



Deliverable 3.3

Final Evaluation and Benchmarking of the Implemented PHY Layer Solutions

March 2025



Co-funded by
the European Union



Contractual Date of Delivery:	February 28, 2025
Actual Date of Delivery:	March 25, 2025
Editor(s):	Israel Koffman (REL)
Author(s)/Contributor(s):	Govindarajan Mohandoss (ARM)
	Anna Umbert, Juan Sánchez-González, Jordi Pérez-Romero,
	Oriol Sallent (UPC)
	Dr. Eli Shasha, Baruch Globen, Israel Koffman (REL)
	German Castellanos (ACC)
	Vladica Sark, Jesús Gutiérrez, Darko Cvetkovski, Mert Özates (IHP)
	J. Xavier Salvat, Jose A. Ayala, Marco Rossanese, Placido Mursia (NEC)
	Esteban Municio (I2CAT)
	Ory Yeger (PW)
	Mir Ghoraishi (GIGASYS)
Work Package	WP3
Target Dissemination Level	Public

This work is supported by the Smart Networks and Services Joint Undertaking (SNS JU) under the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101097083, BeGREEN project. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or SNS-JU. Neither the European Union nor the granting authority can be held responsible for them.

Revision History

Revision	Date	Editor / Commentator	Description of Edits
0.1	2024-11-12	Israel Koffman (REL)	Initial document created
0.2	2024-12-22	Israel Koffman (REL)	Section 3.2 initial content added – RU energy efficiency approach
0.3	2025-01-08	Anna Umbert (UPC)	Add relay interference study (4.5)
0.4	2025-01-12	Ory Eger (PW)	Section 2.1 content added
0.5	2025-01-13	Esteban Municio (i2CAT)	Section 4.4 added
0.6	2024-01-15	J. Xavier Salvat (NEC)	Section 4.2 added
0.7	2025-02-01	German Castellanos (ACC)	Section 2.3 and 3.1 added
0.8	2025-02-04	Vladica Sark (IHP)	Section 4.1 added
0.9	2025-02-05	Mert Özates (IHP)	Added content to section 4.1
0.91	2025-02-28	Israel Koffman (REL)	Final review
1.0	2025-03-25	Jesús Gutiérrez (IHP) Mir Ghoraishi (GIGASYS) Simon Pryor (ACC)	Final revision, proof reading and submission

Table of Contents

List of Acronyms	8
Executive Summary	11
1 Introduction	12
2 DU and CU Acceleration Final Evaluation and Benchmarking.....	13
2.1 GPU DU acceleration architecture	13
2.2 DU LDPC decoder research results analysis and benchmarking.....	15
2.3 CU/RIC acceleration research results analysis and benchmarking.....	16
2.3.1 ARM testing - performance	17
2.3.2 ARM testing - latency	21
2.3.3 Results	23
2.4 Collaborative insights.....	24
3 RU Energy Consumption Optimization Research Results.....	25
3.1 RU Management (xApps/rApps) research results analysis and benchmarking.....	25
3.1.1 Test Definition	25
3.1.2 ARM results	27
3.1.3 x86 Results.....	28
3.1.4 Comparison and Conclusions	29
3.2 RU power consumption reduction research results analysis and benchmarking	29
3.2.1 Power consumption reduction.....	33
4 PHY Layer Enhancements for Improving RAN Power Efficiency – Final Research Results and Benchmarking	37
4.1 ISAC assisted physical layer improvements results and benchmarking	37
4.1.1 Angular and ranging resolution of the sensing system	37
4.1.2 Referent sensing tests in an anechoic chamber.....	38
4.1.3 Sensing assisted beam forming	43
4.1.3.1 System model	43
4.1.3.2 Sensing-assisted Beamforming.....	44
4.1.3.3 Sensing-assisted Beam Codebook Design	44
4.1.3.4 Numerical Results.....	45
4.2 Self-configuring RIS final research results and benchmarking	49
4.3 Integration concept for ISAC and RIS into O-RAN.....	53
4.4 Frequency assignment strategies in relay-enhanced scenarios	55
4.4.1 Frequency assignment strategies.....	55
4.4.2 Scenario under study.....	57
4.4.3 Performance assessment	59
4.4.4 Conclusions.....	63
5 Summary and Conclusion	65

List of Figures

Figure 2-1 LDPC threading concept	14
Figure 2-2 CPU/GPU memory management	14
Figure 2-3 GPU load measurement	15
Figure 2-4 Grafana dashboard for the ARM performance testing	18
Figure 2-5 Implementation of the xApp via dashboard	18
Figure 2-6 dRAX dashboard showing the implemented xApp.....	18
Figure 2-7 Grafana for latency testing.....	22
Figure 2-8 Power consumption comparison for ARM and x86. (left) Average (right) Distribution.....	23
Figure 2-9 Average comparison for ARM and x86. (left) CPU usage. (right) Memory usage	23
Figure 2-10 Latency distribution and comparison for ARM and x86.....	24
Figure 3-1 xApp/rApp system used for testing including x86 server measurements.	25
Figure 3-2 Scenario for 6 Cells 20 UEs (up) and Scenario for 20 Cells 100UEs (down).....	26
Figure 3-3 General dashboard for the energy savings system including the server power consumption metric.....	27
Figure 3-4 ARM Server power consumption vs. energy savings metric.	28
Figure 3-5 x86 server power consumption vs. energy saving metric	29
Figure 3-6 Private 5G network architecture.....	31
Figure 3-7 Sparq-2025-ORU-3.5G Internal view.....	31
Figure 3-8 Sparq-2025-ORU GPS, Power and Communication Interfaces.....	31
Figure 3-9 Amarisoft SW Package Components.....	32
Figure 3-10 PA blanking test setup.....	34
Figure 3-11 RU PA blanking test at RunEL Lab	34
Figure 3-12 PA blanking method	35
Figure 3-13 Energy save at a blanked symbol	35
Figure 3-14 Slot before and after REB's Arrangement	35
Figure 3-15 Western Europe traffic Distribution	36
Figure 3-16 RF PA Block Diagram and Physical Structure.....	36
Figure 4-1 Angle and distance of the detected object in polar coordinates	38
Figure 4-2 Angular and range resolution.....	38
Figure 4-3 Anechoic chamber at IHP with installed ISAC system	39
Figure 4-4 Photo of the anechoic chamber ISAC setup with a person in front of the sensing system	40
Figure 4-5 Sensing heat map in polar coordinates showing a single person.....	40
Figure 4-6 Sensing scenario with two persons in front of the sensing system	41
Figure 4-7 Photo of the sensing scenario with two persons, the second person is taking the photo	42
Figure 4-8 Photo of the computer showing the heat map of the two-person scenario	42
Figure 4-9 Heat map of the two-person scenario	43
Figure 4-10 Success rate of sensing-assisted beamforming and beamforming without sensing information for different codebook sizes	46
Figure 4-11 Success rate versus SNR for sensing-assisted beamforming and uniform codebook	47
Figure 4-12 Success rate versus codebook size for non-uniform user distribution	47
Figure 4-13 Bit error rate comparison of proposed sensing assisted beamforming, uniform codebook and no beamforming case.....	48
Figure 4-14 RMSE comparison of sensing assisted beamforming and uniform codebook	48
Figure 4-15 Example of power profile and corresponding estimated peak for different codebook sizes L in a multi-UE scenario with $K = 14$ UEs and $N = 32$ HRIS elements.	50
Figure 4-16 Radiation pattern at the HRIS along the azimuth directions obtained with O-MARISA, O-wMARISA, and SoA (left), and radiation patterns obtained with Q bits of phase quantization (right).	52
Figure 4-17 Average sum-rate in a multi-UE scenario obtained by solving Problem 3 with perfect CSI and by SoA [4] against the number of UEs K for different network areas and when the number of HRIS elements is $N=32$	52
Figure 4-18 Average sum-rate in a multi-UE scenario obtained by solving Problem 3 with perfect CSI and by SoA [4] against the number of HRIS elements N , the number of UEs K . The network area is $A=50$ "m" $\times$ 50 "m"	53
Figure 4-19 O-RAN based BeGREEN architecture focusing on the RIS/ISAC components	54
Figure 4-20 Frequency assignment strategies in a single-cell scenario.....	56

Figure 4-21 Time-frequency resource usage for the in-band strategy with half-duplex.....	56
Figure 4-22 Uplink and downlink use in the two slots for the in-band strategy with half-duplex	57
Figure 4-23 Scenario under study	57
Figure 4-24 Transmitted power at the BS (a) and relay (b) when 1 or 2 UEs are connected through the relay	59
Figure 4-25 Total power consumption (a) and energy efficiency (b) when 1 or 2 UEs are connected through the relay.....	60
Figure 4-26 Maximum number of supported users as a function of the required bit rate R_k	60
Figure 4-27 Energy Efficiency as a function of the required bit rate R_k	61
Figure 4-28 Total power consumption for all combinations of power consumption values	62
Figure 4-29 Comparison with Spatial Beam-based of: total power consumption (a), and total energy efficiency (b)	63

List of Tables

Table 2-1 LDPC Scenarios Parameters.....	15
Table 2-2 LDPC Processing Comparing ARM and GPU Running 32 CBs in Parallel.....	16
Table 2-3 LDPC Processing Comparing ARM and GPU Running 128 CBs in Parallel.....	16
Table 3-1 Power Consumption ARM vs x86	29
Table 3-2 Radio Unit (Sparq-2025-ORU) Interfaces.....	32
Table 3-3 PA Blanking Tests Results	36
Table 4-1 Simulation setup and parameters	51
Table 4-2 Configuration Parameters	58
Table 4-3 Spectrum Distribution	58
Table 4-4 Combinations of Power Consumption Parameters	58

List of Acronyms

3GPP	3rd Generation Partnership Project
5GC	5G Core
5G NR	5G New Radio
AI	Artificial Intelligence
AoA	Angle-of-Arrival
API	Application Programming Interface
AWGN	Additive White Gaussian Noise
B5G	Beyond 5G
BB	Baseband
BPSK	Binary Phase-Shift Keying
BS	Base Station
CNI	Container Network Interface
COTS	Commercial Off-The-Shelf
CP	Control Plane
CPE	Customer-premises Equipment
CRC	Cyclic Redundancy Check
CSI	Channel State Information
CU	Centralized Unit
CUDA	Compute Unified Device Architecture
CU-CP	Centralized Unit – Control Plane
CU-UP	Centralized Unit – User Plane
DL	Downlink
DNN	Digital Neural Network
DoA	Direction of Arrival
DoW	Description of Work
DPD	Digital Pre-Distortion
DU	Distributed Unit
E2E	End-to-End
EIRP	Equivalent Isotropically Radiated Power
ES	Energy Saving
ET	Envelope Tracking
FFT	Fast Fourier Transform
FMCW	Frequency-Modulated Continuous Wave
FP	Fractional Program
GPU	Graphics Processing Unit
HRIS	Hybrid Reconfigurable Intelligent Surface
HW	Hardware
IFFT	Inverse Fast Fourier Transform
IOS	Internet-of-Surfaces
InfluxDB	Influx Database
IP	Internet Protocol
ISAC	Integrated Sensing and Communication
ISM	Industrial Scientific and Medical
JSON	JavaScript Object Notation
KPI	Key Performance Indicator
L2	Layer 2

