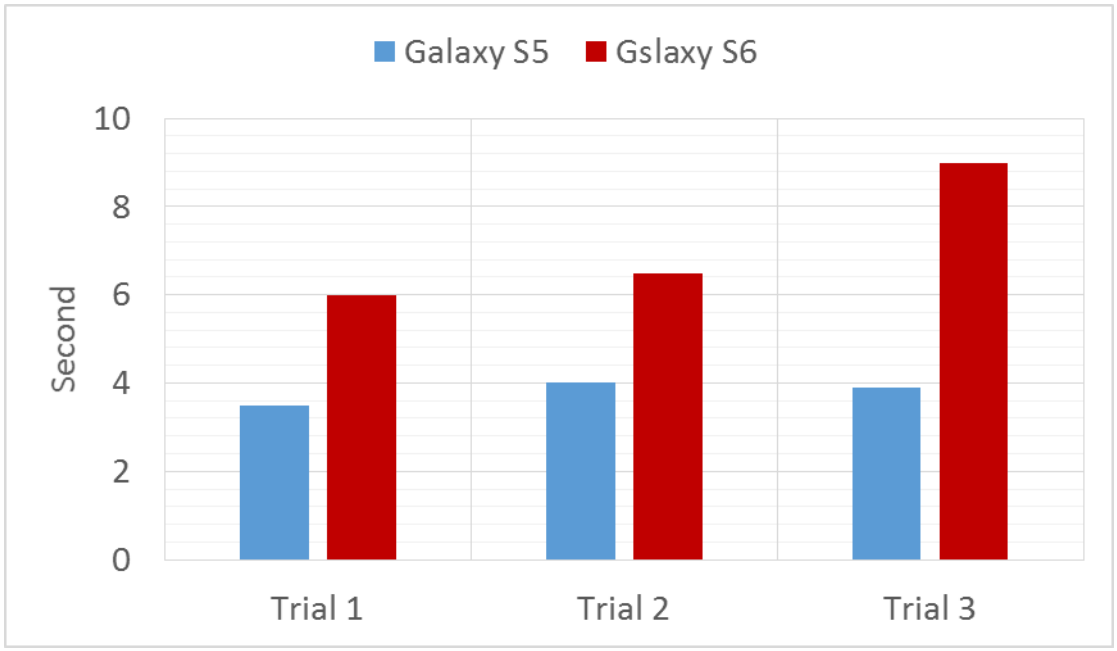


Figure 60: User attachment time measurement results.

7.1.5 UE perceived performance

In this experiment, we compare two conditions to exploit the advantage of applying the phantom cell concept. First of all, the phantom cell can be deployed at a higher frequency band and can utilize larger bandwidth to boost the user-plane performance when compared with the macro cell case. Hence, we compare two different cases: (1) the phantom cell is deployed at band 7 (2.6GHz) and can utilize a larger bandwidth (i.e., 10MHz) and (2) the macro cell deployed at band 13 (750MHz) and can only use a smaller bandwidth (i.e., 5MHz). The measured good-put and jitter at the UE side are provided in Figure 61 with the Nexus 6p COTS UE that receives the UDP traffic in the downlink direction. We can see that the phantom cell case shows a better user experience via exploiting a larger available bandwidth. It is noted that such benefit can be even boosted when the phantom cell uses a much higher carrier frequency (e.g., mmWave) with a larger bandwidth.

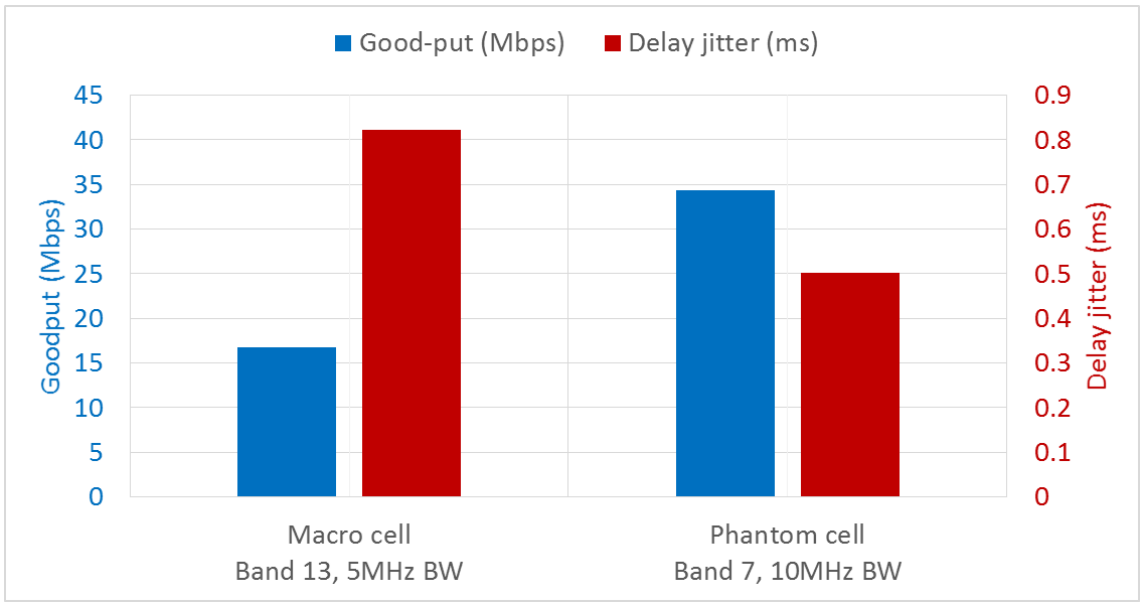


Figure 61: Measured good-put and delay jitter under two different bandwidths for macro/phantom cells.

Moreover, we consider another benefit of the phantom cell. For the macro cell, it shall support more users as it has a larger coverage area; however, the phantom cell can serve users in a limited range with regard to the macro cell thus boosting the experience of nearby users. To show such a benefit, we consider two different cases: (1) the phantom cell is deployed at band 7 (2.6GHz) with 5MHz bandwidth to serve a single UE (i.e., UE1) and (2) the macro cell is deployed at band 13 (750MHz) with 5MHz bandwidth to serve two UEs at the same time (i.e., UE1 and UE2). Here, we use two identical Nexus 6p COTS UEs for fairness comparison and the downlink UDP traffic is transported toward the UEs. In Figure 62, the results are shown for these two cases. We can see that both UE1 and UE2 in the macro cell individually will have a lower good-put when compared with the phantom cell case. Even though the sum of good-put from these two users in the macro cell is close to the one in the phantom cell case; however, a larger delay jitter is still seen for these two users. To sum up, the benefit of phantom cell can boost the user-plane performance based on the spectrum allocation results by the SMA application.

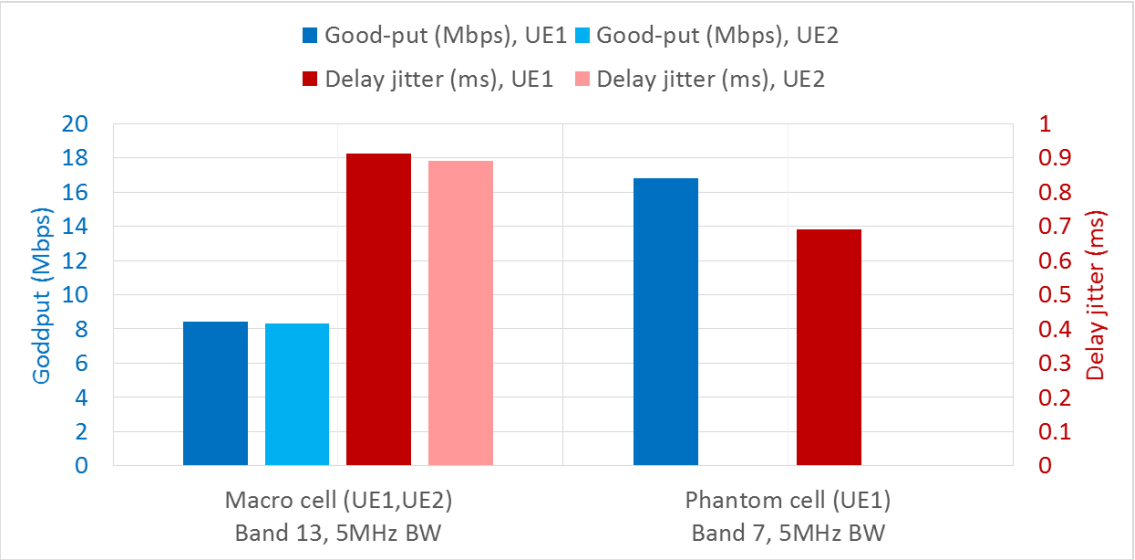


Figure 62: Measured goodput and delay jitter under different number of users for macro/phantom cells.

7.1.6 Summary

In this section, we provide the overall evaluation results of the devised SMA application. Such an application can dynamically provide the most suitable spectrum according to the inputs revealed from the COHERENT RTC as well as SDK. Moreover, we examine also the UE-perceived performance in terms of the user attachment delay and user-plane performance metric when applying different spectrum at the phantom cell. Further, based on the feasibility and advantages of the SMA application, the chaining of SMA and RRM application can be further evaluated to provide a more sophisticated control logic.

7.2 Spectrum sharing using Licensed Shared Access

7.2.1 Interference protection

The LSA controller (C1 and C2) in Figure 18 is a software implementation responsible of polling parameters from the eNodeBs, such as GPS coordinates, possible transmit power levels, current transmit channel, link status etc., and committing the new operational parameters to eNodeBs. The communication between LSA Repository and LSA Controllers contain operator priority class, used channel, tx power, I/N criteria, bandwidth, noise figure (NF), and adjacent channel leakage ratio (ACLR).

The LSA Repository manages the timing and threading of the necessary operations and assigns priority classes to the operators. The priority classes can be determined beforehand by NRA and inserted to the LSA Repository, or agreed through a mutual agreement between the operators.

The information exchange and interference protection in the trial are described below:

1. The Repository initiates each Controller to poll the parameter values from the eNodeBs every 60 seconds. The Controller chooses the best channel (eirp = *max*(**eirp**), ch = *arg max*(**eirp**)), where **eirp** is a vector containing the maximum EIRP values for that geographical location. The EIRP values are based on path losses using the Extended Hata propagation model between the interference source (eNodeB) and the protection contour of the other operators. Specifically, the EIRP value is specified as:

$$axEirp\ 10log10(BW\,[Hz])\ m \leq\ NF + I/N + PL(d) - G_t - G_r,$$

   where NF is the noise figure, $PL(d)$ is the path loss with a distance 'd' between the eNodeB transmit antenna and the protection contour of the protected mobile network, 'I/N' is the interference to noise ratio, 'G_r'/'G_t' are the transmit and receive antenna gains. The protection contour can be for example specified as the limit where the useful signal level of -80 dBm/10MHz is observed.
2. The Repository provides the available incumbent data to the Controller. Since there are no other operators using the spectrum according to the provided incumbent data, the EIRP limit is set to predetermined maximum, i.e., 47 dBm for each channel. The controller then assigns the first available frequency (2620-2630 MHz) as the transmit channel with an appropriate EIRP for the eNodeB. The Controller commits the new transmit parameters to the eNodeB.
3. Subsequently (or simultaneously), the Controller polls the values from the second eNodeB and synchronizes the incumbent data from the Repository. In this case, the protection contour of the first eNodeB is calculated and its transmit channel is taken into account in calculating the eirp list, if the priority class of the first operator is higher than that of the second operator. If the priority class of the second operator is higher than the priority class of the first operator the second Controller can assign the channel used by the lower priority operator to the higher priority operator. In this case, the second Controller bypasses all the lower priority protection contours in the calculation of the eirp list for the second eNodeB. The controller then chooses the best channel for the second eNodeB, and notifies the Repository. The Repository further informs other Controller(s) to make a new calculation of the eirp list based on the provided parameters and adjust their operational parameters.
4. Simultaneously, the Repository manages each Controller to poll the parameter values from the eNodeBs every 60 seconds, and the process begins from the start (item 1 above). If there are no parameter changes, e.g., no priority class changes or location changes, or there are no new Controllers connected to the network, the system remains stable (no commits are made to the eNodeBs), and the eNodeBs continue to transmit on their current allocated channels and powers.

7.2.2 Processing Flow

The signal flow between the elements that takes place during the PoC is described in Figure 63. The LSA controller has a repository that keeps necessary elements with regards to the priority that has to provide to the 2 eNodeBs. The LSA controller communicates with the management system of the two eNodeBs, for polling and collecting technical parameters of the eNodeBs that belong conceptually to two different operators. The LSA controller controls the channels between the eNodeBs of the two operators and gives priority of accessing an eNodeB to the primary operator.

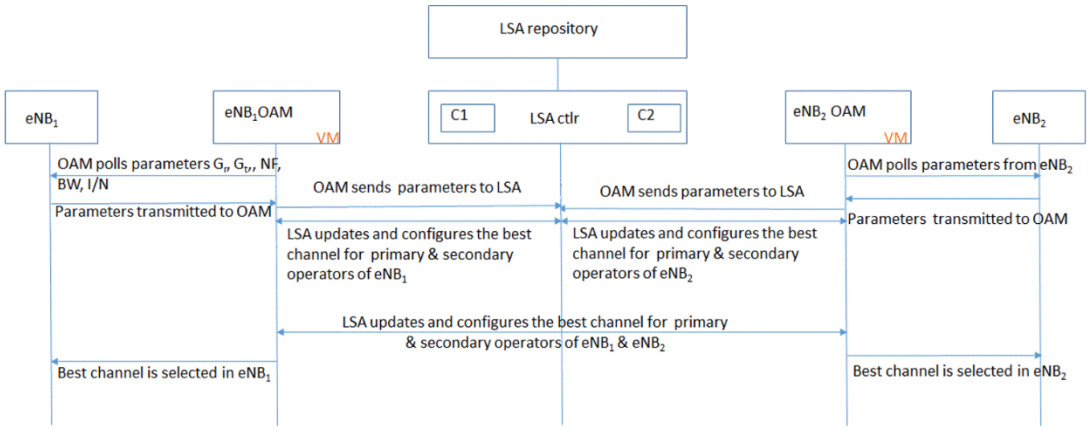


Figure 63: Signal flow when the LSA controller provides priority to one of the two eNodeBs Demonstration

The trial begins by setting lower priority private LTE eNodeB on center frequency 2625 MHz with 10 MHz BW with 24 dBm transmit power. The spectrum is monitored with Tektronix spectrum analyzer connected to a laptop. In the following figures, 2600-2700 MHz band is shown (f_c 2650 MHz, span 100 MHz, rBW 500 KHz). The spectrum mask of the private LTE is shown on 2620-2630 MHz. On the right hand side of the spectrum at around 2670-2690 MHz band is unknown traffic not relevant for this trial. The higher priority mobile network operator comes in the system and chooses the transmission channel currently operated by the private LTE. The private LTE is forced to choose another channel, and due to the adjacent channel leakage power the second best channel is at 2645 MHz. The above spectrum phases are shown in Figure 64.