LDPC	Low-Density Parity Check
LLR	Log Likelihood Ratio
LMFW	Linear Modulated Frequency Waveform
LO	Local Oscillator
LoS	Line-of-Sight
LTE	Long Term Evolution
MAC	Medium Access layer
MARISA	Metasurface Absorption and Reflection for Intelligent Surface Applications
ML	Machine Learning
MIMO	Multiple-Input Multiple-Output
mMIMO	Massive Multiple-Input Multiple-Output
MMSE	Minimum Mean Square Error
mmWave	Millimetre Wave
MNO	Mobile Network Operator
MR	Maximum Ratio
MUSIC	MULTiple Signal Classification
near-RT	near Real-Time
NFV	Network Function Virtualisation
NLoS	Non-Line-of-Sight
NMS	Network Management System
non-RT	non-Real-Time
O-RAN	Open RAN
O-FH	Open Fronthaul
OFDM	Orthogonal Frequency Division Multiplexing
OFDMA	Orthogonal Frequency-Division Multiple Access
PA	Power Amplifier
PAPR	Peak-to-Average Power Ratio
PDCCP	Packet Data Convergence Protocol
PHY	Physical layer
PLA	Planar Linear Array
PoC	Proof-of-Concept
RMSE	Root Mean Square Error
PRB	Physical Resource Block
QAM	Quadrature Amplitude Modulation
QoS	Quality of Service
RAN	Radio Access Network
RB	Resource Block
RIA	Relay's Incidence Area
RIC	Radio Interface Controller
RRC	Radio Resource Control
RRM	Radio Resource Management
RSRP	Reference Signal Received Power
RSS	Received Signal Strength
RTT	Round-Trip Time
RU	Radio Unit
SB	Spatial Beam
SCS	Subcarrier spacing
SDAP	Service Data Adaptation Protocol

SDK	Software Development Kit
SDR	Software Defined Radio
SIMD	Single Instruction Multiple Data
SINR	Signal-to-Interference-plus-Noise Ratio
SLA	Service Level Assurance
SMO	Service and Management Orchestrator
SNR	Signal-to-Noise Ratio
SotA	State-of-the-Art
SVE	Scalable Vector Extension
ToA	Time of Arrival
ToF	Time-of-Flight
TW	Telemetry Gateway
UE	User Equipment
UL	Uplink
ULA	Uniform Linear Array
UPA	Uniform Planar Array
UPS	User Plane Stack
URLLC	Ultra-Reliable Low Latency Communication
UWB	Ultra-Wideband
WR	White Rabbit
XDP	eXpress Data Path
WP	Work Package

Executive Summary

The primary aim of BeGREEN D3.3 is to present the final developments and evaluation of the proposed physical (PHY) layer enhancements and optimization strategies. In BeGREEN D3.1, the State-of-the-Art (SotA) on PHY mechanisms energy consumption was described and innovative solutions and enhancements were suggested. In BeGREEN D3.2, the details of the specific solutions and algorithms developed in BeGREEN were captured. Additionally, their potential in improving energy efficiency (EE) in cellular networks and power measurement results were shown. A wide range of solutions and modules were discussed, each with its unique angle on the O-RAN/3GPP framework for 5G deployments, as well as forward looking to 6G. BeGREEN D3.3 presents the final achievements for improving the EE of PHY layer functions and using advanced PHY layer technologies, like Integrated Sensing and Communication (ISAC) or Reconfigurable Intelligent Surfaces (RISs), as an enabler for improving the EE of the network. It is shown how acceleration, using advanced architectures, can be used to decrease the power consumption of modules with high computational complexity in both the distributed unit (DU) and the centralized unit (CU). Novel radio unit (RU)-related power optimizations are also presented, that is power amplifier (PA) blanking, envelope tracking, and artificial intelligence (AI)-based digital pre-distortion (DPD) algorithms. Also, the coverage extension by using ISAC for optimal allocation of network resources and power efficient RIS for replacing RUs are demonstrated. Finally, interference management methods that can strongly affect the EE of the network utilizing both base stations (BSs) and relays are analysed. Combining the different methods in next generation networks has the potential of being a significant factor in improving their power efficiency, making them greener and environmentally friendly.

1 Introduction

BeGREEN main goal is to develop new strategies for reducing power consumption in all layers and components of the Radio Access Network (RAN). Specifically, Work Package 3 (WP3), puts the focus on developing enablers to significantly reduce the power consumption of the lower layers of the radio stack (L1/L2/L3) in 5G and beyond 5G (B5G) systems. The lower layers of the radio stack in RAN systems encompass the most resource and energy hungry functions of the RAN. They can account for up to 80% of energy consumption, a large part of which is used for idling. The transmitter power amplifier (PA) has the highest energy consumption followed by the complex baseband (BB) processing. Additional energy is spent on cooling the used devices, which almost doubles the used energy. Therefore, the main efforts in WP3 are focused towards reducing the energy consumption in the RU and in the BB units, as well as using new technologies like ISAC and RIS to sense the environment and reduce the idling time, based on the expected traffic.

BeGREEN D3.1 [5] conducted a survey of State-of-the-Art (SotA) energy consumption for L1/L2/L3. The performance of current solutions for the lower layers of the radio stack were analysed, namely: (i) the performance of current Hardware (HW) acceleration for Distributed Unit (DU) and Centralized Unit (CU) processing, (ii) the different kinds of Radio Unit (RU) energy saving features, and (iii) the energy consumption of xApps and rApps for controlling the RU. We also introduced the idea of using Integrated Sensing and Communication (ISAC), reconfigurable intelligent surfaces (RISs) and relays for improving network energy efficiency (EE). Finally, it was highlighted that the basic concepts, ideas, directions and architectures for energy and power saving that will be developed within WP3.

BeGREEN D3.2 [1] proposed, implemented and tested the initial solutions for reducing the energy consumption in the physical (PHY) layer. These results showed that significant energy savings can be achieved with the proposed algorithms. Additionally, new technologies like ISAC and RIS were implemented and tested. It was shown that using ISAC, detection of potential users, i.e. UEs, can be reliably performed and used for reduction of the idling by dynamic and optimal network resource assignment. Finally, it was shown that RIS is a viable solution to increase the coverage in an urban environment, without deployment of additional base stations (BSs), which will increase the energy consumption.

This document, BeGREEN D3.3, describes the final results and the achieved performances of the different enablers proposed in WP3. It also reports the achieved energy savings with the different PHY layer technologies addressed in WP3.

BeGREEN D3.3 is organised as follows. Section 2 shows the achievable reduction in power consumption when running the Low-Density Parity Check (LDPC) decoder and CU-UP (User Plane) algorithms on an ARM and NVIDIA GPU-based platform compared to an x86-based platform. Section 3 shows the developed RU power consumption reduction methods which use RU on/off schemes, PA blanking schemes, Peak-to-Average Power Ratio (PAPR) reduction methods, Digital Pre-Distortion (DPD) and Envelope Tracking (ET), and the achieved energy savings. Section 4 presents the final results of the developed approaches for enhancing coverage in radio cells using ISAC, RIS, and relays. And Section 5 is the summary and conclusions driven from the discussions presented in the document.

2 DU and CU Acceleration Final Evaluation and Benchmarking

The DU defined in the Open RAN (O-RAN) standard is a network component that performs the layer 1 (L1) and layer 2 (L2) processing of the communication protocol. In BeGREEN we are researching methods to increase its EE and reduce its power consumption. Specifically, we focus on the L1 receiver tasks. These tasks need to be done in real-time, and some of the modules there require very high computational power. One of the most common implementations for these algorithms is dedicated HW. The main disadvantage of these implementation is the lack of flexibility and scalability. Even when the tasks are not needed due to low network loads, the HW is still there and cannot be used for other tasks. This has driven the industry and O-RAN towards SW implementation-based on commercial off-the-shelf (COTS) servers. Usually, legacy x86 architecture are used, and in BeGREEN we examine using other novel architectures. We have looked into the most computationally complex algorithms in the DU, which are the Low-Density Parity Check (LDPC) decoder and the sphere decoder. The power consumption for these algorithms can be significant, and in BeGREEN we are suggesting methods of lower them by optimizing the algorithm, the implementation and introducing new architectures. In BeGREEN D3.2 [1] we have shown how running the LDPC algorithm on an ARM-based platform leads to ~20% power reduction in average compared to legacy x86 platforms. Here, we will show results when accelerating the LDPC algorithm on a Graphical Processing Unit (GPU).

2.1 GPU DU acceleration architecture

The architecture algorithm being developed in BeGREEN as well as the GPU-based architecture were described in [5]. The Log likelihood Ratios (LLRs) that are the output of the demapper represent the estimated bits, where their sign represents the bit value, and their amplitude represent the confidence. The BeGREEN LDPC decoder is a “soft” decoder, where the confidence levels of the LLRs are used to improve the decoding performance and to be able to know how to treat the demapped bits. These LLRs are then being processed within the LDPC decoder by the belief propagation message passing algorithm, where the messages are iteratively being passed between the check nodes and bit nodes. The 5G LDPC code is described in the 3GPP standard [6]. There are two base graphs that define the two parity check matrices. Each entry of these parity check matrices is not a single bit but represents a square matrix to accommodate various block sizes. The size of the matrix is $Z \times Z$, where Z is called the lifting size, and the matrix is a cyclic shifted unity matrix. There are 51 options for the lifting size Z , from 2 to 384. This structure allows efficient parallelization according to the lifting size, as the calculations there are independent and can be done simultaneously. In addition, the main Transport Block (TB) for each UE is composed of several Code Blocks (CBs), and different UEs will have different TBs to carry their information. The BeGREEN LDPC decoder parallelizes according to the lifting size and between CBs. The intent is to be able to utilize the full capacity of the GPU, which can perform a very large number of similar operations concurrently. The general concept of the BeGREEN implementation is that the GPU program is multithreaded to utilize the capabilities of the GPU where all threads share the same functionality, and each thread is responsible to process a different data segment. When running the LDPC on GPU we expect either the LLRs to be computed on a CPU when using a low complexity Minimum Mean Square Error (MMSE) receiver, or on a GPU when using a highly complex sphere decoder receiver. The interface between the host (the CPU) and the device (the GPU) in the case where the LLRs are computed in the CPU is shown in Figure 2-1. The Host and the device have separate memory systems. Hence, to be accessible, data must be transferred between host memory and device memory. Note that accessing global memory is slow and constant memory or shared memory are preferred. This is shown in Figure 2-2.

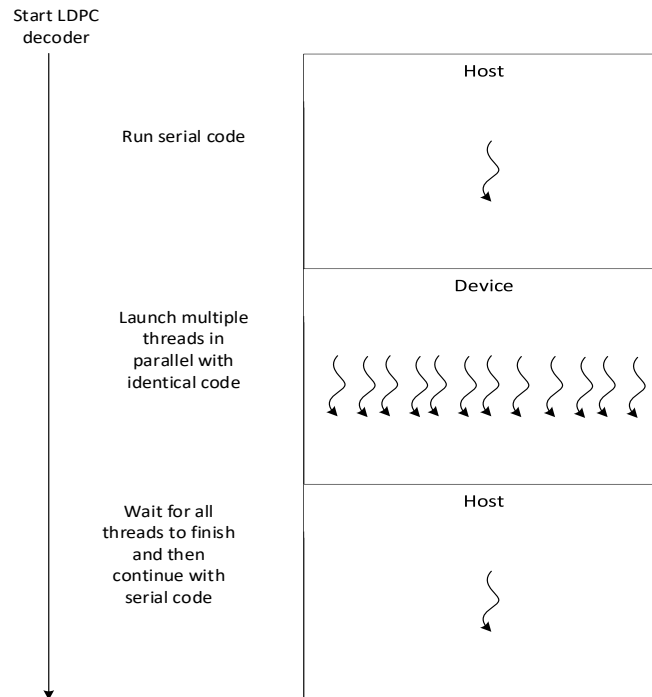


Figure 2-1 LDPC threading concept

Compute Unified Device Architecture (CUDA)¹ is a proprietary parallel computing platform and application programming interface (API) for accelerated general-purpose processing over Nvidia² GPUs. In the LDPC development within BeGREEN, two approaches for CUDA multithreading were investigated. The first is to launch threads which contain minimal functionality, with the advantage that threads with complex functionality tend to be more time consuming. The other is to launch threads with more significant functionality, allowing the use of shared memory instead of global memory, which reduces time overhead of memory access and reducing the number of threads launches and overhead. There is a trade-off between these two approaches, being a priori not clear which approach is better in terms of utilization. Measuring the two solutions indicates that second approach is faster in all scenarios.

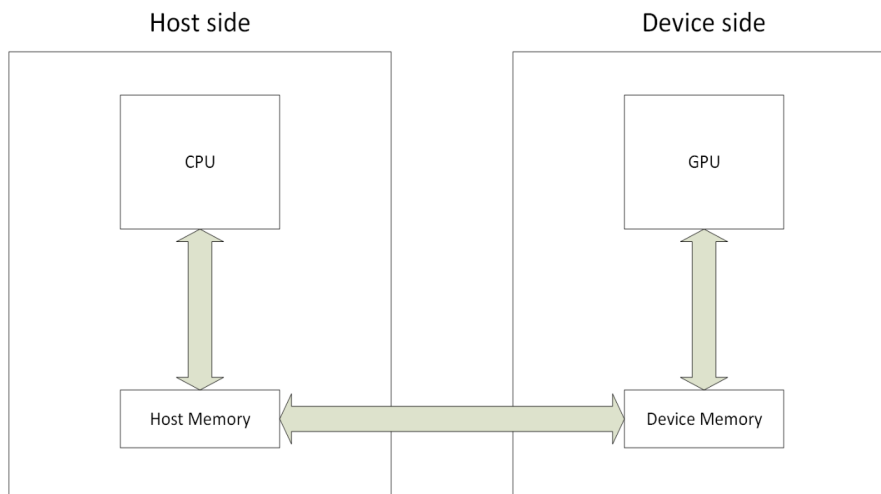


Figure 2-2 CPU/GPU memory management

¹ <https://developer.nvidia.com/cuda-toolkit>

² <https://www.nvidia.com/>

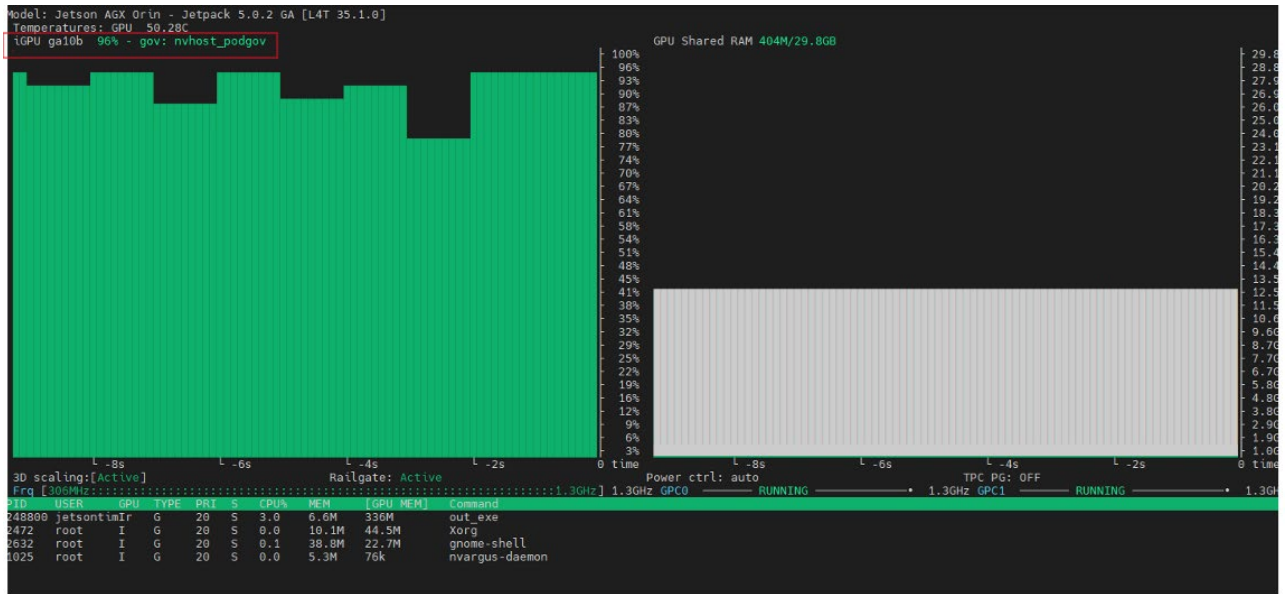


Figure 2-3 GPU load measurement

2.2 DU LDPC decoder research results analysis and benchmarking

The BeGREEN LDPC GPU implementation strive to fully utilize the GPU. An example GPU load when running the decoder is shown in Figure 2-3, where we can see that it is almost fully utilized (96%). We have checked how many threads can be run in parallel while still improving utilization. We have observed that, after a threshold of 32 CBs and the maximal lifting size, utilization stops to improve. GPU is fully loaded at this point and additional threads are pending in queue. This is supported by jetson-stats that indicate loading state of the GPU is close to 100%. Hence, above the threshold running more threads in parallel **increases** average power consumption. Therefore, it was found that for large CBs (high parallelism within the CB), processing as much as 32 CBs in parallel is preferred, and for small CBs (low parallelism within the CB), processing more than 32 CBs in preferred, potentially the whole transport block.

For the purpose of evaluating the usage of a GPU for processing the LDPC decoder and comparing it to the ARM Marvel Octeon 10 platform presented in [1], we have used the NVIDIA Jetson AGX Orin 64GB³ platform that includes a 2048-core NVIDIA Ampere GPU with 64 Tensor Cores. The same 8 scenarios that were tested in [1] were now also run on the GPU. The parameters for these scenarios are shown in Table 2-1.

Table 2-1 LDPC Scenarios Parameters

#	Modulation	LDPC Code Rate	#RBs	TB size (bits)	# of CBs per TB	CB size (bits)	Lifting size Z	Base Graph
1	QPSK	193/1024	273	14856	4	3714	114	2
2	16QAM	658/1024	273	100392	12	8366	109	1
3	64QAM	567/1024	273	131176	16	8199	113	1
4	256QAM	682.5/1024	273	208976	25	8359	109	1
5	64QAM	567/1024	10	4736	1	4736	87	1
6	256QAM	682.5/1024	3	1128	1	1128	99	2
7	QPSK	193/1024	10	552	1	552	87	2
8	QPSK	193/1024	1	48	1	48	58	2

³ <https://www.nvidia.com/en-us/autonomous-machines/embedded-systems/jetson-orin/>

Results comparing ARM and GPU energy scores for each of these scenarios where GPU is running 32 CBs in parallel and 128 CBs in parallel are shown in Table 2-2 and Table 2-3 respectively. The Octeon in these runs use 23 out of its available 24 (as one core is used for the operating system).

Table 2-2 LDPC Processing Comparing ARM and GPU Running 32 CBs in Parallel

#	GPU Max Load (%)	ARM SVE2 Energy kbits/Joule	GPU Energy Score kbits/Joule	GPU to ARM Ratio (%)
1	96	2994	474	16
2	97	8306	1403	17
3	97	7027	1209	17
4	97	9026	1502	17
5	96	6694	1096	16
6	67	9982	1297	13
7	87	9770	259	3
8	66	1039	60	6

Table 2-3 LDPC Processing Comparing ARM and GPU Running 128 CBs in Parallel

#	GPU Max Load (%)	ARM SVE2 Energy kbits/Joule	GPU Energy score kbits/Joule	GPU to ARM ratio (%)
1	98	2994	408	14
2	98	8306	1308	16
3	99	7027	1085	15
4	98	9026	1429	16
5	98	6694	1084	16
6	80	9982	1187	12
7	93	9770	400	4
8	70	1039	158	15

When using all its resources, ARM SVE2 LDPC decoder implementation running on Marvel Octeon significantly outperforms GPU implementation in terms of EE. We can see that, for the larger TB sizes, the difference is smaller. Also, for the smallest TB sizes running 128 CBs on the GPU in parallel gives some improvement over 32 CBs, but in more cases 32 CBs perform better.

Using 23 cores of Octeon for LDPC decoder processing it is ~6 times more energy efficient than Jetson (maximum load case). However, when comparing the GPU to just using 1 ARM core the GPU is much more energy efficient. Usually, the platform will have GPU and some CPU cores (like ARM cores) and, in this case, we will be able to run the LDPC over the GPU leaving many CPU cores for many other purposes. Hence, GPUs can be efficiently utilized for LDPC when available.

2.3 CU/RIC acceleration research results analysis and benchmarking

Following the initial system setup and measurements in deliverable D3.2, the next phase of the study focuses on comparing and testing ARM with x86 platforms for the implementation of the RIC to perform a series of comparative analyses. The primary objective is to assess the performance and power consumption of applications running on these platforms within the RIC environment.

The research postulates that the complexity of xApps directly correlates with increased power consumption. To validate this hypothesis, two testing scenarios were designed: one focusing on ARM performance testing and the other on latency analysis. These tests aim to evaluate the impacts of performance and latency on the system's capabilities.

The study outlines detailed procedures for conducting tests within the ARM environment and compares these with x86-based results. By analyzing the findings, the research provides insights into the performance characteristics, scalability, and EE of the platforms, laying the groundwork for further system enhancements.