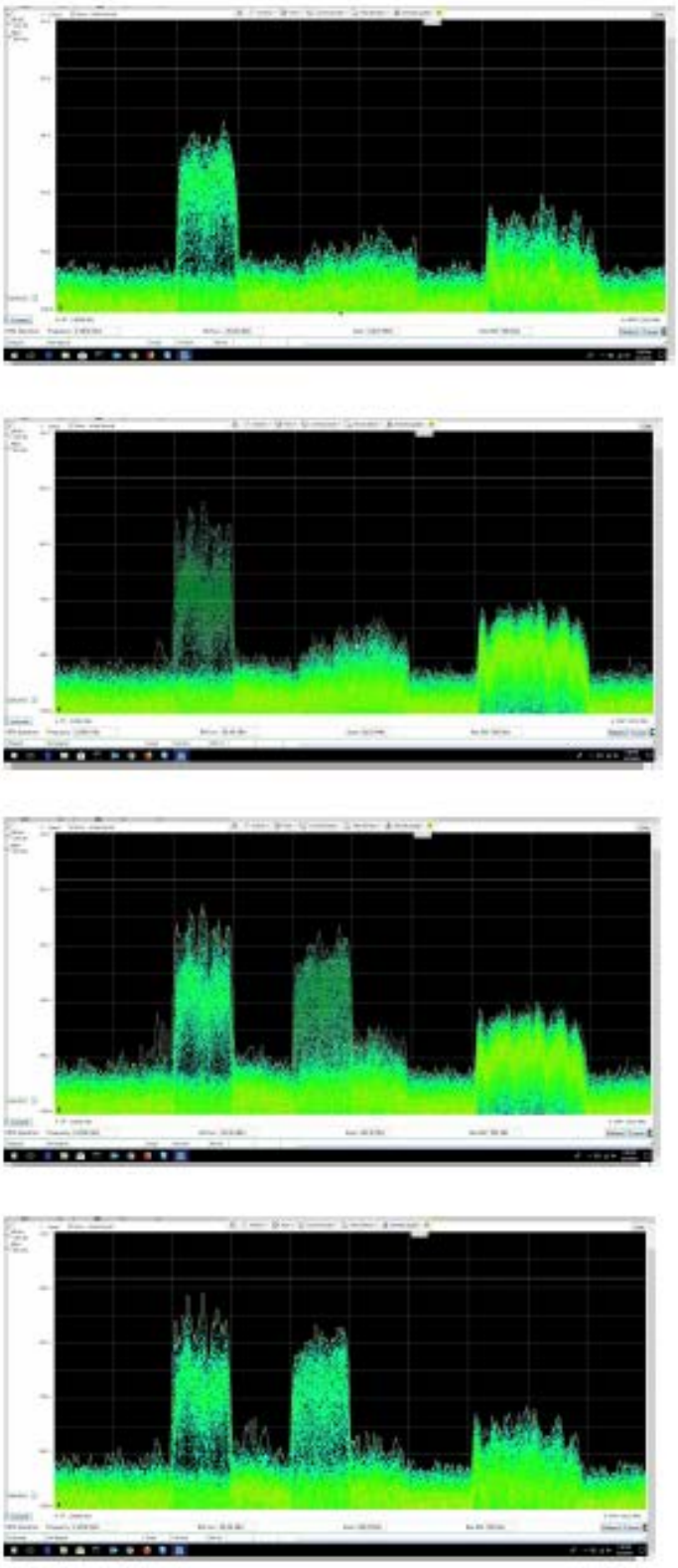


Figure 64: Spectrum sharing between primary and secondary users

7.2.3 Summary

This section shows the spectrum sharing trial results conducted in the real environment. In this trial, the private LTE eNodeB shares the 2.6GHz spectrum with the incumbent mobile operator. The LSA Controller observes the spectrum usage information and dynamically reallocate the spectrum to the low priority spectrum users to avoid interference to the incumbent mobile operator.

8. Distributed Antenna System – Evaluation Results

This work focusses on the measurements performed to validate our work in [7] in order to attain a reliable throughput and performance by RB reuse at different UE locations of a DAS system by efficient pairing of UEs with RRHs based on the probability of coverage metric. As the probability of coverage metric, given in [17], depends on the location of UEs and RRHs, transmitted powers and targeted SINR, the measurements are performed varying each of these parameters to understand their effect on the performance of the DAS system.

8.1 Measurement Setup

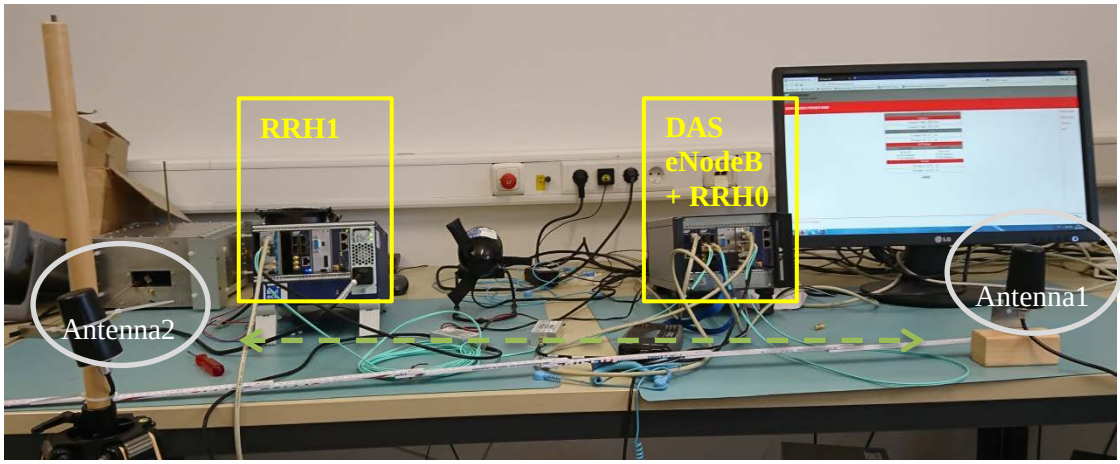


Figure 65: DAS eNodeB with 2 RRHs

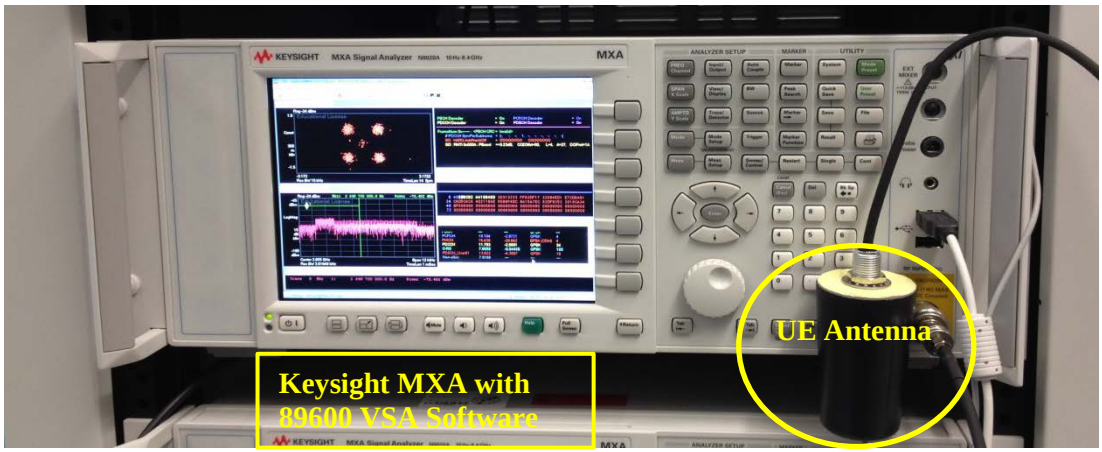


Figure 66: Keysight MXA Signal Analyzer

Regarding the trials with real-time LTE based DAS testbed Figure 65 & Figure 66, we have made use of Keysight MXA signal analyzer (Figure 66) running 89600 VSA software for LTE PHY layer analysis, which provides in-depth analysis LTE system performance of the decoded channels in terms of EVM.

Figure 65 shows the testbed which is a microTCA chassis that contains a Dual TI C6670 DSP board and RRH board apart from the microTCA controller. The DSP and RRH are connected via CPRI using optical cable. The current implementation is a selection transmission DAS technique i.e., to pair UEs based on distance of the UE. In this approach the nearest user to the RRH would be served by that RRH. To validate the pairing, support of overlapping RB allocation for UEs is implemented in the PHY processing, in contrast to standard LTE non-overlapping RB allocation. For the measurement, LTE single stream (SISO) transmission is configured for both the UEs. Different Radio network temporary identifiers (RNTIs) are configured for the UEs. On the UE receiver side, a Keysight MXA signal analyser is used. Keysight VSA software running on MXA signal analyser

(Figure 66) is capable of real-time capturing and decoding of time domain LTE signals over the air. Figure 66 shows the decoded IQ constellation along with the allocated RBs provided by the VSA software. The measurement setup can decode the DLSCH channels transmitted through the implemented DAS Engine, which verifies the real-time LTE based DAS testbed implementation.

Figure 67 shows the measurement setup in the lab (indoor environment), where the measurements are performed, with RRH0 and RRH1 connected with antennas via long CPRI cables so that they can be moved in different positions in order to mimic the real DAS and to understand the impact of RRH and UE placements on the EVM measurements. Figure 67 shows a closer view of the RRH0, RRH1 and UE antenna connected to MXA. RRH1 antenna is placed on a tripod to facilitate easy placement of the RRH1 in different positions. The power values of RRH0 and RRH1 can be easily modified using graphical user interface (GUI) of the RRHs provided by Fraunhofer, Heinrich Hertz Institute, Berlin (HHI). One can enter the required transmit power values in the GUI to control the RRH. In practice it will be controlled by the DAS eNodeB in conjunction with C3/RTC.

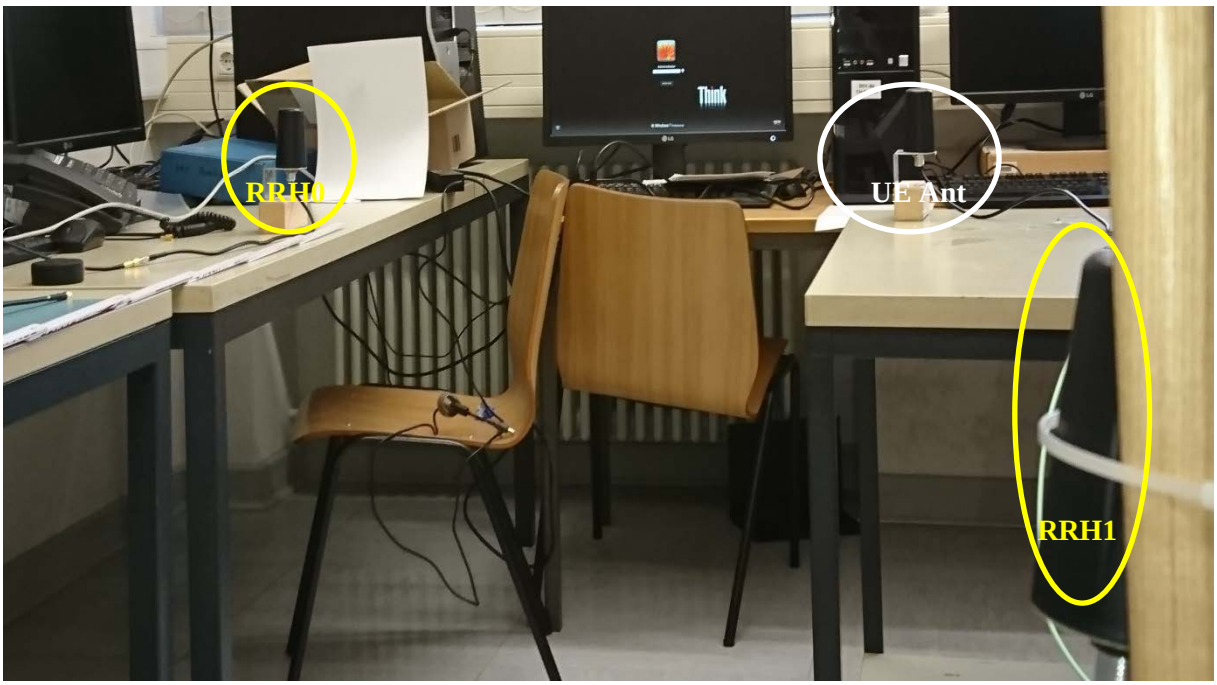


Figure 67: Measurement Setup

Figure 68 is a pictorial description of different rays emitted by both the RRHs targeting different UEs. Omni-directional antennas produce unwanted transmission or interference, because of that the non-serving RRH1 is also received at the UE. In our experimentation we have measured primarily the error vector magnitude (EVM) for monitoring the performance at the UE. 3GPP LTE [26] defines the limits of error vector magnitude (EVM) Table 10, the error vector magnitude in terms of percentage of RMS for different modulations.