2.3.1 ARM testing - performance

The xApp is already built for both ARM and x86 platforms. The xApp can be deployed and then activated over the xApp API. It runs two operations: matrix multiplication and hashing. Operational tests can be triggered, and their duration specified. Multiple experimental tests can also be scheduled manually. The xApp communicates directly with the dRAX Influx Database (InfluxDB), monitoring CPU and RAM usage, as well as the execution time of the operation, storing the results in InfluxDB.

A Grafana Dashboard has been built for the xApp, which is available for download via the xApp API and can be imported into Grafana. This Grafana Dashboard queries InfluxDB to display CPU and RAM usage results. It also monitors the logs of the xApp, providing traceability for when tests start, complete, or are stopped manually using the API.

- **Operational Modes:** The xApp tests performance in terms of CPU and RAM usage, with two modes of operation: a) Matrix Multiplication b) Hashing.
- **xApp APIs:** Performance tests are triggered using the xApp API. The xApp API includes three endpoints:
 - POST /api/start
 - POST /api/stop
 - GET /api/status
- **Results:** Test results are written into the dRAX KPM/PM Time Series (InfluxDB) Database, using the *db_5g* database and the *performance metrics* measurement label.
- **Grafana Dashboard:** The xApp includes a Grafana Dashboard to visualize the results. The JavaScript Object Notation (JSON) configuration for the dashboard can be downloaded via the xApp API endpoint.

```
GET /api/grafana
```

Example:

```
curl -X GET http://10.0.100.70:30248/api/grafana -o grafana_dashboard.json
```

To import the dashboard into Grafana, navigate to the dRAX Grafana web page at <drax-host>:30300. Login using the provided credentials, click on the left-hand menu, and select Dashboards. Click the "New" button and select "Import" as the option. Paste the full JSON from the xApp into the form on Grafana and click "Load." Finally, click the "Import" button to complete the process. The xApp Grafana Dashboard will then be available in the Dashboard menu as shown in the Figure 2-4.



You can check the status of your xApp in the **xApp Overview** page of the **dRAX Dashboard** as shown in the Figure 2-6 below. Make sure it is in **Running** state and note the **NodePort** for the **xapp-api** Service.

Figure 2-5 Implementation of the xApp via dashboard

Figure 2-6 dRAX dashboard showing the implemented xApp

1. Click Submit and check the xApp status on the xApp Overview page of the dRAX Dashboard.

Deploying on Specific Kubernetes Nodes

Using Affinity:

Prepare a `values.yaml` file with the following content:

```
affinity:
  nodeAffinity:
    requiredDuringSchedulingIgnoredDuringExecution:
      nodeSelectorTerms:
        -
          -
            operator: In
            values: ["amd64"]
            key:
              matchExpressions:
                "kubernetes.io/arch"
```

For ARM architecture, change the values to:

```
values: ["arm64"]
```

Upload the `values.yaml` file along with the Helm Chart during deployment.

Use values.yaml on xApp Deployment

To deploy the xApp, navigate to the xApp Deployment page on the dRAX Dashboard and click the xApp Deployment button. From the Deploy Method drop-down menu, select Upload Helm Chart and provide the required metadata, including the xApp name. Under the Upload Helm Chart section, click the Choose a file button to upload the `.tgz` xApp Helm Chart file. If applicable, upload the `values.yaml` file by clicking the Choose a file button next to the Optional Values file field. Once all the necessary files are uploaded, click the Submit button to complete the deployment process.

Example API Requests

To discover the xApp API, access the xApp Overview page on the dRAX Dashboard, click the Show button in the Services column for the xApp, locate the `xapp-api` service, and note the NodePort.

Start a Fixed-Duration Matrix Multiplication Test

This test runs a matrix multiplication workload with a matrix size of 2000x2000 for 5 minutes (300 seconds) across 4 processes.

```
curl -X POST http://10.0.100.70:30248/api/start \
-H "Content-Type: application/json" \
-d '{
  "workload": "matrix_multiplication",
  "matrix_size": 2000,
  "duration": 300,
  "processes": 4
}'
```

Start a Hashing Test with 50 Iterations

This test runs a hashing workload where each task hashes 20 MB of data for 50 iterations using 2 processes.

```
curl -X POST http://10.0.100.70:30248/api/start \
-H "Content-Type: application/json" \
-d '{
  "workload": "hashing",
  "data_size": 20,
  "iterations": 50,
  "processes": 2
}'
```

Check Test Status

```
curl -X GET http://10.0.100.70:30248/api/status
```

Stop the Test

```
curl -X POST http://10.0.100.70:30248/api/stop
```

Testing Matrix Multiplication:

This section describes three series of matrix multiplication tests designed to evaluate computational performance across different configurations. The tests include a single-threaded baseline efficiency test, a multi-threaded scaling performance test for both ARM and x86 architectures, and a larger matrix stress test to assess memory bandwidth and peak computational capacity. Each test provides insights into the efficiency, scalability, and limits of the platforms under varying workloads.

Single-Threaded Test (Baseline Efficiency):

- Matrix Size: 5000.
- Duration: 5 minutes (300 sec).
- Processes: 1 (single-threaded).
- Purpose: Compare raw computational efficiency of a single core.

```
curl -X POST http://10.0.138.31:31853/api/start \
-H "Content-Type: application/json" \
-d '{
  "workload": "matrix_multiplication",
  "matrix_size": 5000,
  "duration": 300,
  "processes": 1
}'
```

Multi-Threaded Test (Scaling Performance):

- Matrix Size: 10,000.
- Processes:
 - ARM: 32 (to utilize all cores).
 - x86: 16 (to utilize all physical cores or 32 to use hyper-threading).
- Duration: 10 min (600 sec).
- Purpose: Measure how well each platform scales with multi-threading.

For ARM:

```
curl -X POST http://10.0.138.31:31853/api/start \
-H "Content-Type: application/json" \
-d '{
  "workload": "matrix_multiplication",
  "matrix_size": 10000,
  "duration": 600,
  "processes": 32
}'
```

For x86:

```
curl -X POST http://10.0.138.31:31853/api/start \
-H "Content-Type: application/json" \
-d '{
  "workload": "matrix_multiplication",
  "matrix_size": 10000,
  "duration": 600,
  "processes": 16
}'
```

Larger Matrix Stress Test (Peak Performance):

- Matrix Size: 20,000 or higher.
- Processes:
 - ARM: 32.
 - x86: 32 (to use hyper-threading).
- Duration: 15 min (900 sec)
- Purpose: Stress-test memory bandwidth and peak computational throughput.

```
curl -X POST http://10.0.138.31:31853/api/start \
-H "Content-Type: application/json" \
-d '{
  "workload": "matrix_multiplication",
  "matrix_size": 20000,
  "duration": 900,
  "processes": 32
}'
```

2.3.2 ARM testing - latency

This xApp tests the latency of receiving messages from the dRAX Databus (Kafka) from the dRAX RAN (CU). The xApp writes the results into the dRAX KPM/PM Time Series Database (InfluxDB). It uses the db_5g database and the latency_metrics measurement label. The latency is calculated for every message and stored in microseconds along with the topics on which the message was received.

The stored data in InfluxDB is as follows:

```
{
```

```

    "measurement": "latency_metrics",
    "tags": {
      "platform": platform_architecture,
      "topic": topic
    },
    "fields": {
      "latency": latency
    }
  }
}

```

xApp Configuration

The xApp uses xApp configuration to enable or disable the logging of data to InfluxDB. Configuration can be modified on the dRAX Dashboard by clicking on the Details button of the xApp. Expand the Configuration Parameters and locate the log_to_influx configuration option. Set it to "Yes" to enable logging.

xApp Grafana Dashboard

The xApp also includes a dRAX Grafana Dashboard, which can be imported into Grafana to visualize the results.

The JSON for the Grafana Dashboard can be downloaded using the xApp API endpoint:

```
GET /api/grafana
```

Example:

```
curl -X GET http://10.0.100.70:30248/api/grafana -o grafana_dashboard.json
```

To import the dashboard into Grafana, navigate to the dRAX Grafana web page at <drax-host>:30300. Login using the provided credentials, select Dashboards from the left-hand menu, and click the "New" button. Select the "Import" option, paste the JSON from the xApp into the form, and click "Load." Finally, click "Import" to complete the process.

The xApp Grafana Dashboard is now imported and can be accessed by selecting it from the Dashboards menu as shown in Figure 2-7.

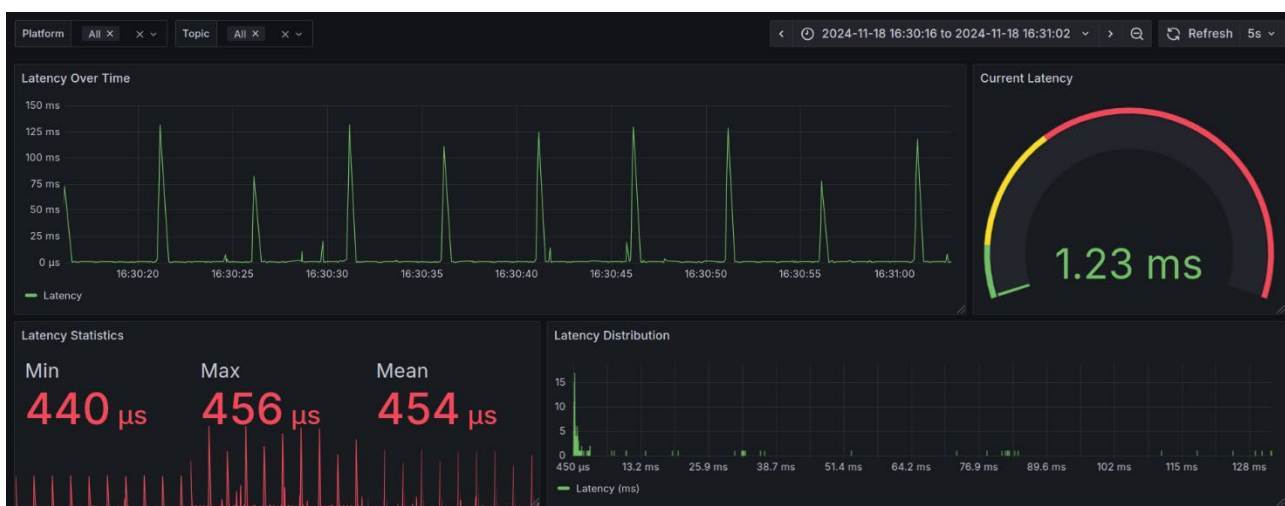


Figure 2-7 Grafana for latency testing

2.3.3 Results

The results presented in this section are derived from the ARM testing detailed previously in Section 2.3.1. In terms of average power consumption, the ARM server demonstrates significantly better efficiency compared to the x86 server. On average, the ARM server consumes one-third less power than the x86, meaning the x86 is approximately 50% more power-hungry. The power consumption varies notably with the size of matrix multiplications. As shown in Figure 2-8 (left), the power consumption for the ARM server increases from an average of 140 watts to 175 watts when processing larger matrix sizes. Similarly, the x86 server sees its power usage rise from an average of 210 watts to approximately 276 watts under the same conditions, showing an increment of 25% in the ARM server and 31% for the x86. Additionally, in Figure 2-8 (right) the distribution of power measurements is presented. Here the power of the x86 is much more spread than the one for the ARM server showing increments up to 66% more than the average values.