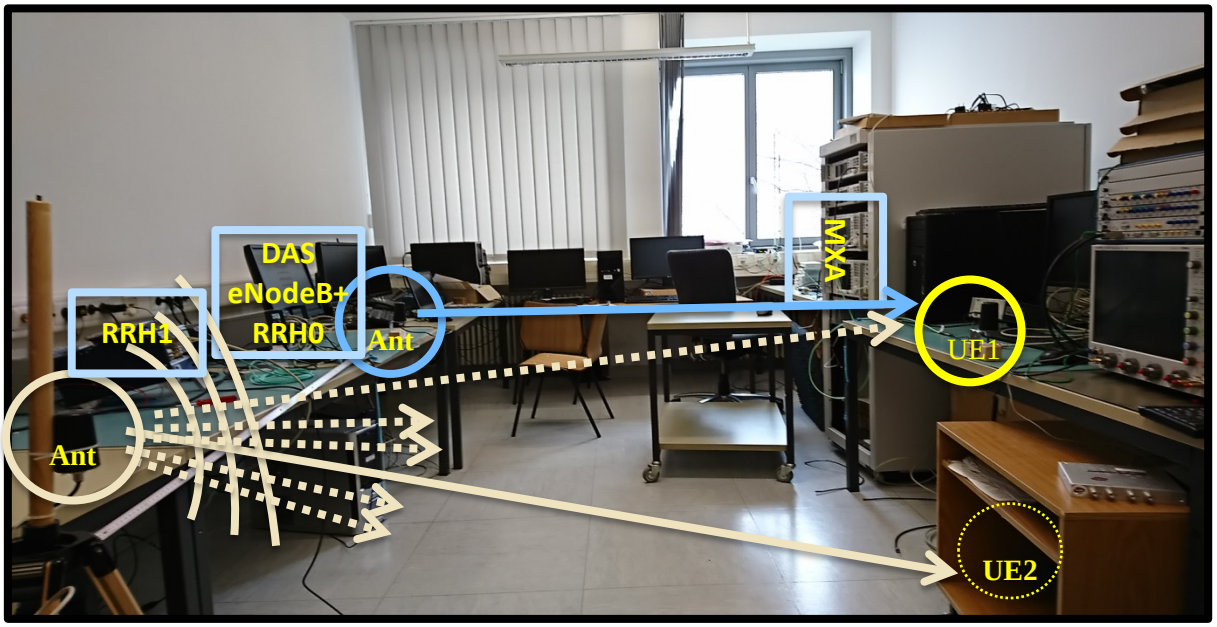


Figure 68: Pictorial representation of rays from the RRHs

The EVM is defined in LTE [26] as a measure of the difference between the ideal symbols and the measured symbols after the equalization. This difference is called the error vector. The EVM result is defined as the square root of the ratio of the mean error vector power to the mean reference power expressed in percent.

Modulation scheme of PDSCH	Required EVM [%]
QPSK	18.5%
16QAM	13.5%
64QAM	9%

Table 10: 3GPP LTE EVM requirements[26]

The details about measurement setup are given in Table 11. As explained before, EVM is the observed metric at UE1 location, which provides the quality of the received signal at the UE1 location while RRH1 transmits a fully overlapping RBs scheduled for UE2. The parameters varied in different scenarios are basically the separation distance between RRHs, transmitted power and modulation scheme. The effects of these parameters are observed through the observed EVM.

Scenario	Details
No. of UEs	2
No. of RRHs	2
No. of Antennas per RRH	1 (SISO transmission)
Measurement equipment	Agilent MXA
DL/UL	Downlink
Measurement metric	EVM on VSA software
Parameters	Separation distance between RRHs, Transmitted power and Scheduling of RBs and MCS

Table 11: Measurement setup

Scenario	Details

No. of RRHs	2
No. of Antennas per RRH	1 (SISO transmission)
Bandwidth	10 MHz
DL Resource Blocks scheduled for UE1	50 RBs / All RBs
DL Resource Blocks scheduled for UE2	50 RBs / All RBs
MCS used	9(QPSK),11(16 QAM)
Transmitted power RRH0	-8 dBm
Transmitted power RRH1	-8 dBm
RNTI of UE1	10
RNTI of UE1	20
DL Frequency	2.655 GHz
FSPL at 1 meter	41 dB
FSPL at 2 meter	47 dB
FSPL at 3 meter	50 dB

Table 12: LTE PHY configuration – Common Parameters

Table 12 provides the common parameters of the LTE PHY configuration deployed in order to perform the measurements. The description below describes the measurement scenarios investigated at our lab and the corresponding measurement results.

Scenario 1 and Scenario 2 show a typical case where the UEs are quite close to the distributed RRHs. The impact of performance by varying RRH placements and RRH power is studied. Scenario 3 and Scenario 4 depict the case where an UE is positioned equidistant to both of RRHs, which is the worst case scenario for DAS. Here, the effect of varying modulation schemes is studied as well.

8.2 Scenario 1

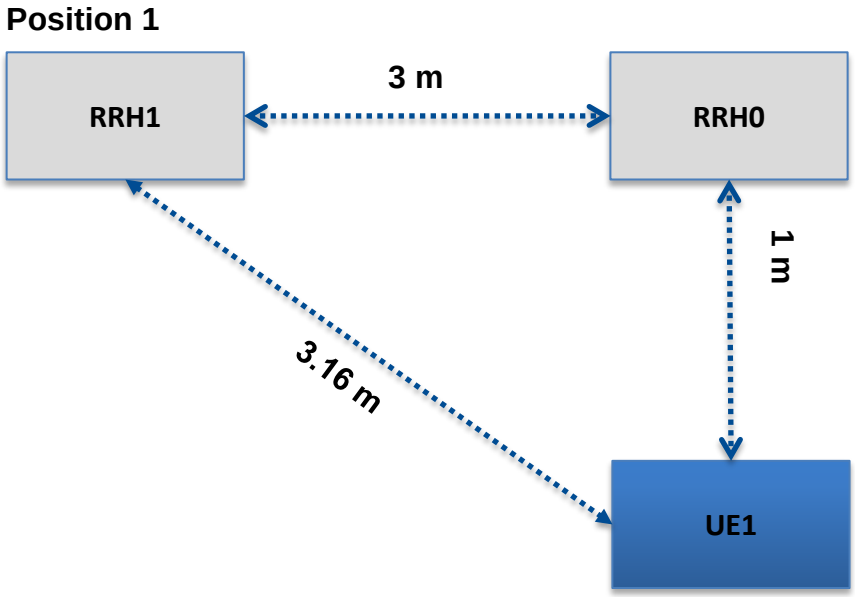


Figure 69: Scenario 1: RRHS seperated by 3 m, RRH1 at Position 1

In the scenario 1 of Figure 69, both RRHs are separated by 3m and the UE1, which is the Keysight's MXA, is at a distance of 1m directly in front of RRH0. Table 13 provides the observation results. Figure 70 & Figure 71 show some snapshots of VSA software measurements of DL decode info, EVM and constellation diagram for different SIR values. DL decode info shows the PDCCH channel decoded information, i.e., 50 RB allocation, RNTI value of the UE, CRC pass information. EVM and constellation diagrams captured on the VSA software are provided in these figures for some of the trails based on signal-to-interference-ratios (SIR). From the constellation diagrams we can see the effect of interference on the QPSK constellation and how it got improves with increase in SIR. From the measurement results, the minimum required signal to interference ratio (SIR) to meet the LTE requirements for QPSK is around 23 dB.

RRHs seperation distance (m)	RX Power at UE1 due to RRH0 (dBm)	RRH1 Atten. (dB)	RX Power at UE1 due to RRH1 (dBm)	Signal-Interference - Ratio (dB)	CRC/EVM Pass/Fail	EVM measured (in %)	Min. LTE EVM requirement for QPSK (in %)
3	-50	0	-64	14	Pass/Fail	49.3	18.6
3	-50	3	-67	17	Pass/Fail	26.8	18.6
3	-50	6	-70	20	Pass/Fail	20.4	18.6
3	-50	9	-73	23	Pass/Pass	14.5	18.6
3	-50	12	-76	26	Pass/Pass	11.7	18.6
3	-50	15	-79	29	Pass/Pass	8.4	18.6
3	-50	18	-82	32	Pass/Pass	5.9	18.6
3	-50	21	-85	35	Pass/Pass	3.9	18.6
3	-50	24	-88	38	Pass/Pass	3.4	18.6
3	-50	27	-91	41	Pass/Pass	1.6	18.6

Table 13: EVM Measurements for Scenario 1: QPSK modulation

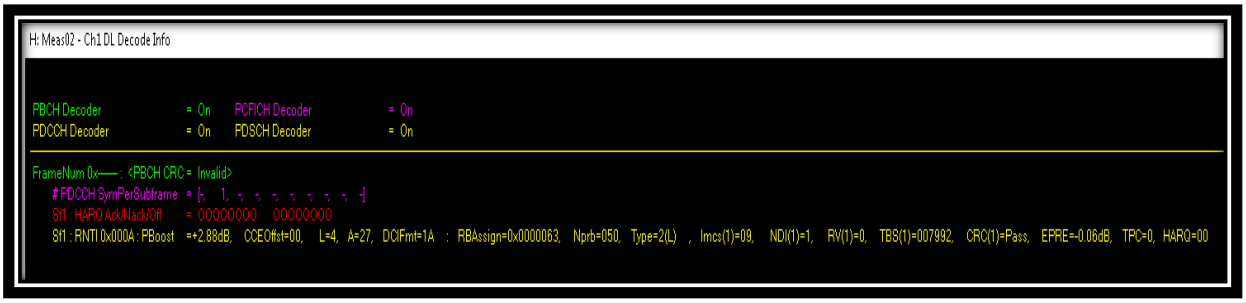


Figure 70: Scenario 1: DL Decode info trace from VSA software

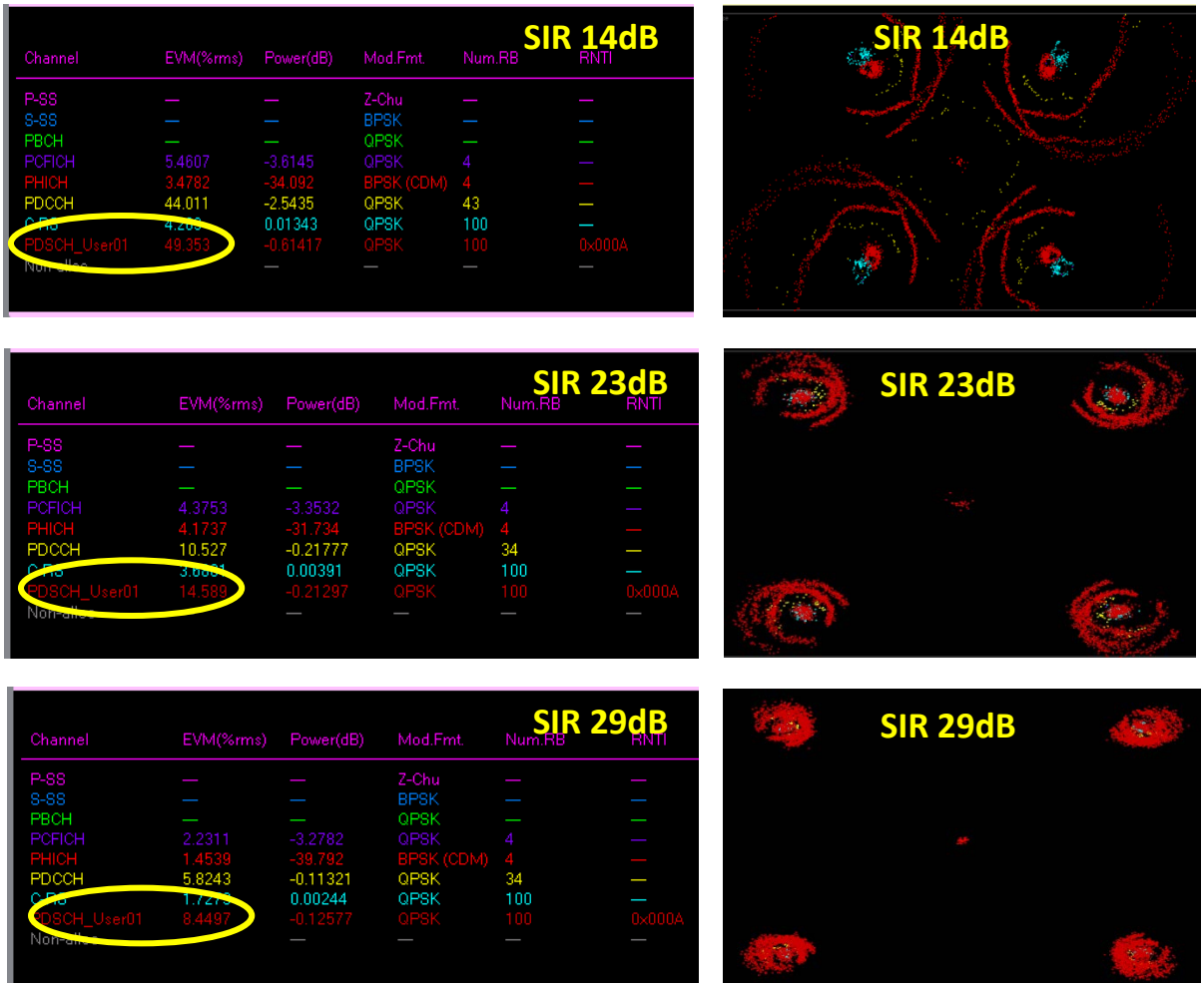


Figure 71: Scenario 1: EVM and constellation diagram from VSA software

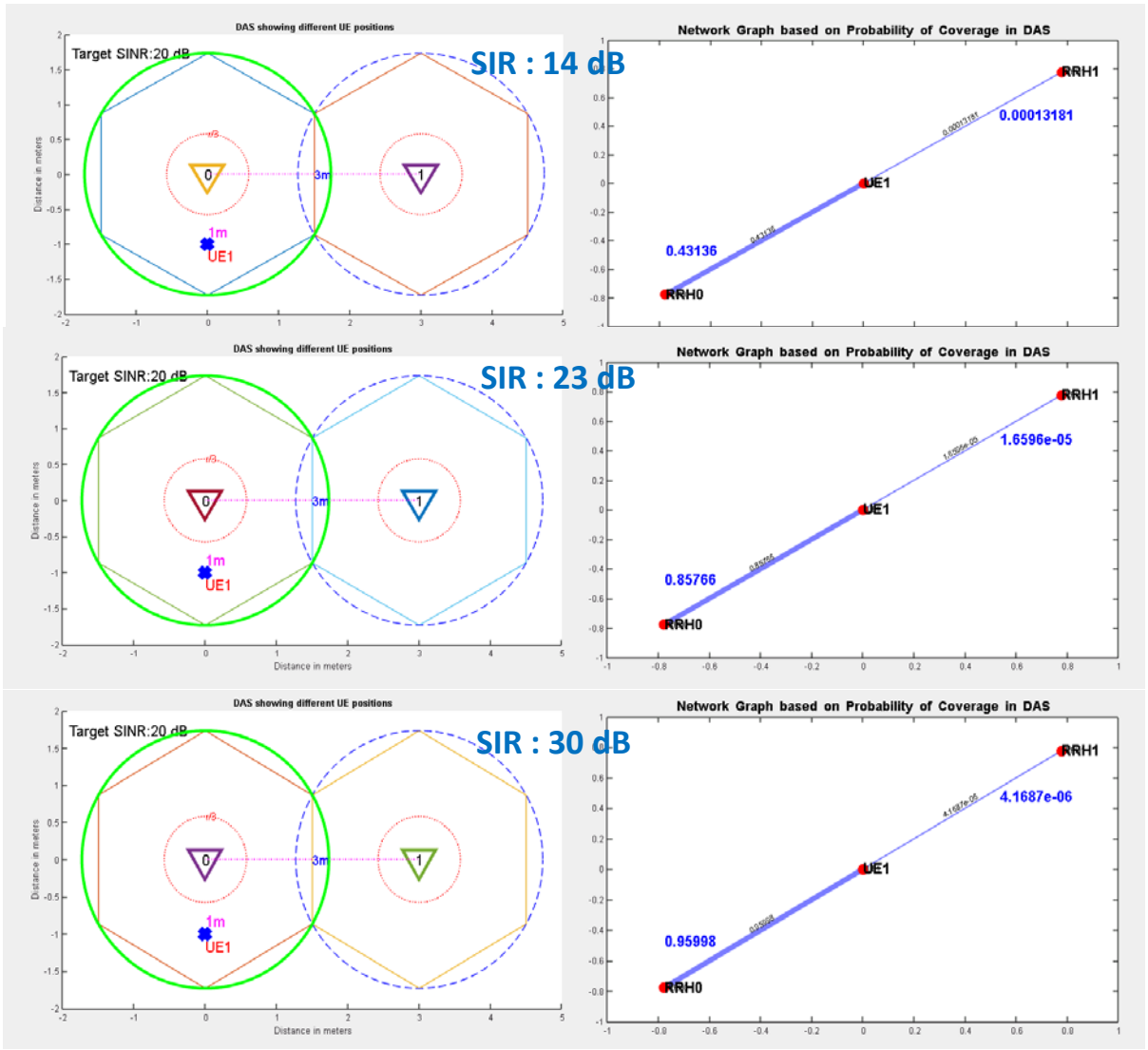


Figure 72: Scenario 1: Probability of coverage based network graph for DAS

Figure 72 shows the network graph generated using the probability of coverage metric implemented on MATLAB. We can see from the theoretical simulation results that for the SIR of 23 dB the probability of coverage at the UE1 location is around 0.85, which is a very high value for making a decision for pairing of UE with RRH0. For SIR of 30 dB the probability of coverage is around 0.95. These results map linearly to the measurement results in **Table 13**.

8.3 Scenario 2

In the scenario 2 shown in Figure 73, RRH1 is moved to position 2, which is 3m directly opposite to UE1. Hence the separation distance between RRHs is around 3.16m. Table 14 provides the observation results. Figure 74 & Figure 75 show some snapshots of VSA software measurements of DL decode info with 50 RBs allocated, TB Size of 7992 bits, MCS of 9, EVM and constellation diagram for different SIR values. From the measurement results, the minimum required signal to interference ratio (SIR) at the UE location to meet the LTE requirements for QPSK in this scenario is around 16 dB.

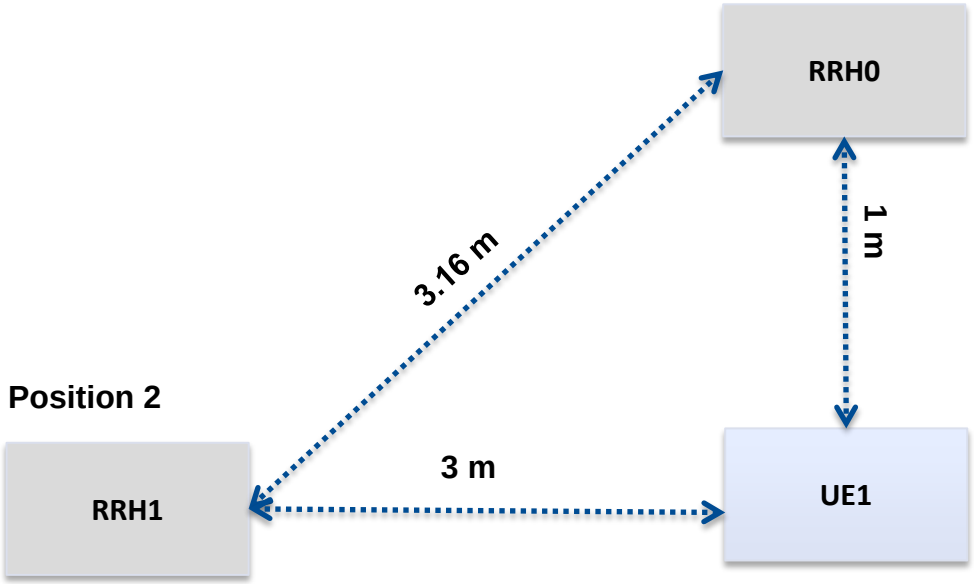


Figure 73: Scenario 2: RRHS seperated by 3.16 m, RRH1 at Position 2

RRHs seperation distance (m)	RX Power at UE1 due to RRH0 (dBm)	RRH1 Atten. (dB)	RX Power at UE1 due to RRH1 (dBm)	Signal-Interference - Ratio (dB)	CRC/EVM Pass/Fail	EVM measured (in %)	Min. LTE EVM requirement for QPSK (in %)
3.16	-50	0	-60	10	Pass/Fail	42.6	18.6
3.16	-50	3	-63	13	Pass/Fail	23.5	18.6
3.16	-50	6	-66	16	Pass/Pass	17.8	18.6
3.16	-50	9	-69	19	Pass/Pass	12.8	18.6
3.16	-50	12	-72	22	Pass/Pass	10.1	18.6
3.16	-50	15	-75	25	Pass/Pass	7.7	18.6
3.16	-50	18	-78	28	Pass/Pass	3.5	18.6
3.16	-50	21	-81	31	Pass/Pass	2.3	18.6

Table 14: EVM Measurements for Scenario 2: QPSK modulation

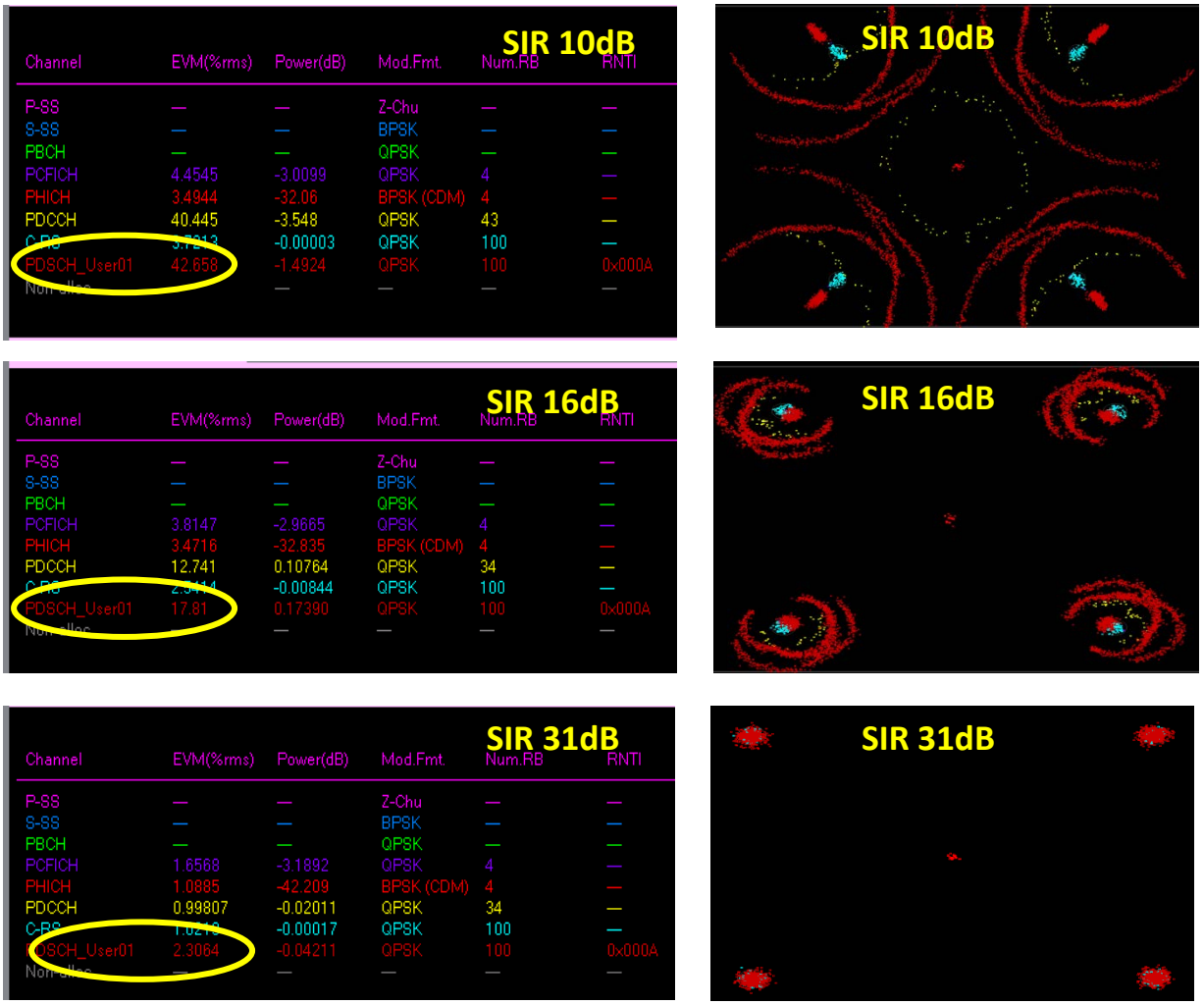


Figure 74: Scenario 2: EVM and constellation diagram from VSA software

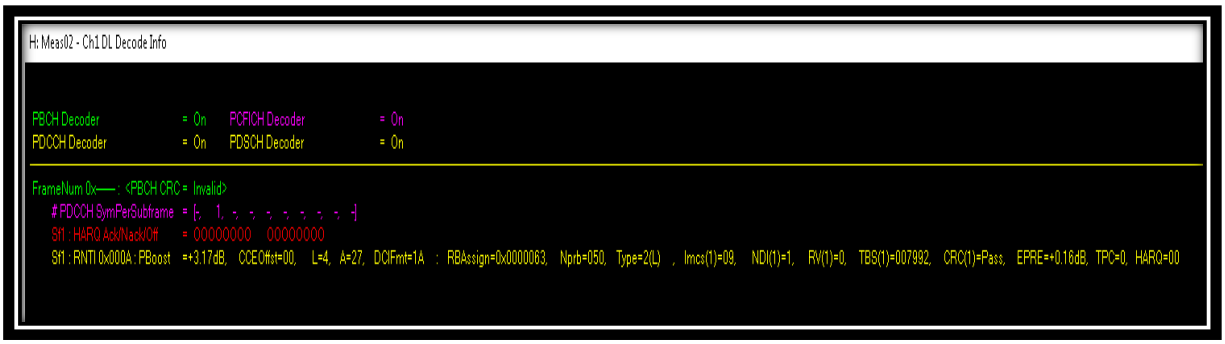


Figure 75: Scenario 2: DL decode info from VSA software

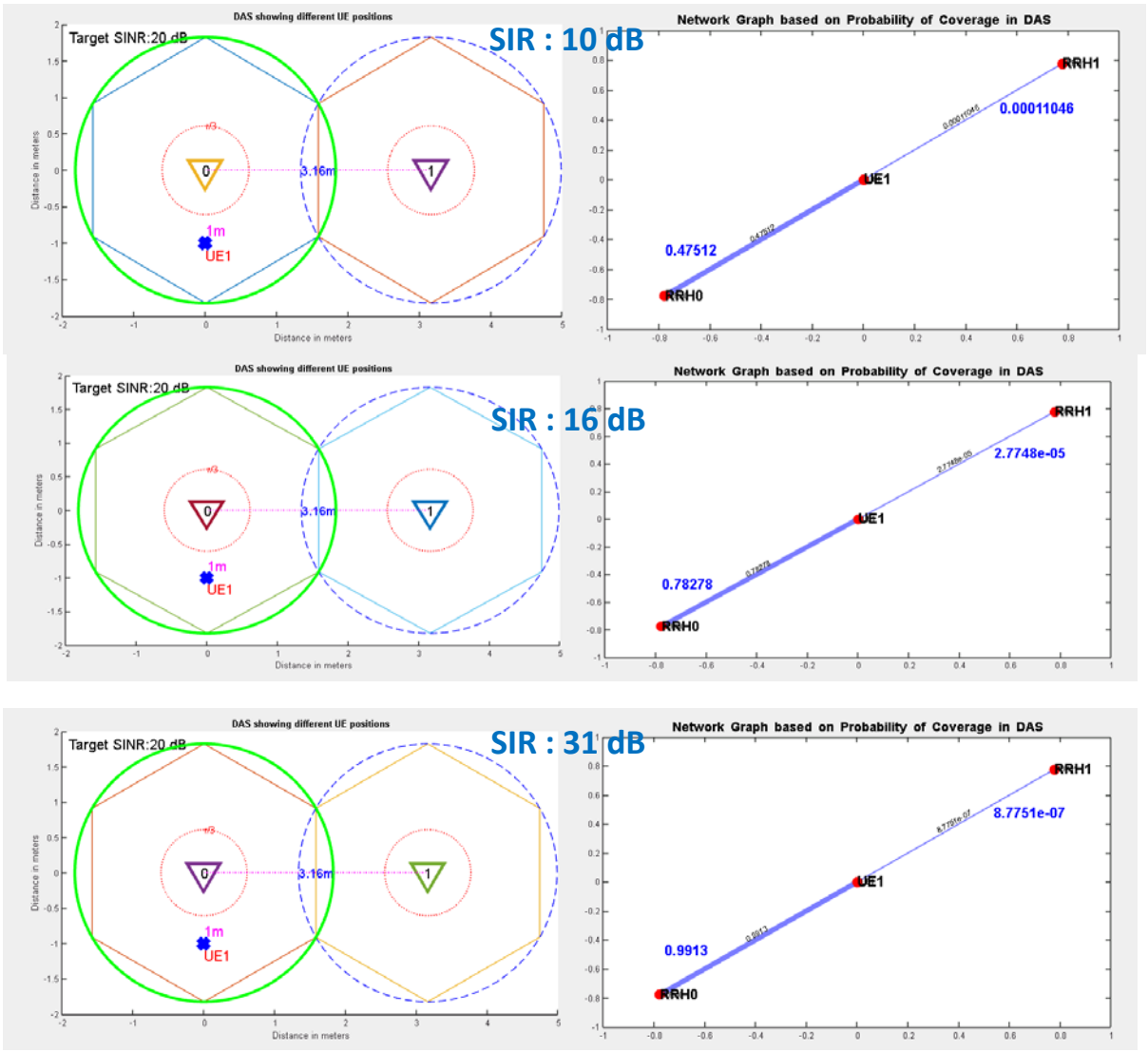


Figure 76: Scenario 2: Probability of coverage based network graph for DAS

Figure 76 shows the network graph generated using the probability of coverage metric implemented on MATLAB. We can see from this theoretical simulation results that for the SIR of 16 dB the probability of coverage at the UE1 location is around 0.78, which is high enough for making a decision for pairing of UE with RRH0. For SIR of 31 dB the probability of coverage is around 0.99. The results map almost linearly to the measurement results in Table 14.