For total CPU usage, the comparison reveals an advantage for the x86 server when performing smaller computational tasks, as it consumes less CPU power. However, this efficiency gap diminishes as the matrix size increases. At higher matrix sizes of 10,000 and 20,000, the CPU usage of the x86 server rises significantly, aligning more closely with that of the ARM server. It can be seen that, for lower matrix sizes, the CPU utilization is inferior in the x86, with almost 50%, but it rapidly increases to 100% for higher matrix sizes, while the ARM increases only slightly, remaining just under the 100% mark, as shown in the Figure 2-9 (left). Similarly to the CPU usage comparison, the memory usage exhibits a comparable pattern as presented in Figure 2-9 (right). At lower matrix sizes, the x86 server slightly outperforms the ARM server. However, as the matrix sizes increase to 10,000 or 20,000, the ARM server demonstrates slightly better performance compared to the x86 server. Additionally, the memory usage scales proportionally with the matrix sizes, reflecting consistent system behavior as workload demands increase.

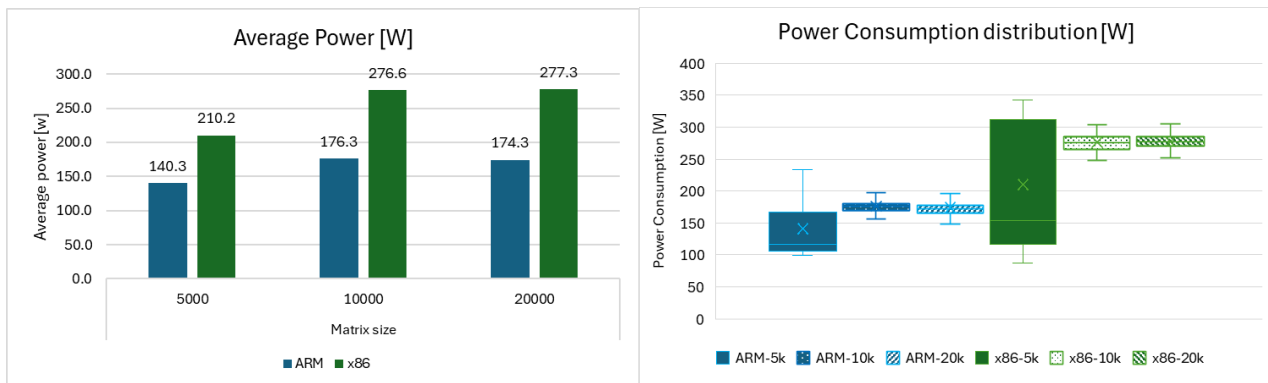


Figure 2-8 Power consumption comparison for ARM and x86. (left) Average (right) Distribution

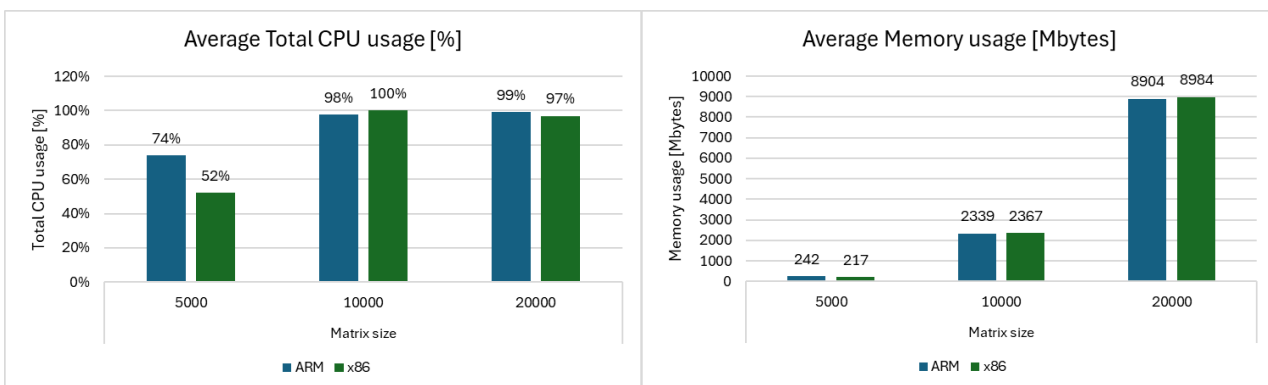


Figure 2-9 Average comparison for ARM and x86. (left) CPU usage. (right) Memory usage

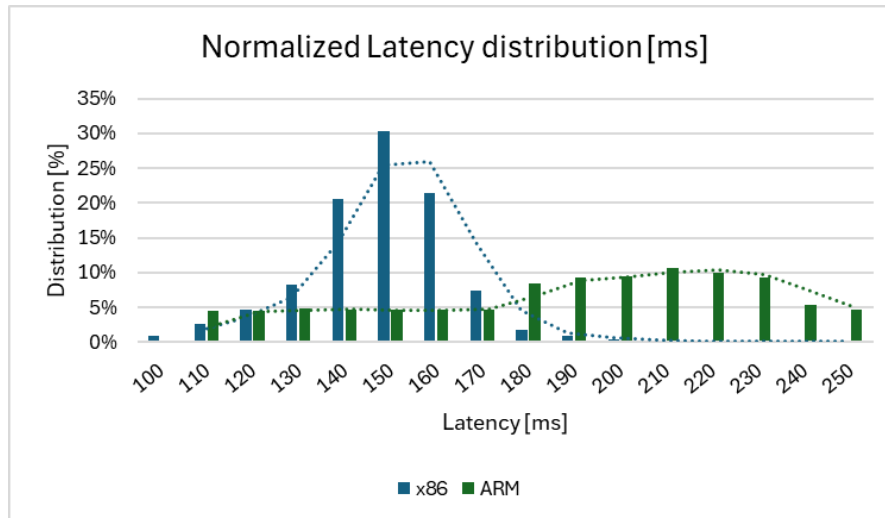


Figure 2-10 Latency distribution and comparison for ARM and x86

In the next test, the ARM and x86 platforms are compared in terms of latency when serving up to 64 UEs in the system. The two xApps defined in the previous section are evaluated, and the results are presented here. The x86 server demonstrates an average latency of 143 milliseconds, with a tightly clustered distribution around this value. In contrast, the ARM server shows a more dispersed latency distribution, with an average value of approximately 225 milliseconds. As illustrated in Figure 2-10, the x86 server consistently achieves lower latency, while the ARM server averages slightly higher at 200 milliseconds. This difference is attributed to the implementation of the xApps on each server and the way resources are utilized by the respective servers.

2.4 Collaborative insights

When evaluating the usage of a GPU for running the CU acceleration, it was found that there is not a high potential the GPU will be efficient for this purpose. As shown in the previous subsections, when comparing the DU algorithms on GPU to the full number of cores, we can also see that for DU acceleration this is the case. Hence, both the CU and the DU indicate that when looking into specifically just the power consumption metric, CPU processing is favourable to GPU processing. However, as already discussed, GPU has advantages (i.e. offloading from the CPU and leaving it free for other parts of the receiver).

3 RU Energy Consumption Optimization Research Results

This section examines the results of RU management research and power consumption reduction strategies, focusing on xApp/rApp performance, EE, and benchmarking within the BeGREEN architecture. The integration of PoC1 algorithms into the PoC3 setup is analysed, with a detailed comparison of ARM and x86 servers in terms of power consumption and performance. Key RIC components, including the Energy Saving (ES) xApp, Smart Handover (SHO) xApp, Telemetry Gateway (TGW), and InfluxDB, are evaluated to identify factors contributing to server power consumption and inform future optimizations. Additionally, the study delves into RU power consumption reduction methods, such as PA blanking, and highlights their effectiveness in reducing energy use during non-allocated symbols. Through testing scenarios with varying cell and UE counts, this section provides valuable insights into energy-saving policies, server behaviour, and architectural improvements for optimal 5G network operations.

3.1 RU Management (xApps/rApps) research results analysis and benchmarking

To integrate PoCs in the context of the BeGREEN architecture, the setup described for PoC3 was tested with the algorithms of the PoC1. In particular, the implementation of the RIC on the ARM and x86 servers and its power consumption values is compared with the ES algorithms described in PoC1 for different policies, where ES percentage policies are sent towards the Near-RT RIC to control the RAN based on these policies (please see deliverable D4.2 Section 2.1 [14], and deliverable D5.2 Section 3.1.1 [15]). The testing aims to identify the elements that make the server to consume more power to define future optimization and, therefore, reduce server power consumption. In particular, the RIC items that are evaluated here are the ES xApp, the SHO xApp, the TGW and the InfluxDB, as sketched in Figure 3-1.

3.1.1 Test Definition

The test uses 2 A1 policies to modify the behaviour of the ES xApp, the first one issues a 100% energy savings for a set of Cell, and the second one issues 0% energy savings for the same set of cells. Two use case where investigated, a) 6 Cells and 20 UEs with a policy scope of 4 Cells, and b) 20 Cells and 100 UEs with a scope of 15 Cells, as presented in Figure 3-2 (up). The power consumption of the servers is measured via internal metric sensors and exported to Grafana for easily plotting and analysis as shown in Figure 3-2 (down). The tests measure the power consumption, when all the cells are active, then when the cells in the policy are off and then when all the cells are back on again.