8.4 Scenario 3

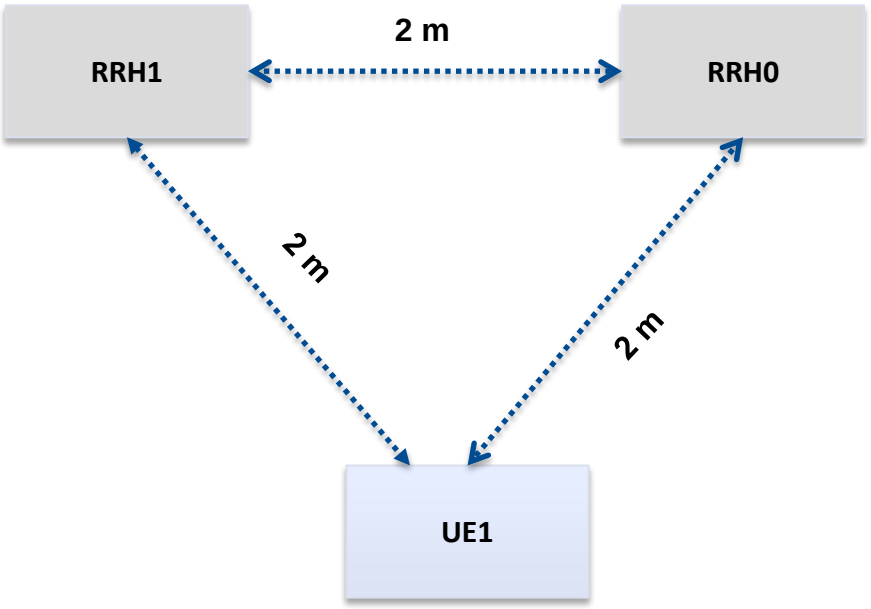


Figure 77: Scenario 3(QPSK): RRHS seperated by 2 m and UE is 2 m equi-distant to RRHs

In the scenario 3 shown in Figure 77, RRH1 is moved to position 1, which is 2m directly opposite to RRH0. Hence the separation distance between RRHs is around 3m. UE is equidistant to both the RRHs. It is kind of a worst case scenario. **Table 15** provides the observation results. Figure 78 & Figure 79 show some snapshots of VSA software measurements of DL decode info, EVM and constellation diagram for different SIR values. We can see that at the SIR starts with 0 dB and then it varied till 30 dB by attenuating the signal of RRH1. The minimum required signal to interference ratio (SIR) to meet the LTE requirements for QPSK in this scenario is around 18 dB.

RRHs seperation distance (m)	RX Power at UE1 due to RRH0 (dBm)	RRH1 Atten. (dB)	RX Power at UE1 due to RRH1 (dBm)	Signal-Interference - Ratio (dB)	CRC/EVM Pass/Fail	EVM measured (in %)	Min. LTE EVM requirement for QPSK (in %)
2	-59	0	-59	0	Fail/Fail	NA	18.6
2	-59	3	-62	3	Fail/Fail	NA	18.6
2	-59	6	-65	6	Fail/Fail	NA	18.6
2	-59	9	-68	9	Pass/Fail	41.7	18.6
2	-59	12	-71	12	Pass/Fail	31.8	18.6
2	-59	15	-74	15	Pass/Fail	21.3	18.6
2	-59	18	-77	18	Pass/Pass	16	18.6
2	-59	21	-80	21	Pass/Pass	11.7	18.6
2	-59	24	-83	24	Pass/Pass	8.1	18.6
2	-59	27	-86	27	Pass/Pass	6.5	18.6
2	-59	30	-89	30	Pass/Pass	5.4	18.6

Table 15: EVM Measurements for Scenario 3: QPSK modulation

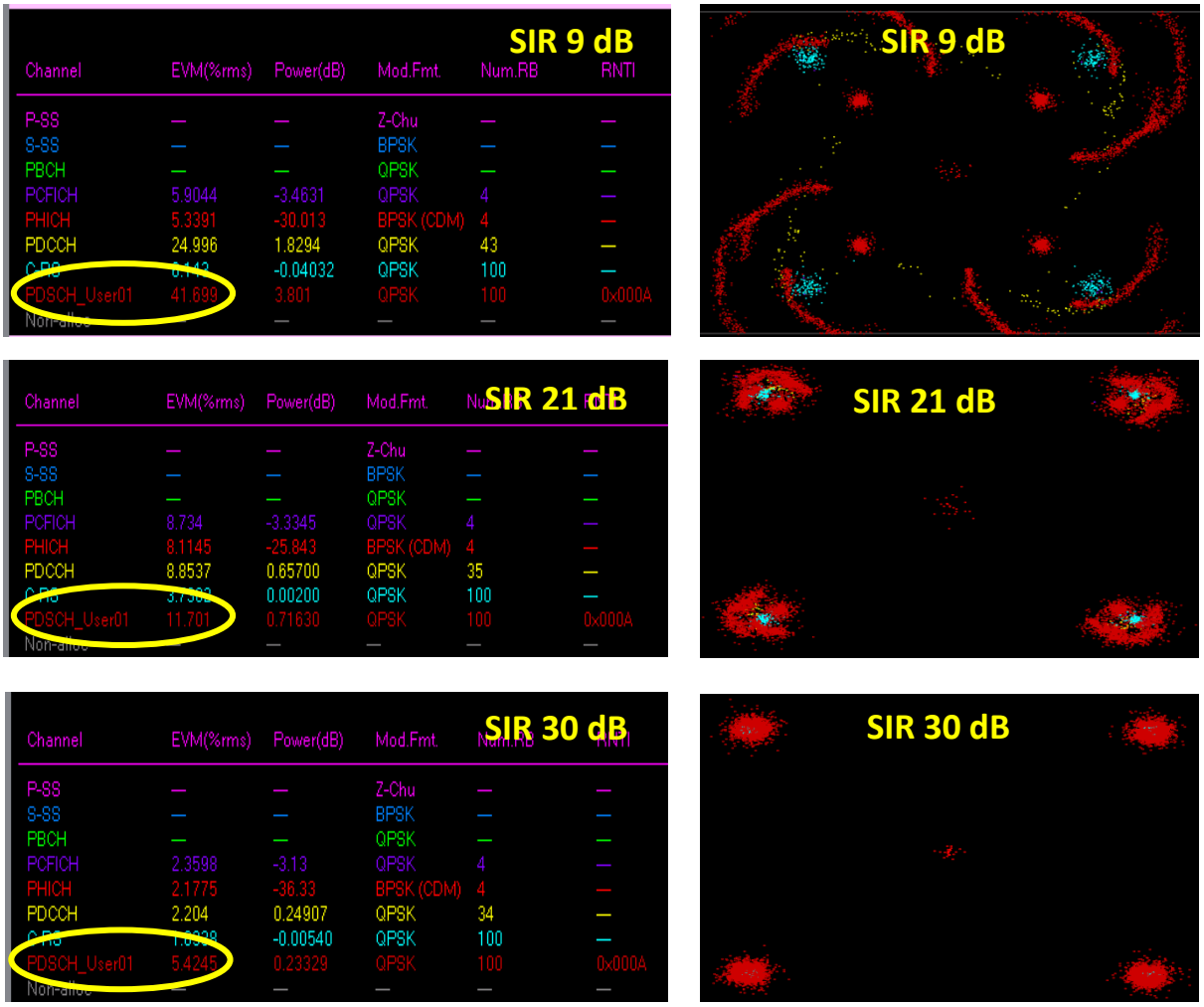


Figure 78: Scenario 3: EVM and constellation diagram from VSA software

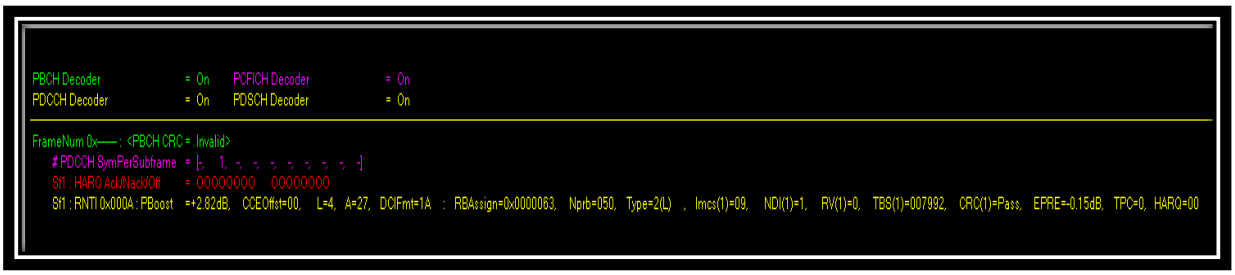


Figure 79: Scenario 3: DL Decode Info from VSA software

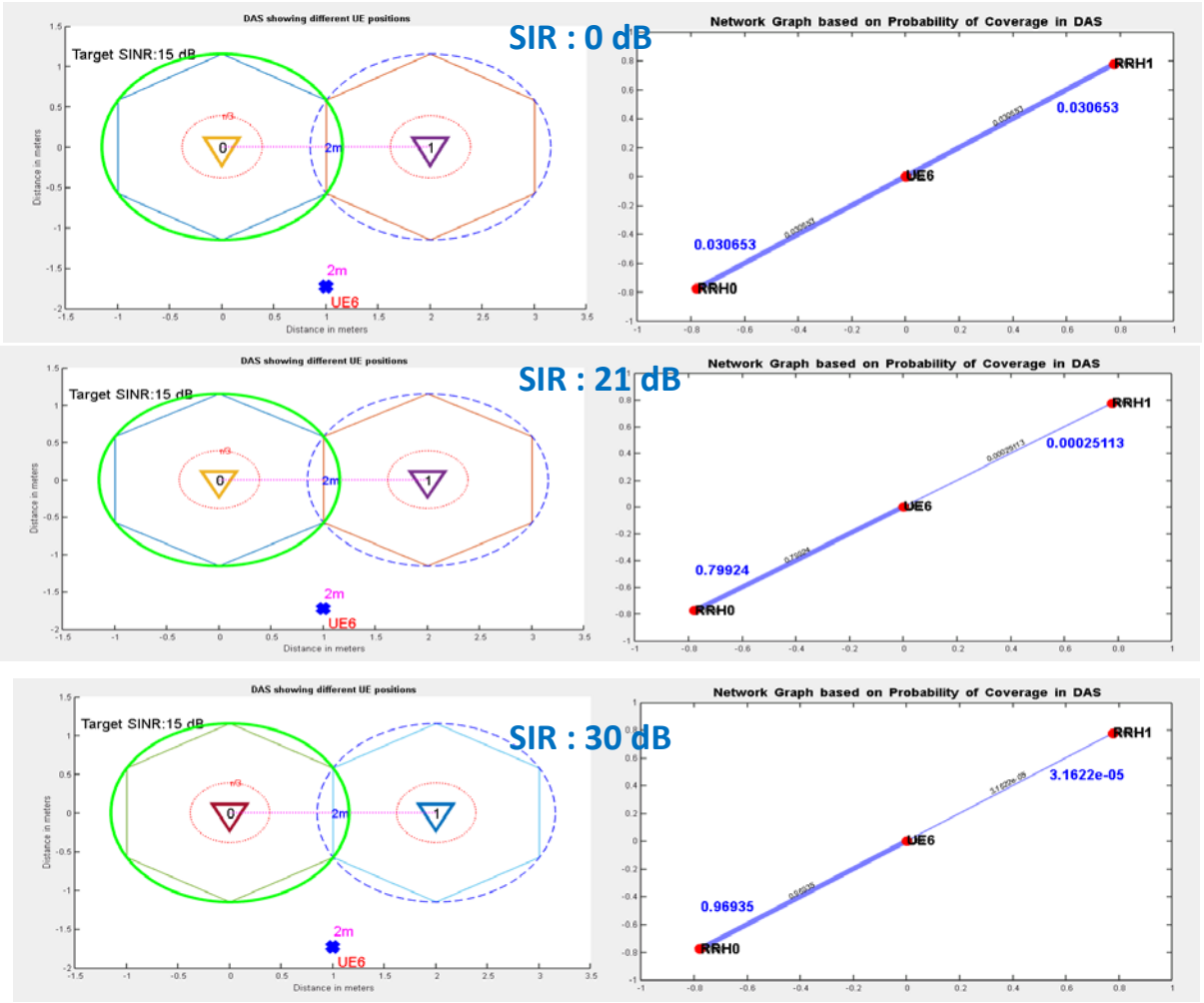


Figure 80: Scenario 3&4: Probability of coverage based network graph for DAS

Figure 80 shows the network graph generated using the probability of coverage metric implemented on MATLAB. The necessary parameters for the MATLAB based GUI for the scenarios 3 and 4 were given to generate this output figure. We can see from these theoretical simulation results that for the SIR of 21 dB the probability of coverage at the UE location is around 0.8, which is high enough for making a decision for the pairing of UE with RRH0. For SIR of 30 dB the probability of coverage is around 0.97. The results map almost linearly to the measurement results in **Table 15** & **Table 16**.The probability of coverage metric is not directly dependent on the modulation technique used, hence the results are same for scenario 3 and scenario 4. Nevertheless, the probability of coverage is dependent indirectly on the modulation scheme through the targeted SINR, which in turn is essential to achieve a decent bit error rate (BER) for the modulation scheme.

8.5 Scenario 4

In the scenario 4 shown in Figure 81, RRH1 is moved to position 1, which is 2m directly opposite to RRH0. Hence the separation distance between RRHs is around 3m. UE is equidistant to both the RRHs. It is kind of a worst case scenario. **Table 16** provides the observation results. Figure 82 & Figure 83 show some snapshots of VSA software measurements of DL decode info with MCS 10 showing 16QAM modulation, EVM and constellation diagram for different SIR values. We can see that at the SIR starts with 0 dB and then it varied till 30 dB by attenuating the signal of RRH1. This scenario is similar to that of scenario 3, but with 16QAM. The minimum required signal to interference ratio (SIR) to meet the LTE requirements for 16QAM in this scenario is around 21 dB.

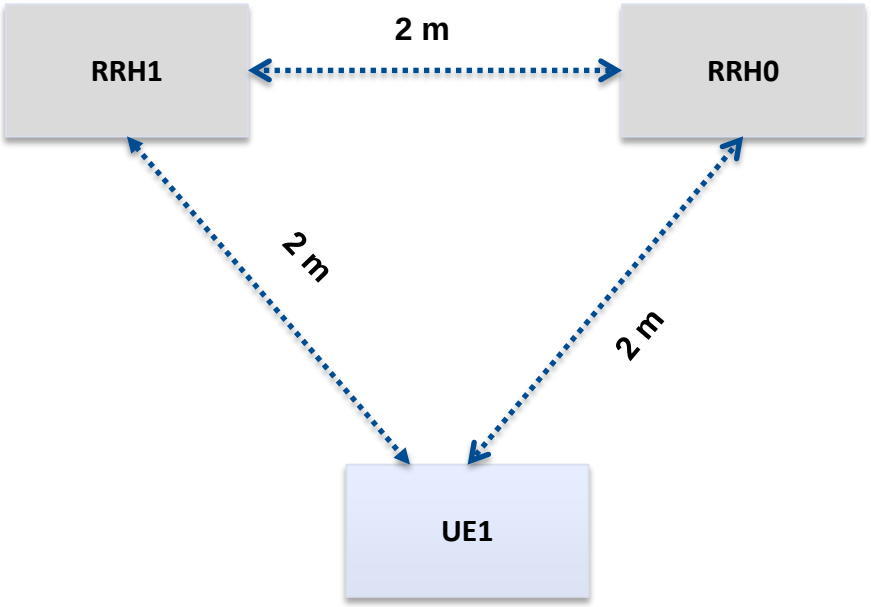


Figure 81: Scenario 4(16QAM): RRHS seperated by 2 m and UE is 2 m equi-distant to RRHs

RRHs seperation distance (m)	RX Power at UE1 due to RRH0 (dBm)	RRH1 Atten. (dB)	RX Power at UE1 due to RRH1 (dBm)	Signal-Interference - Ratio (dB)	CRC/EVM Pass/Fail	EVM measured (in %)	Min. LTE EVM requirement for 16QAM (in %)
2	-59	0	-59	0	Fail/Fail	NA	13.5
2	-59	3	-62	3	Fail/Fail	NA	13.5
2	-59	6	-65	6	Fail/Fail	NA	13.5
2	-59	9	-68	9	Fail/Fail	NA	13.5
2	-59	12	-71	12	Pass/Fail	36.8	13.5
2	-59	15	-74	15	Pass/Fail	27.6	13.5
2	-59	18	-77	18	Pass/Fail	19.6	13.5
2	-59	21	-80	21	Pass/Pass	13.5	13.5
2	-59	21	-83	24	Pass/Pass	9.5	13.5
2	-59	21	-86	27	Pass/Pass	7.5	13.5
2	-59	21	-89	30	Pass/Pass	5.4	13.5

Table 16: EVM Measurements for Scenario 4: 16QAM modulation

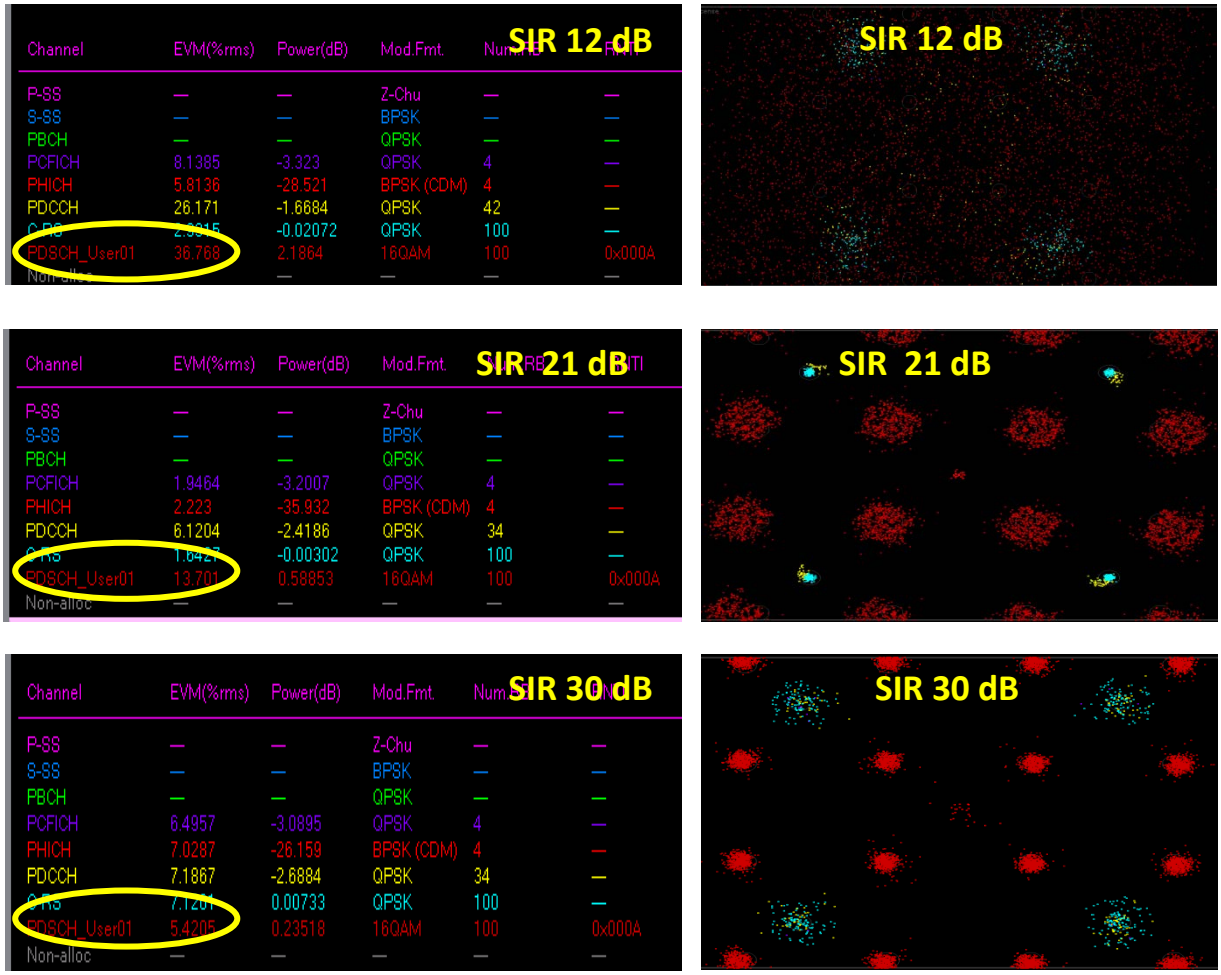


Figure 82: Scenario 4: EVM and constellation diagram from VSA software

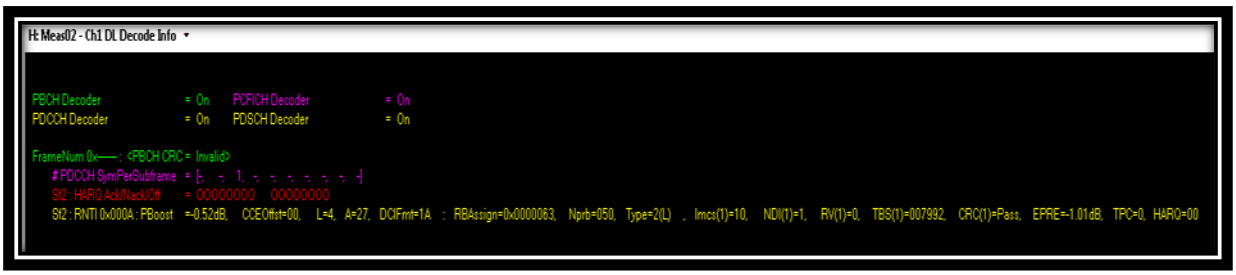


Figure 83: Scenario 4: DL Decode Info from VSA software

8.6 Summary

The real-time LTE PHY based DAS testbed has been designed and implemented on a DSP based SDR platform. The PHY testbed has been successfully tested against commercial Keysight 89600 VSA software running on Keysight N9020A MXA signal analyzer. Selection transmission DAS technique has been investigated in the view of analyzing coordinated pairing and scheduling of RBs in DAS. Based on our research work in COHERENT in WP3 [17], efficient pairing of RRHs and UEs, can exploit the chances of RB reuse in the same cell. In order to efficiently and reliably achieve coordinated pairing and scheduling, a metric called probability of coverage has been derived by UDE in WP3. MATLAB based simulation results of this metric in Rayleigh fading channel provided good results [18].

The goal of our work in WP6 firstly is to design and implement the real-time LTE PHY based DAS testbed and finally to validate the probability of coverage metric using the testbed. In this regard, different scenarios based on the placement of RRHs and UE, different transmit powers and modulation schemes have been investigated and measurement results have been captured on the Keysight VSA software running on Keysight MXA N9020A signal analyzer. EVM is measured on

VSA software in different scenarios by varying the distances, modulation schemes and transmit power of RRHs. EVM metric is chosen by 3GPPP for LTE conformance testing. The results for each scenario have been described in detailed above in the sections related to the scenarios. Almost linear mapping of the simulation results based on probability of coverage to the measurement results has been achieved. This provides the validation of using the probability of coverage metric, which we have derived in WP3 work, for coordinated pairing and scheduling of RBs. From our measurement results, for QPSK modulation an SIR above 18 dB and for 16QAM an SIR above 21 dB at the UE location is required to meet the conformance requirements of LTE [26], and these required SIR values which in turn map to the targeted SINR in the probability of coverage metric provides the reliable means to decide on the pairing of RRHs and UEs in the DAS network.

MATLAB based GUI has been implemented to generate Network graphs based on the probability of coverage metric. These network graphs simulated in MATLAB and presented in this chapter serve as an example of how the network graphs based on the probability of coverage metric are generated at the C3/RTC and how the C3/RTC can decide based on the probability of coverage metric in order to instruct the DAS eNodeB for coordinated pairing and scheduling for efficient RB reuse.

9. Technical and business impact

COHERENT Key Performance Indicators (KPIs) act as quantification for the technical performance of the systems developed. To this end, a number of KPIs are used to evaluate the performance of the corresponding components and functions tested, for every experiment and any functionality demonstrated.

9.1 5G PPP KPIs

5G PPP Program has defined following high-level KPIs (P1-P6) (Table 17). KPIs P1, P2, P3 and P5 are more relevant to COHERENT and have been addressed by the testbeds.

5G PPP KPIs			Relevant to COHERENT
Performance KPIs	**P1**	Providing 1000 times higher wireless area capacity and more varied service capabilities compared to 2010	High
	P2	Saving up to 90% of energy per service provided	Medium
	P3	Reducing the average service creation time cycle from 90 hours to 90 minutes	High
	P4	Creating a secure, reliable and dependable Internet with a “zero perceived” downtime for services provision	Low
	P5	Facilitating very dense deployments of wireless communication links to connect over 7 trillion wireless devices serving over 7 billion people	Medium
	P6	Enabling advanced user controlled privacy	Low

Table 17: High-level KPIs of 5G PPP Program

9.2 COHERENT KPIs

In Table 18 we map the list with the important KPIs of the evaluation process presented in Chapter 6 of D6.1 to the specific COHERENT use cases. These KPIs can be mapped to the corresponding PoC as highlighted in bold texts.

	KPI ID	COHERENT KPI Name	Target Values	Final value and Related PoC
High-level Application level & System KPI	**1.1**	Ability for a control app to request and reserve radio resources, or bitrate/latency for a slice	Binary value = Yes	Yes, related to **RAN sharing** PoC
	1.2	RAN Spectrum efficiencies and radio resource partitioning	Binary value = Yes	Yes, related to **RAN sharing** and **Spectrum management** PoCs.
	1.3	Handover functional to small cell or Wi-Fi load on testbed	Binary value = Yes	Yes, related to **Load balancing** and **RAN Monitoring and Proactive Control** PoCs.
	1.4	Run slices on top of testbeds. Measures the ability of the COHERENT framework to operate with concurrent vertical slices	Binary value = Yes	Yes, related to **RAN sharing** and **Spectrum management** PoCs.
	1.5	Multi-domain network provisioning	Binary value = Yes	Yes, related to **Load balancing** PoC.
	1.6	RTC requirements satisfied. Related also to protocol messaging exchange between RTC and C3	Binary value = Yes	Yes, it is inherent from both FlexRAN Controller and 5G-Empower platform and all PoCs are related.
	1.7	Probabilistic throughput / attainable throughput requirement satisfied. (Used to trigger proactive control and management actions and select between candidate links to implement this action)	Binary value = Yes	Yes, related to **RAN Monitoring and Proactive Control** PoC.
RAN KPI	**2.1**	RAN latency, can be improved with better scheduling techniques	30-100 ms (depending on the channel quality and traffic load)	Yes (see Annex 11.2), it is inherent from OAI platform and related to **RAN sharing** and **Spectrum management** PoCs.
	2.2	RAN jitter, delay variation introduced by the radio access networks, it will be useful when exploring scheduling techniques	<10 ms	1.3-1.9 ms (see Annex 11.2), it is inherent from OAI platform and related to **RAN sharing** and **Spectrum management** PoCs.
	2.3	Ability to dynamically change MAC scheduling	Binary value = Yes	Yes, through RTC
	2.4	Ability to increase PMR slice capacity/throughput and/or robustness from Normal slice resources.	Binary value = Yes	Yes, related to **RAN sharing** PoC
	2.5	Ability to increase Normal slice capacity/throughput and/or robustness without decreasing throughput and robustness for PMR slice	Binary value = Yes	Yes, related to **RAN sharing** PoC
	2.6	Fast deployment time for PMR Network	Binary value = Yes	Yes, related to **RAN sharing** PoC
	2.7	User Attainable Throughput Satisfaction	Throughput QoS requirement satisfied = Yes, within a tolerance	Yes, related to **RAN Monitoring and Proactive Control** and **Load Balancing** PoCs.