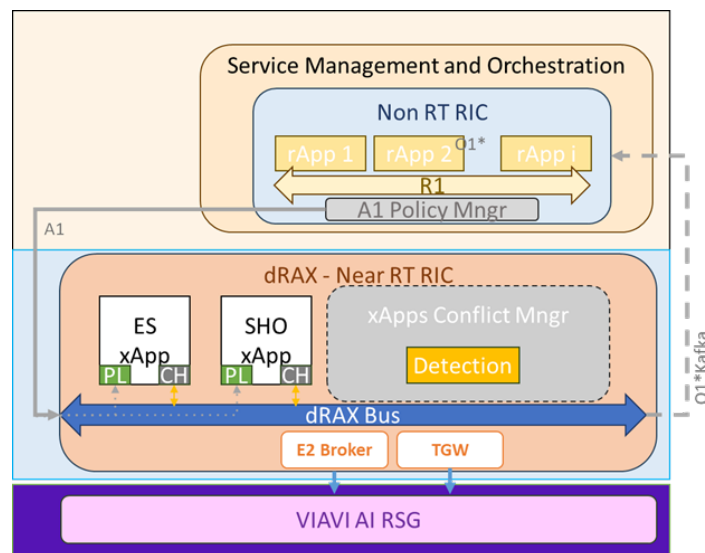


Figure 3-1 xApp/rApp system used for testing including x86 server measurements



Figure 3-2 Scenario for 6 Cells 20 UEs (up) and Scenario for 20 Cells 100UEs (down)

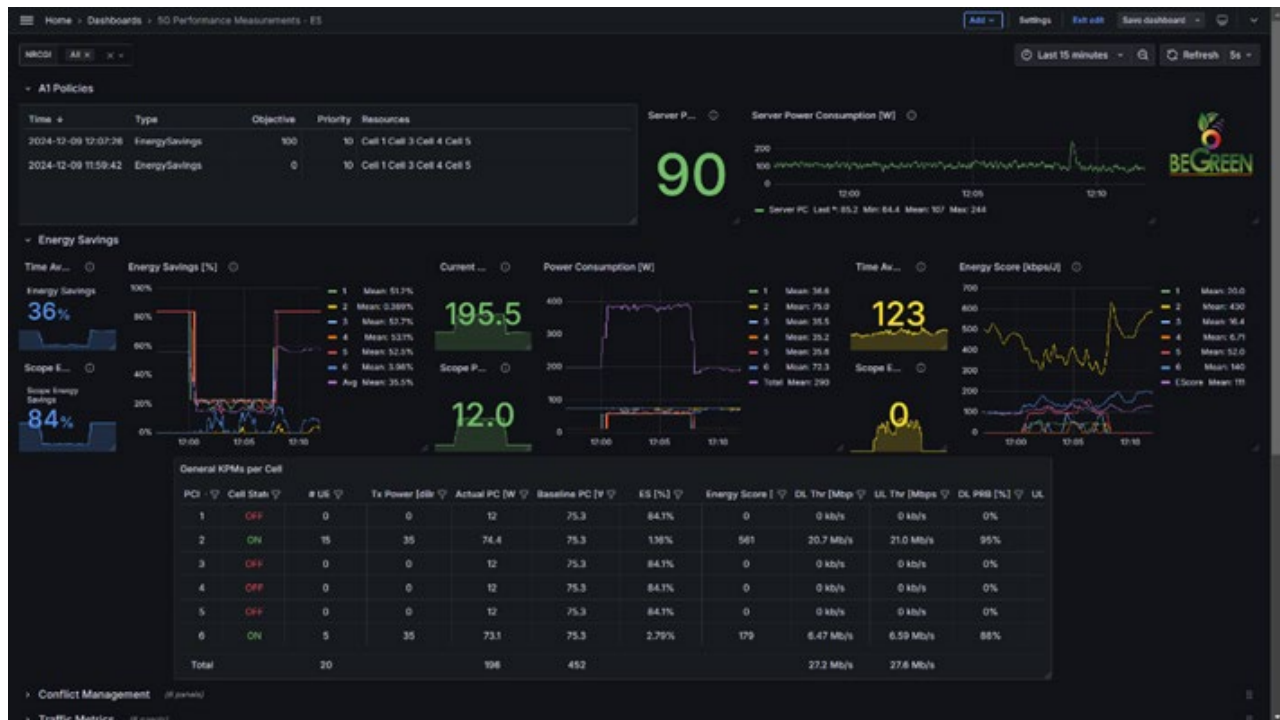


Figure 3-3 General dashboard for the energy savings system including the server power consumption metric

The Grafana dashboard shown in Figure 3-3 is divided into several sections. The A1 policies section shows the policy arrival time, the type of policy (e.g., energy savings), and its objective, which ranges from 0% to 100% of energy savings. It includes the cells or resources affected by the policy, their types, and their priority. The upper right displays server power consumption while running the service. The middle section showcases cell and network metrics. On the left, energy savings for each cell and the total system are displayed in blue. The green section represents power consumption per cell, visible when cells are powered off. The yellow section shows the energy score for each cell and the system average. The bottom section displays general Key Performance Measurements (KPMs) per cell, including the PCI, status, number of connected UEs, transmission power in dBm, power calculation compared to baseline consumption, energy savings (blue section), energy score (yellow section), and traffic metrics such as downlink (DL), uplink (UL), and Physical Resource Block (PRB) utilization.

3.1.2 ARM results

Analysing the CPU percentage utilization of some of the items in the Energy Saving System, the one that most consumes CPU resources is the TGW. Its CPU consumption in the ARM server is the most variable one: it can stay as low as 20% for only 2 Cells on (UC1 ES=100%) where only the Reference Signal Received Power (RSRP) metrics from 2 servers are reported for only 20 UEs, but can rise up to 75% or 90% when all cells are active and the UEs are reporting its RSRP neighbour metrics. Figure 3-4 shows the comparison of the Energy Saving (**green**) and the Server Power Consumption (**yellow**), showing that when the ES is under 20%, the server power consumption is a little higher. Then when the policy of 100% ES is issued, the server power consumption is quite stable in lower values than the average. Also, it is very notable when the Energy Saving is rising (or falling), the server consumption varies drastically. This is due to the fact that the SHO xApp is processing all the new Cell's configuration, activating all the algorithm to move the UEs to better cells, consuming large amounts of energy in short periods of time; i.e., every time that a new cell is powered on or off, the algorithm searches for a new configuration, sends the needed handover messages and then stays on hold until a new configuration is needed. This shows how the ARM server power consumption is reactive to the activities and processes that are running for the Energy Saving system.



Figure 3-4 ARM Server power consumption vs. energy savings metric.

3.1.3 x86 Results

While analysing the server power consumption of the x86 server, it looks like the variations are not related to the Energy Saving system process, since the power consumption values are not affected by the performance of the network. In Figure 3-5, we show how the differences of the power consumption are not related to the energy savings system, and that are quite constant over the whole test. Despite significant variations in CPU usage, the power consumption of the server remains relatively constant during workload execution. This behaviour is likely attributed to advanced dynamic power management features in modern x86 architectures, such as Intel SpeedStep⁴ and AMD Cool'n'Quiet⁵, which optimize voltage and frequency independently of CPU load to ensure EE. Additionally, memory and cache behaviour play a critical role; after a period of intense activity, the system may stabilize memory access patterns and reduce cache thrashing, indirectly lowering energy usage. Consequently, after a certain period of sustained operation, the server's power consumption decreases by 10-15%, reflecting the combined effects of these optimizations as shown in Figure 3-5 at 12:35pm testing time.

⁴ <https://www.intel.com/content/www/us/en/support/articles/000007073/processors.html>

⁵ <https://mytechencounters.wordpress.com/2010/03/15/overclock-your-phenom-cpu-and-save-power-simultaneously/>



Figure 3-5 x86 server power consumption vs. energy saving metric

3.1.4 Comparison and Conclusions

As can be seen from Table 3-1, the ARM server has a variable power consumption depending on the activities from the Energy Savings system. Despite that its general power consumption fluctuates the minimum (119-90W), its differences with and without ES are the higher. On the other hand, the power consumption of the x86 server is much higher in general terms, and fluctuations are independent of the Energy savings application. Hence, for the x86 the influence of the Energy saving is almost depreciable (less than 1%) compared with the ARM case that rise to 5%. Also, the in the ARM server while the ES policy of 100% is in place, the power consumption is less than in the x86, proving the dynamic power allocation of ARM servers.

Table 3-1 Power Consumption ARM vs x86

	Use case 1 (6Cells)			Use Case 2 (20 Cells)			MAX	MIN
	ES 0%	ES 100%	Diff	ES 0%	ES 100%	Diff		
ARM Server	102.5W	99.8W	2.7W	105.4W	100.6W	4.8W	119W	90W
X86 Server	102.7W	102.3W	0.3W	104.3W	103.1W	1.2W	140W	83.2W

This test shows the differences between the ARM and x86 servers and some of internal architecture optimizations. The ARM server is prone to dynamically modify its internal power consumption based on the CPU usage of all the nodes/items that are deployed internally, whereas the x86 cannot modify its power consumption, but has other energy saving mechanisms that can reduce the power consumption in diverse moments of the server usage. Other tests to dive into more details of the internal server's architecture to create optimal xApp algorithms to take advantage of each server capabilities could help to understand and optimize even more the power consumption of each server.

3.2 RU power consumption reduction research results analysis and benchmarking

The RU is the highest power consumer in the cellular network infrastructure. In this chapter we investigate an innovative module called the Power Amplifier (PA) Blanking Module and its ability to substantially improve

the RU power savings. The RU is the 5G RAN key element. It is designed for coverage flexibility to meet diversity of target scenarios.

RunEL Sparq-2025-ORU-3.5G ORUs that will be used for this research provides an adaptable solution, allowing complementary interoperability with other vendors O-RAN 5G devices. The Sparq-2025-ORU-3.5G main features are:

- All-in-one integrated packaging of 5G RF and Baseband (Low PHY) components.
- Full compliance with 3GPP Release 15 Standard.
- Frequency Bands: 3.3GHz to 3.8GHz (n78 5G NR Frequency band) - other bands are optional.
- Supports MIMO 2x2 or MIMO 4x4.
- Beam Forming of up to 4 dual polarized beams.
- Antenna support - model dependent: either four external antennas or one beam forming internal antenna with 4 dual polarized beams.
- Support for internal GPS receiver for TDD synchronization.
- IEEE-1588 synchronization.
- Flexible coverage capabilities – greater coverage area or greater penetration capabilities.
- Small footprint, single-handed quick installation and simple provisioning.
- Fast roll-out for service providers.
- Seamless and cost-effective integration with 5G DU by ORAN Interface (Option 7.2 Category B)
- High performance with Quality of Service (QoS).
- Remote NMS management via RunEL's NMS application

The Sparq-2025-ORU-3.5G ORU perfectly supports compact solution for Private 5G Network deployed in cases which need high performance 5G networks for Ultra-Reliable Low Latency Communication (URLLC) applications.

The Private 5G Network consist of two types of modules:

- Sparq-2025- Network in a Box (NIB).
- Sparq-2025-O-RU.

Figure 3-6 shows the block diagram and end-to-end architecture of the RunEL Private 5G Network that includes the Sparq-2025-NIB and the Sparq-2025-O-RU in two different frequency bands (3.5 GHz and 28 GHz). The system supports the O-RAN split PHY architecture option 7.2 category B with O-RAN Interface between the O-RU and the NIB.

The NIB includes all the 5G network high level components including O-DU, CU, Next Generation Core (NgCore) and Mobile Edge Cloud (MEC) from the High PHY (OSI layer-1) to the Application Layer (OSI Layer-7).