			level	
	2.8	Attainable Throughput enhancement	Up to 2 times increase in sum throughput can be achieved for DAS with 2 RRHs in some scenarios, based on standard LTE	Yes, related to **Distributed Antenna System** PoC.
EPC KPI	**3.1**	Ability of EPC and scheduling to support a range of service profiles	Up to 5 different profiles	N/A
	3.2	EPS Attach Success Rate (EASR), which can be used to determine how COHERENT can improve connectivity.	>99%	90% (see Annex 11.2), it is inherent from OAI platform.
	3.3	Dedicated EPS Bearer Creation Success Rate	>90%	80% (see Annex 11.2), it is inherent from OAI platform.
	3.4	Dedicated Bearer Set-up Time by MME (DBSTM)	< 1sec	228ms (see Annex 11.2), it is inherent from OAI platform.
	3.5	Service Request Success Rate (SRSR), it determines the success rate of the times that the UE goes from idle to connected (for instance after paging occasions)	>90%	N/A
	3.6	Inter-RAT Handover Success Rate (IRATHOSR), to characterize the performance of seamless handover	>90%	N/A
Traffic KPI	**4.1**	End To End delay at transport level (UE to S-GW)	<150 ms	35-59 ms (see Annex 11.2), it is inherent from OAI platform and related to **RAN sharing** PoC.
	4.2	Jitter for Channel	< 5 ms	N/A
	4.3	Packet Loss rate for channel (UE to server endpoint)	<0.8 %	N/A
	4.4	Data rate: Network Layer	15 Mb/s (for 5MHz bandwidth, single user) fairness: 0.3-0.4	Yes (see Annex 11.2) it is inherent from OAI platform and related to **RAN sharing** and **Spectrum management** PoCs.
	4.5	Reduced packet loss for PMR slice	Binary value = Yes	Yes, related to **RAN sharing** PoC
	4.6	Reduced end-to-end delay for PMR slice	Binary value = Yes	Yes, related to **RAN sharing** PoC

Table 18: Technical KPIs used for system evaluation

10. Conclusions

The design, deployment and evaluation procedures in WP6 aims to provide a proof of principle for the COHERENT architecture and to evaluate the practical feasibility and performance of the technical solution proposed. The test cases specified in the Deliverable D6.1 [1], namely Spectrum Sharing, RAN Sharing, Load balancing Distributed Antenna System have been designed taking into consideration the functionalities which were proposed by the system architecture in Deliverable D2.2 [2]. This deliverable describes in detail the execution, results and validation of the use cases.

Two reference implementations of the COHERENT C3 as anchors for all experiments: the FlexRAN controller and the 5G-EmPOWER platform has been used. Using COHERENT C3, multiple PoCs have been built during the project: 1) NIF for LTE as a part of the monitoring and management component of the COHERENT network architecture to gather information about the current state of the network from each network element and aggregate them in the network graph to provide a consolidated, probabilistic, RAT-agnostic view of the network state (PoC: RAN Monitoring and Proactive Control); 2) Inter-app communications of RAN control apps (PoC: RAN Inter-app Communications) 3) RAN slicing for PS application (PoC: RAN Sharing); 4) an open programmable platform used for LTE and Wi-Fi load balancing scenarios (PoC: Load Balancing); 5) Spectrum Management Application with the purpose of dynamically switching on or off the so-called phantom-cell and spectrum sharing between the operators and vertical sectors (PoC: Spectrum management); 6) Distributed Antenna System with coordinated UE pairing for improving/maintaining per user throughput and coverage extension (PoC: Distributed Antenna System).

All experiments have provided quantifiable evaluation results aligned with 5G-PPP KPIs and project objectives. They prove the COHERENT solution feasibility and efficiency and represent a basis for further dissemination and exploitation activities.

11. Annex

11.1 Setup for the RAN Slicing Demonstration

The setup with the required material and the required procedures to run the RAN PMR Slicing PoC are described below. This description is not restrictive, other hardware deployments could be used.

11.1.1 Required Material

Required material (both Hardware and Software):

- 2 computers:
  - 1 PC (laptop PC1) for controlling the eNodeB;
  - 1 PC (laptop PC2) for:
    - EPC including several components such as HSS, PGW/SGW, MME (in other potential implementation the EPC can be deported on another PC representing a CN separated from the RTC controller, e.g., on PC3, not represented here);
    - FlexRAN controller (developed by EURE);
    - RRM application for RAN slicing (developed by TCS in Python).
- 1 USRP B210 used as RF eNodeB module and connected to the PC used as eNodeB (i.e., PC1) and implementing L1, L2, L3;
- 1 USB cable, 1 USRP B210 antenna for Band 7, in a 5 (or 10) MHz Band, 2.6 GHz, FDD (according to the eNodeB configuration file);
- 1 (single) Ethernet cable between PC1 and PC2, on top of which both ProtoBuf interface (used between RTC FlexRAN and eNodeB Agent) and S1 interfaces (S1-U and S1-MME) will run;
- (at least) 2 slices are created on the eNodeB side;
- (currently) 2 COTS UEs: 1 connecting to a slice, 1 connecting to the other slice;
- 2 reconfigurable SIM cards for UE-HSS interaction in EPC;
- RRM app interacts with the FlexRAN controller through Yaml and JSON interfaces connected through a Pistache Web Server and Pistache Web Client both located on PC2 (in other potential implementation the client can be deported on another PC representing a distant server, e.g., on PC4, not represented here)
- 1 RTC FlexRAN Controller on PC2;
- 1 SW Agent on PC1;
- PC2 can be connected (or not) to an external IP interface (internet or other) through e.g., Wi-Fi
- Iperf directly deployed on UE side and on PC2 side is further used to evaluate throughput & transmission performance in real time.

11.1.2 Required Procedures to Run the PoC

1. Connection Phase (mandatory):
   a. Connect USRP to PC1 through USB cable;
   b. Connect PC1 to PC2 through Ethernet cable in peer-to-peer mode;
2. Check-up Phase (optional):
   a. Check the installation of OAI software + the Agent on PC1;
      i. https://gitlab.eurecom.fr/oai/openairinterface5g/tree/develop or
      ii. https://gitlab.eurecom.fr/oai/openairinterface5g/commits/develop_integration_2018_w10
   b. Check the configuration of the eNodeB on PC1 (IP address for connection with EPC, frequency band, etc.);
   c. Check the installation of HSS, MME, S/P-GW in the EPC on PC2 and the configurations (IP address for connection with EPC, HSS certificate for the Diameter protocol using to communicate with the MME, etc.);
      i. https://gitlab.eurecom.fr/oai/openair-cn/tree/develop

d. Check the installations of RTC FlexRAN controller and RRM Slicing Application on PC2;
   i. https://gitlab.eurecom.fr/flexran/flexran-rtc/tree/develop-uplink
   ii. https://gitlab.eurecom.fr/mosaic5g/store/tree/develop
e. Check the installation of the interfaces (ProtoBuf, Yaml, JSON);

3. Initialization of the testbed before running the demonstration (mandatory):
   a. Start HSS;
      i. gnome-terminal -t HSS --
      ii. command="/home/coherent2/Projects/openair-cn/SCRIPTS/run_hss"
   b. Start MME;
      i. gnome-terminal -t MME --
      ii. command="/home/coherent2/Projects/openair-cn/SCRIPTS/run_mme"
   c. Start S/P-GW;
      i. gnome-terminal -t SPGW --
      ii. command="/home/coherent2/Projects/openair-cn/SCRIPTS/run_spgw"
   d. Start eNodeB with Agent;
      i. #!/bin/bash
      ii. cd Projects/openairinterface5gflexran/
      iii. source oaienv
      iv. cd cmake_targets/tools
      v. ./make_agent_cache
      vi. cd ../../targets/bin
      vii. sudo -E ./lte-softmodem.Rel14 -O ../PROJECTS/GENERIC-LTE-EPC/CONF/enb.band7.flexran.usrpb210.conf
   e. Start RTC FlexRAN Controller;
      i. cd flexran-rtc
      ii. ./run_flexran_rtc.sh
   f. Start RRM Slice Application (Python);
      i. cd flexran-rtc/sdk
         1. Test mode : python rrm_kpi_app.py --op-mode test
         2. Live mode : python rrm_kpi_app.py –-op-mode sdk
   g. Create 2 slices using Yaml (with required parametrization defined in the previous section);
   h. Check the functionality of Yaml and policy application (information recovered through JSON);
   i. Connect 2 UEs (one will connect by default to the first slice and the second one to the second slice);
   j. Apply different strategies for PMR slice and normal slice and verify the behavior of UE1 and UE2.

11.2 Additional KPI measurement

In order to measure the System KPIs, a set of different tools were used depending on the evaluation metrics and performance indicators that we wanted to collect from in-the-field deployment.

11.2.1 Attachment and session establishment success rates KPIs

In order to measure the system KPIs, we used one UE at different locations with respect to the eNodeB as depicted in the figure below. Two procedures are considered, namely **Initial Attach**, and **EPS dedicated bearer setup** (E_RAB_SETUP). Note that in the latter case a UE can have more than one communication channel with the base station, each one responsible for a different type of traffic. In the conducted experiments two additional channels are established.

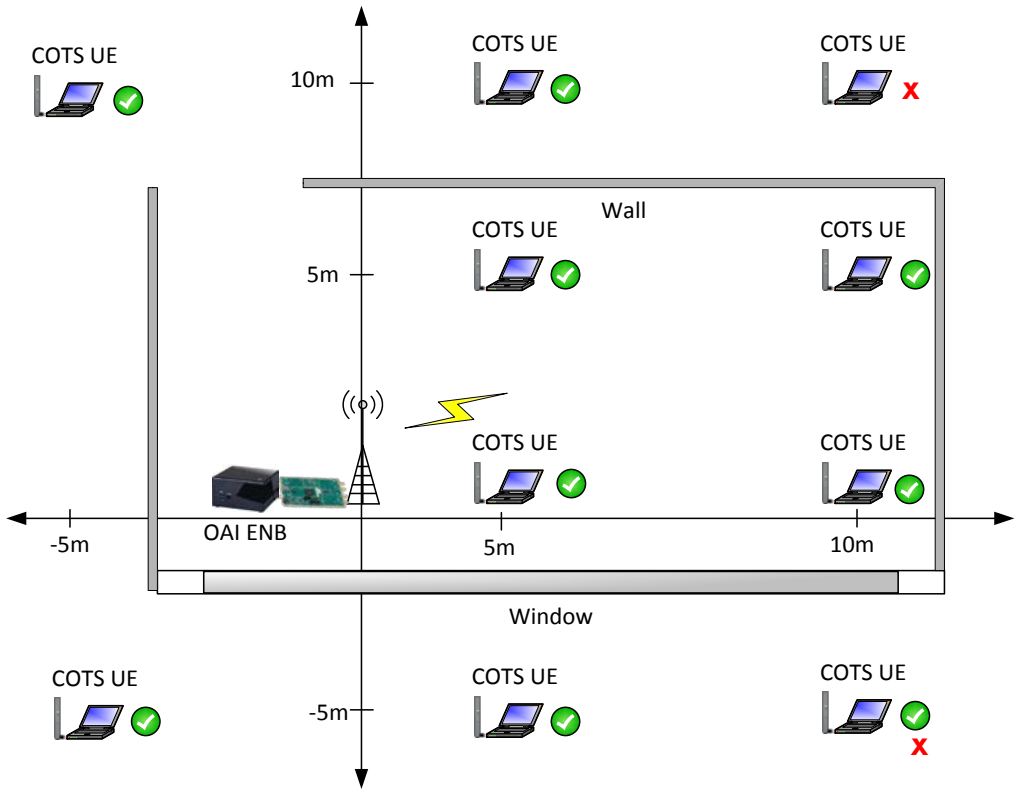


Figure 84: Measurement setup for the system KPIs

The initial attachment procedure would fail only in cases where the user is experiencing a deep fading, e.g., due to walls, and thus the user *attach request* message is not detected or failed (overloaded scenarios and channel access contention are not considered in these experiments). As for the dedicated bearer session, two cases are observed: either failure due to failure of the attach procedure, or failure after a successful initial attach procedure. The latter occurs primarily due to uplink errors. The measurements have shown a success rate of **90% for the initial attach procedure, and 80% for the EPS radio access bearer establishment**.

11.2.2 Attachment and session establishment delay KPIs

For measuring the Attachment and Session delay KPIs, we used Wireshark so that we can measure delays based on the timestamps in the collected traffic traces. Wireshark traces were taken from both ENB and EPC. The following figure describes the signaling that takes place for the Initial attach and the EPS dedicated bearer setup procedures.

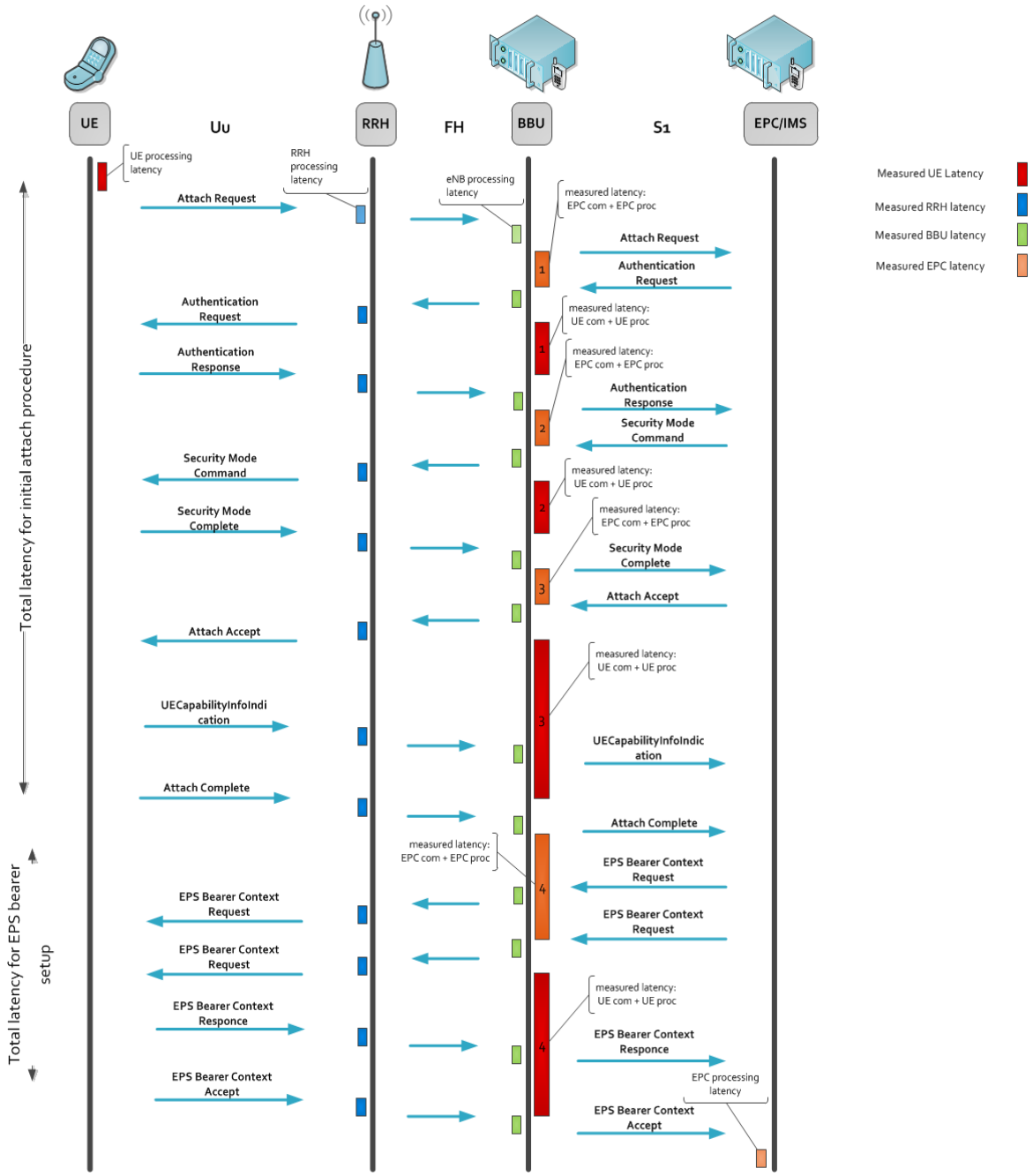


Figure 85: Initial attach and EPS dedicated bearer procedure latency breakdown

Having as a reference the figure above and using the traces acquired by Wireshark, the following tables with performance indicators are derived:

Latency (ms)	1	2	3	Total
UE communication and processing	315 (0.3147)	18 (0.0178)	250 (0.2502)	583
EPC communication and processing	40 (0.0402)	2 (0.0021)	8 (0.0083)	50
Total	355	20	258	633

Table 19: Initial attach procedure latency (eNodeB traces)

Latency (ms)	1	2	3	Total
UE communication and processing	315 (0.3149)	18 (0.0181)	250 (0.2502)	583
EPC communication and processing	39 (0.0399)	2 (0.0019)	8 (0.0081)	49
Total	354	20	258	632

Table 20: Initial attach procedure latency (EPC/IMS traces)

eNodeB Perspective	
Latency (ms)	**4**
UE communication and processing	219 (0.2190)
EPC communication and processing	9 (0.0090)
Total	228

EPC/IMS Perspective	
Latency (ms)	**4**
UE communication and processing	219 (0.2193)
EPC communication and processing	9 (0.0087)
Total	228

Table 21: EPS dedicated bearer procedure latency

For both procedures, we can see that the results from both probing points (ENB and EPC) are quite similar and their differences, if any, are very limited and in the order of 1ms. For the initial attach procedure, the total control-plane latency measured is 633 ms and, for the EPS dedicated bearer set up, is 228 ms. In both cases the radio access network is the one that has been experimentally found to introduce the largest part of the measured latency.

11.2.3 Data plane QoS KPIs

For measuring the Data plane QoS we used iperf, a client-server based tool for measuring bandwidth between two endpoints. The data traffic used is of type UDP and the data rate employed is 15Mbit/sec and 30Mbit/sec depending on the bandwidth configuration at eNodeB. We used two PCs: in one of them iperf ran in client mode and was generating traffic for the UE (LTE dongle), which was connected to another PC in which iperf was running in server mode and was just discarding the traffic.

For the Data plane QoS it is interesting not only to benchmark the current implementation of the RAN service, but also to compare the altered RAN architecture with the performance results achieved by the original one (up to M36, before the project extension period). Additionally, it is important to investigate the behavior of the extended RAN architecture under different circumstances; the fact that the I/Q samples are packetized and transported via Ethernet makes hard and sometimes impossible to meet the frame/subframe and HARQ timing constraints imposed by LTE. As result, in the following figures we choose to have the throughput results for the following cases:

- **eNodeB_1, 5 MHz/10 MHz -** refers to the monolithic eNodeB.
- **eNodeB_2, 5MHz/10MHz -** refers to disaggregated eNodeB with local RRU (remote radio unit).
- **eNodeB_3, 5 MHz/10MHz -** refer to disaggregated eNodeB with remote RRU with split 8 (two distinct physical hardware and the temporal I/Q samples are transported over Ethernet)

The table and figure below present the average jitter, drop rate, and goodput of eNodeB in the three considered scenarios for 5MHz and 10MHz channel bandwidth. It can be seen that the same performance is achievable with and without RRU, confirming the feasibility of packet-based fronthaul network based on Ethernet. In other words, when having a 5 or 10 MHz radio bandwidth the I/Q samples along with the data introduced by the fronthaul headers are able to fit in a 1Gbit link and arrive on-time to the eNodeB (LTE stack timing is respected).

Metric / Scenario	eNodeB_1	eNodeB_2	eNodeB_3
average jitter values (ms) 5MHz	1.998ms	1.939ms	1.943ms
average jitter values (ms) 10MHz	1.611ms	1.393ms	1.403ms
average drop rate values (%) 5MHz	7.5170e-006%	2.3232e-005%	3.8787e-005%
average drop rate values (%) 10MHz	9.5807e-004%	2.3646e-005%	0.0011%

Table 22: Jitter and drop rate of eNodeB in different scenarios

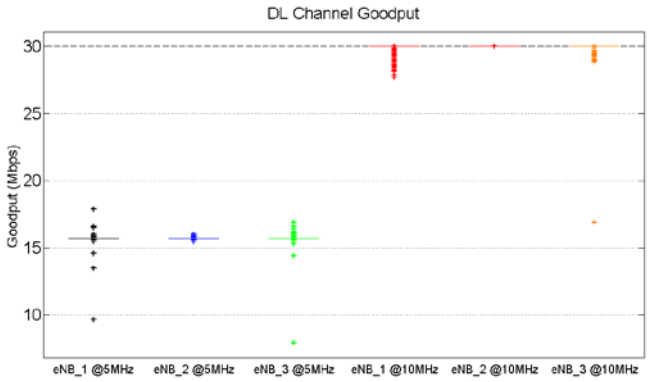
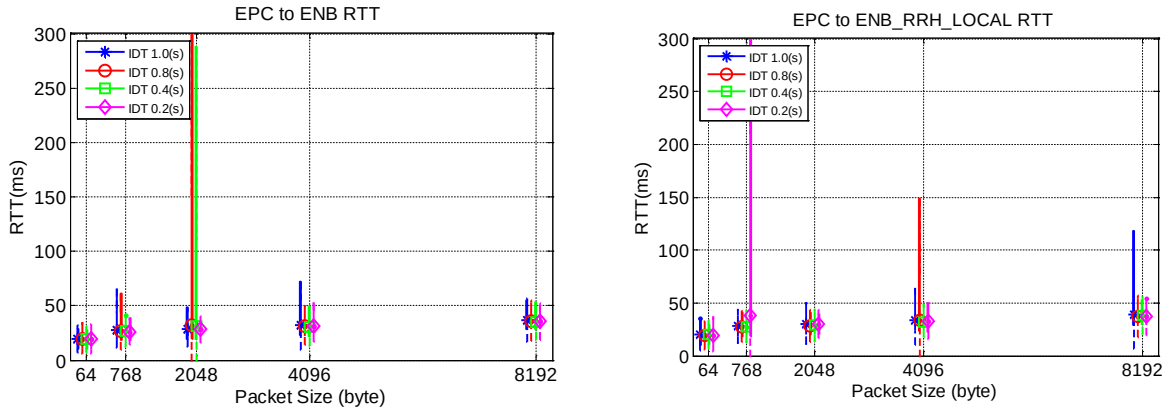


Figure 86: Measured goodput at the UE for different scenarios

11.2.4 Data plane delay KPIs

For the data plane delay KPI, we measure the application Round Trip Time (RTT) over the default radio bearer. In particular, RTT is evaluated through different traffic patterns generated by the *ping* tool with 64, 768, 2048, 4096, 8192 Packet Sizes (PS) in byte, and 1, 0.8, 0.4, and 0.2 Inter-Departure Time (IDT) in seconds. While our focus is on the evaluation of the RAN service, a third party EPC at EURECOM is used to measure the RTT at the application level.

The figure below presents the measured RTT as a function of PS and IDT for the three considered cases, namely standalone eNodeB, eNodeB with local RRU, and eNodeB with remote RRU. As expected, higher RTT values are observed when PS and IDT increase. It can be also noted that the RTT trend is similar in the three cases, and thus confirming the feasibility of RRU.



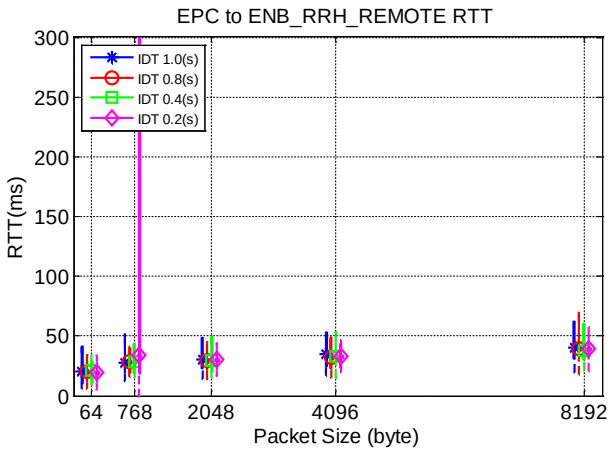


Figure 87: Data-plane round trip time for the three considered scenarios

Summary of System KPIs

The table below summarizes the experimental results about the system KPIs for the eNodeB located at EURECOM.

KPIs	Units	KPI measurement
Registration success rate	Rate (%)	90%
Session establishment success rate	Rate (%)	80%
Session drop rate	N/A	N/A
Attachment delay	Time (ms)	633 ms
Session establishment delay	Time(ms)	228 ms
Average Data plane QoS	Time (ms) Rate (%)	average jitter: • 5MHZ: 1.939ms • 10MHz: 1.393ms average packet loss: • 5MHZ: 2.3232e-005% • 10MHz: 2.3646e-005% average throughput: • 5MHZ: 16MBps • 10MHz: 30MBps
Data plane delay	Time(ms)	Round trip delay: • 64 bytes/packet: 35 ms • 768 bytes/packet: 45 ms • 2048 bytes/packet: 44 ms • 4096 bytes/packet: 48ms • 8092 bytes/packet: 59 ms

Table 23: OpenAirInterface ENB System KPIs

REFERENCES

[1] D6.1, Draft report on technical validation, COHERENT project deliverable, 2017.

[2] D2.2, System architecture and abstractions for mobile networks, COHERENT project deliverable, 2016.

[3] CREATE-NET, 5G-EmPOWER Multi-access Edge Computing Operating System. On line: http://empower.create-net.org/

[4] CREATE-NET, 5G-EmPOWER, Documentation. On line: https://github.com/5g-empower/5g-empower.github.io/wiki

[5] CREATE-NET, 5G-EmPOWER, Source code. On line: https://github.com/5g-empower/

[6] EURECOM, OpenAirInterface. On line: https://www.openairinterface.org/

[7] D3.2 Final report on Physical and MAC layer modelling and abstraction, COHERENT project deliverable, 2017.

[8] Git repository for NIF code, https://github.com/nigsics/aquamet

[9] A. Rao, R. Steinert. “Probabilistic multi-RAT performance abstractions”. Submitted for review in the third IFIP/IEEE International Workshop on Management of 5G Networks (5GMan 2018).

[10] Git repository for NIF on FlexRAN SDK

[11] Git repository for NIF on EmPOWER SDK

[12] 3GPP TS 36.212 V9.2.0}, Multiplexing and channel Coding (Release 9)

[13] 3GPP TS 36.213 V9.2.0}, Physical layer procedures (Release 9)

[14] 3GPP TS 36.211 V9.1.0}, Physical Channels and Modulation (Release 9)

[15] 3GPP TS 36.214 V9.1.0}, Physical layer - Measurements (Release 9)

[16] 3GPP TS 36.101 V9.3.0}, User Equipment (UE) radio transmission and reception (Release 9)

[17] D. Peethala, N. Zarifeh and T. Kaiser, "Probability of coverage based analysis of distributed antenna system and its implementation on LTE based real-time-testbed," *2017 9th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*, Munich, 2017, pp. 226-231.

[18] COHERENT, Deliverable D4.2 “Final report on flexible spectrum management”, Coordinated Control and Spectrum Management for 5G Heterogeneous Radio Access Networks, H2020 COHERENT project under Grant Agreement No.: 671639

[19] H. Ishii, Y. Kishiyama and H. Takahashi, "A novel architecture for LTE-B :C-plane/U-plane split and Phantom Cell concept," 2012 IEEE Globecom Workshops, Anaheim, CA, 2012, pp. 624-630, doi: 10.1109/GLOCOMW.2012.6477646

[20] D2.4, Draft report on technical validation, COHERENT project deliverable, 2017.

[21] D5.2, Final Specification and Implementation of the Algorithm from Programmable Radio Access Networks, COHERENT project deliverable, 2017

[22] D2.3, Initial release of the SDK, COHERENT project deliverable, 2016

[23] X. Foukas, N. Nikaein, M. Kassem, M. Marina and K. Kontovasilis. "FlexRAN: A flexible and programmable platform for software-defined radio access networks." In Proceedings of the 12th International on Conference on emerging Networking EXperiments and Technologies, pp. 427-441. ACM, 2016.

[24] JSON-RPC 2.0 Specification, available at http://www.jsonrpc.org, last checked 18 April 2018

[25] Travelping, erGW - Erlang implementations of GGSN or P-GW, https://github.com/travelping/ergw

[26] 3GPP TS 36.141 V9.8.0, LTE, EUTRA Base station conformance testing (Release 9), Table 6.5.2.5-1

[27] Docker image of erGW PDN-GW control plane, https://hub.docker.com/r/ergw/ergw-pgw-c-node/

[28] Docker image of erGW PDN-GW user plane, https://hub.docker.com/r/erw/ergw-pgw-u-node/