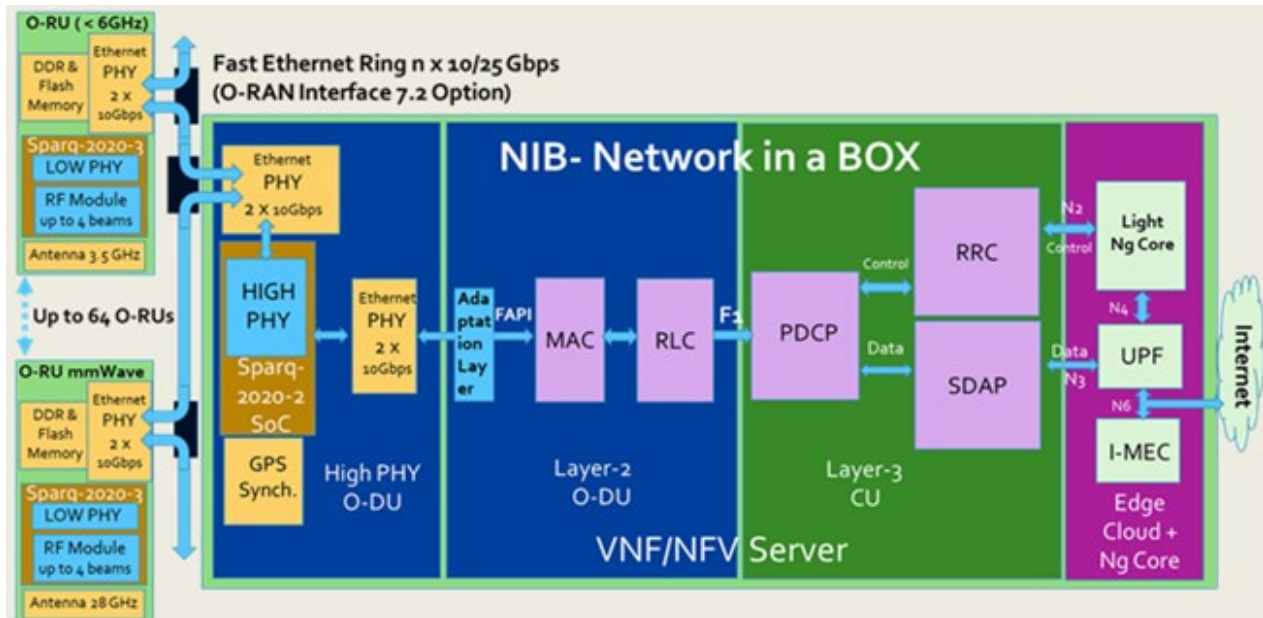


Figure 3-6 Private 5G network architecture

Figure 3-7 illustrates the Sparq-2025-ORU-3.5G components, antenna module will be replaced. The Sparq-2025-ORU interfaces are allocated over two panels. Each of these panels is referred to according to the corresponding interfaces. The GPS, Power and Communication Interfaces are depicted in Figure 3-8.

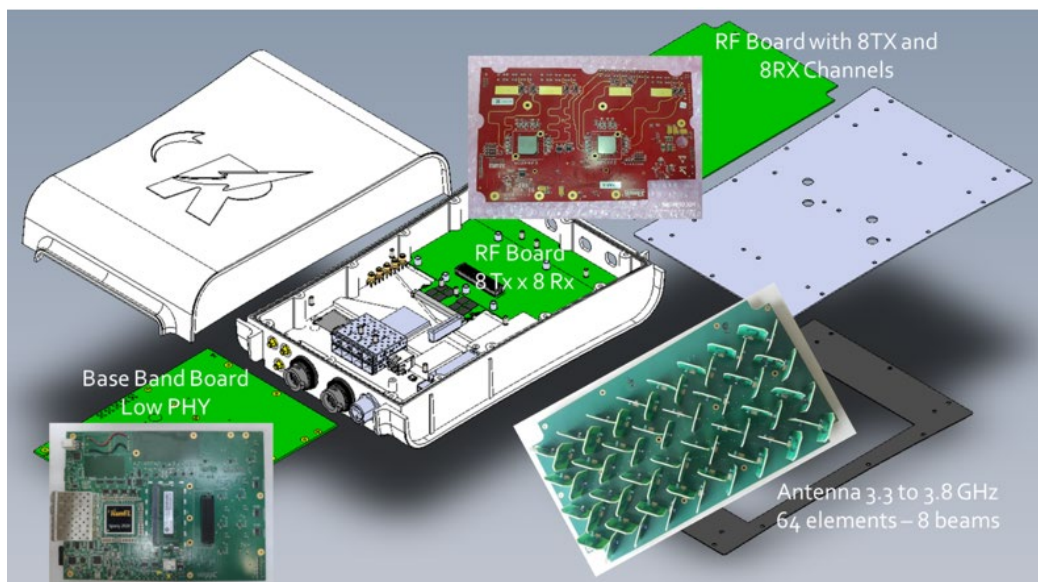


Figure 3-7 Sparq-2025-ORU-3.5G Internal view

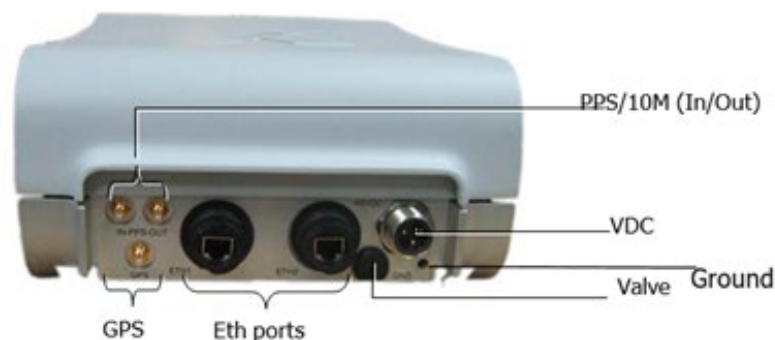


Figure 3-8 Sparq-2025-ORU GPS, Power and Communication Interfaces

The Sparq-2025-ORU interfaces are detailed in Table 3-2.

Table 3-2 Radio Unit (Sparq-2025-ORU) Interfaces

Connector	Description
GPS	Connects to an external (optional) GPS antenna. The GPS antenna is ordered separately. Connector Type: SMA male sealed Industrial
ETH1	Primary Fast Ethernet connector- 8 x 10G SFP+ (Aggregated 40 Gigabit Ethernet connection to Backhaul)- See Note -2 below
ETH2	Second Ethernet port 1G for local and out-of-band management, RJ-45
PPS/10M	PPS In and PPS Out or 10 MHz in or Out (Configurable) can be used for synchronization of multiple sectors. Connector Type: SMA male sealed Industrial
-48/24/12VDC	Power connector. External DC power connector (-48VDC configurable to 12VDC or 24VDC- 2 Amperes) for outdoor deployment. Mating Connector Type: Conec 41-00001
GND	Ground blind hole connector. In normal installations, connect to the pole on which the unit is mounted. (The ORU unit does not include a lightning arrester.)

The system under test, including the DU, CU and Core parts, is detailed in the next section. The DU CU and Core parts are based on Amarisoft software package running on a server. Enabling wireless connectivity through innovative technology.

For this project Amarisoft delivers gNB software. Complementing this software suite is a compact 5G core network. This package content is presented in the next diagram. The Low PHY is implemented by the Sparq-2025-O-RU connected via O-RAN 7.2 interface. The Channel Simulator is not included.

This SW package supports:

- NR 3GPP release 17 compliant.
- Implement one 5GC with built-in AMF, AUSF, SMF, UPF, UDM and 5G-EIR.
- Support several gNodeBs, ng-eNBs or N3IWFs with standard NG interface (NGAP and GTP-U protocols).

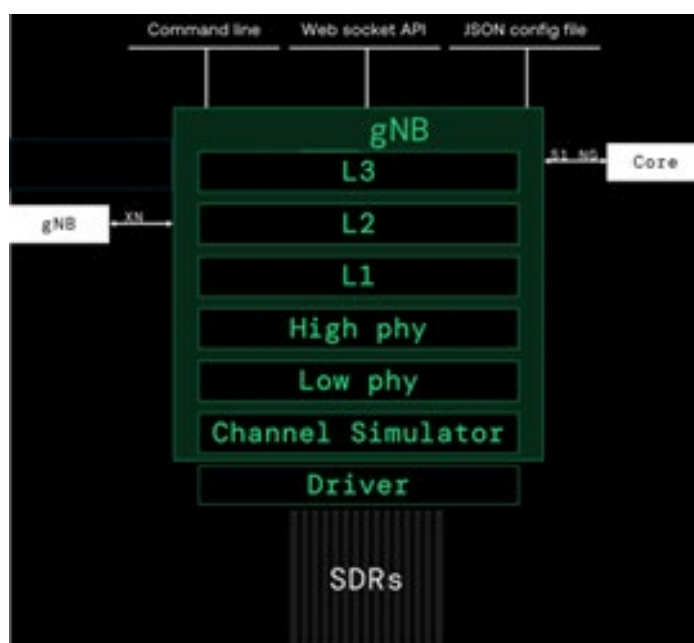


Figure 3-9 Amarisoft SW Package Components

- NAS integrity check and encryption using the AES, Snow3G and ZUC algorithms. Ciphering support is now subject to export rules for your country.
- Support of USIM cards using the XOR, Milenage or TUAK 5G-AKA authentication algorithm.
- Handling of UE procedures: registration, authentication, security configuration, deregistration, service access, radio bearer establishment, paging.
- Multi PDU sessions support and built-in dynamic QoS flow setup for easy VoNR/IMS testing.
- Transparent access to the IP network (no external UPF is necessary).
- Configurable access point name, IP range, DNS and QoS flows.
- IPv4, IPv4v6, IPv6 and unstructured PDUs support.
- Configurable logging system for all channels with built-in text decoders.
- Remote API using WebSocket.
- Command line monitor.
- MICO, active time and eDRX support.
- Supports several IMS servers with Rx interface.
- Support of NB-IoT, LTE and non-3GPP RAT.
- User management via internal database without any external HSS.
- Support of broadcast and multicast PDU session options.
- Support sending of Public Warning System messages (ETWS/CMAS).
- Support of N12 interface with external AUSF.
- Support of N8 and N13 interface with external UDM.
- Support of N17 interface with external 5G-EIR.
- Support of N20 interface with external SMSF.
- Support of N50 interface with external CBC.
- Support of N62 interface with external AF.
- Support of network slicing.
- Support of control plane Clot 5GS optimization.
- Non-IP data delivery Clot feature.
- Support of NL1 interface.
- Support of N5 interface with the following restriction: IMS_SBI feature is declared by the PCF, credit management is not supported.
- Support of Ethernet PDU sessions.

3.2.1 Power consumption reduction

Several power consumption reduction methods were implemented. The RU PA Blanking method that was implemented and demonstrated in a lab test and the AI based DPD and ET methods that were verified by extensive simulations (look at BeGREEN D3.2, Section 3.2.2.4 [1]).

Figure 3-10 presents the PA blanking test-set.

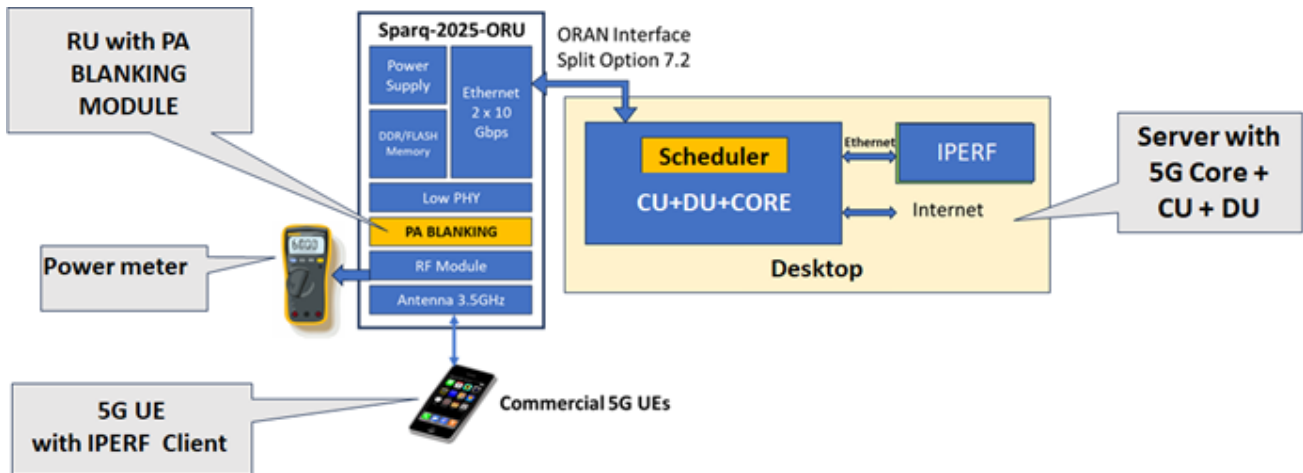


Figure 3-10 PA blanking test setup

The RU under test is connected to the Server which carries the DU, the CU and the Core software packages. The connection is via O-RAN 7.2 interface. The whole assembly implements a 5G BS. The BS is connected to a commercial UE that was used as the edge device User Equipment (UE). An iPerf tool, installed over the UE (client) and the 5G Core (server), was used to stream data over the link from the 5G BS system to the EU.

A power meter was used to check the effectiveness of the PA blanking running at the RU under test.

By September 2024 the test was carried out at RunEL lab. Figure 3-11 screenshots were taken during the tests at RunEL.

The RU, at the drawing left side, is connected on its right side to the CU-DU-Core via O-RAN7.2 interface and via 5G air interface to the UE.

The PA blanking mechanism strives to concentrate the allocated Resource Elements Blocks (REB) at a set of symbols leaving the rest of the slot symbols without allocated REB's. This method is illustrated in Figure 3-12.

At the time of non-allocated symbols, the linear RF PA is blanked to save energy consumption. A fast-switching mechanism implemented at the RU makes the transition from an active symbol to a non-active symbol very fast. As can be seen from Figure 3-13, over 90%, of the energy at non-allocated symbols time can be saved. For maximum energy saving the DU Scheduler arranges the down link data REB's (Resource Blocks) in active symbols trying to leave maximum non-allocated symbols at an active slot.

Figure 3-14 below depicts this principle. As can be observed after arranging the slot allocations 6 more symbols can be blanked to save energy. In Table 3-3 the results from the lab measurements of the PA Blanking Module are shown.

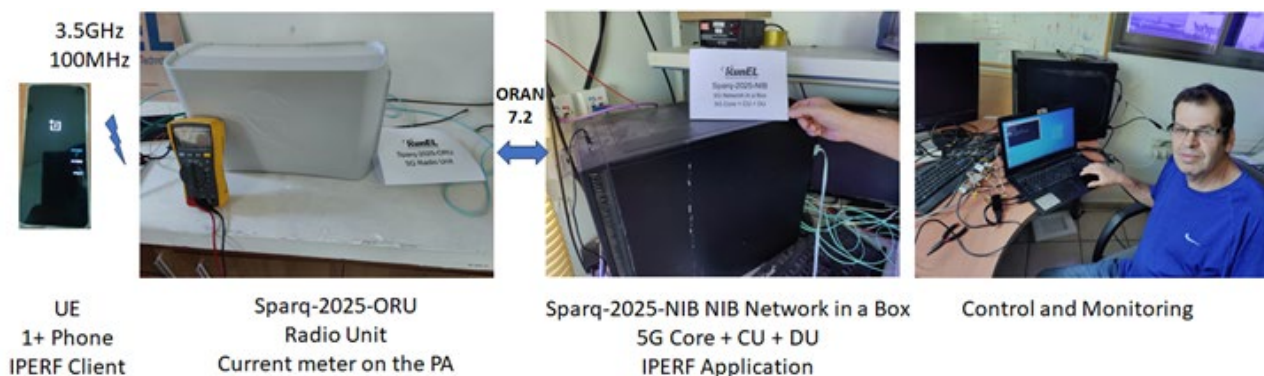


Figure 3-11 RU PA blanking test at RunEL Lab

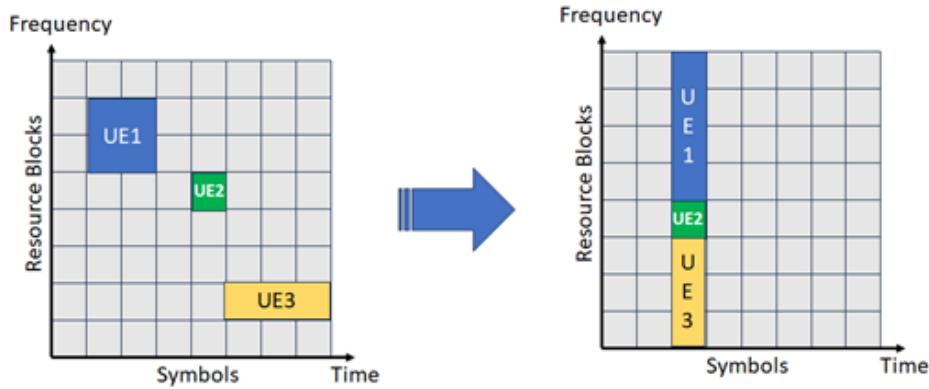


Figure 3-12 PA blanking method

33 Microsec = symbol at 30KHz Subcarrier spacing
 PA switching time < 1 Microsec
 93% of the time blanking

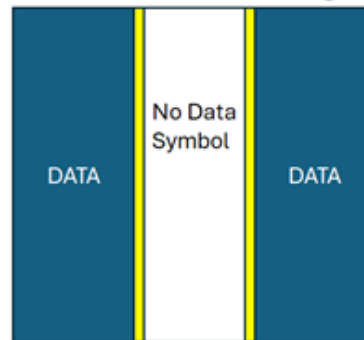


Figure 3-13 Energy save at a blanked symbol

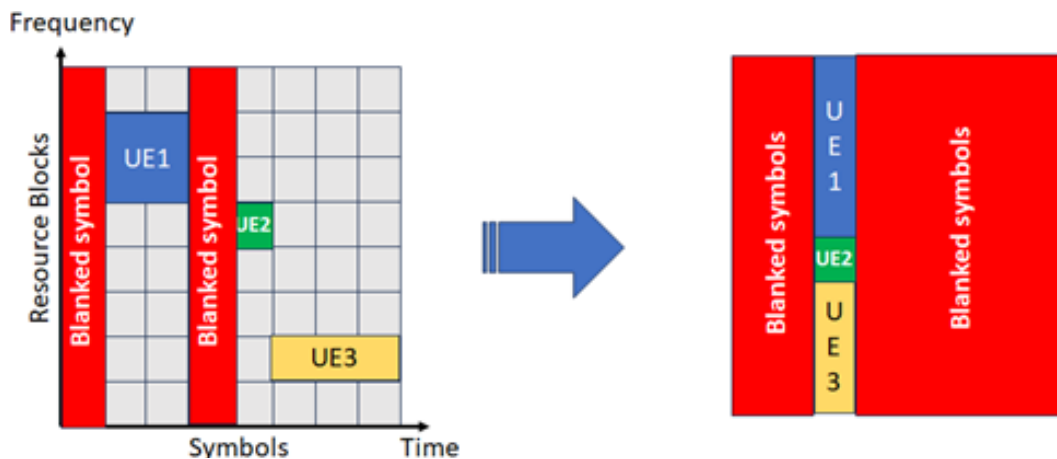


Figure 3-14 Slot before and after REB's Arrangement

As can be observed at low traffic load the saved energy is over 80%. For busy hour traffic the saved energy is over 30% and for average traffic load we can save over 50%. The above traffic assumptions are based on the following traffic distribution⁶ charts.

⁶ Google Analytics by Jonathan Ellins – November 2024

Table 3-3 PA Blanking Tests Results

	#	DL Throughput (Mbps)	PA Current with PA Blanking (mAmps)	PA Current without PA Blanking (mAmps)	RU Power with PA Blanking (Watts)	PA Power without PA Blanking (Watts)
Low Traffic 6h AVG Traffic 12h Peak Traffic 6h	1	No Data	24	250	0.48	5
	2	10	40	250	0.8	5
	3	100	120	260	2.4	5.2
	4	1000	214	311	4.28	6.22

The Peak to Low Traffic ratio in Dense Urban- Western Europe is ~10. As presented above the main task was implemented at the DU scheduler setup. To minimize the energy consumption at a non-allocated symbol the RU linear Power Amplifier operation is stopped for the blank time. To accomplish this task, we used the RU RF PA pin 16 blank control as indicated in Figure 3-16.

The PA switching time is less than a microsecond which yields effective blanking at 93% of the symbol time at 0.5 millisecond slot (at 30KHz SCS).

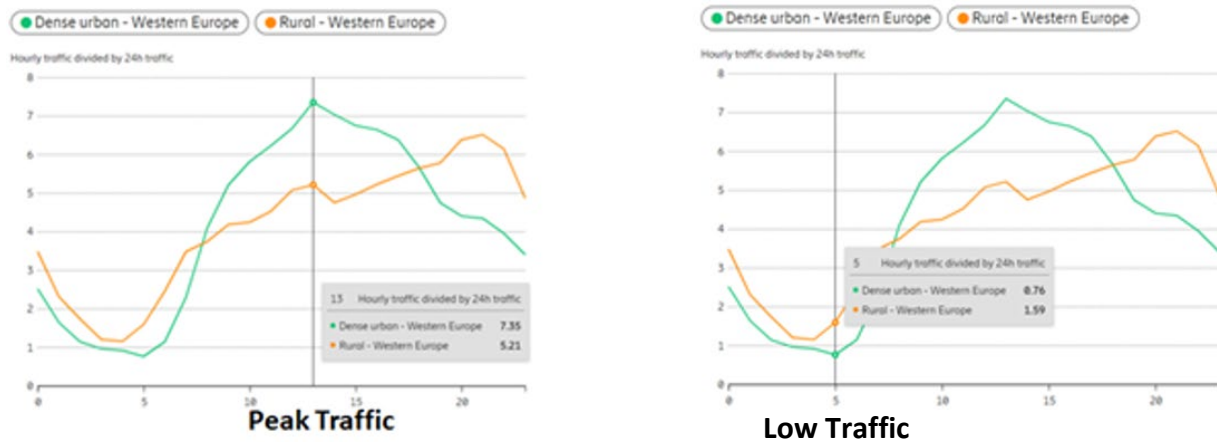


Figure 3-15 Western Europe traffic Distribution

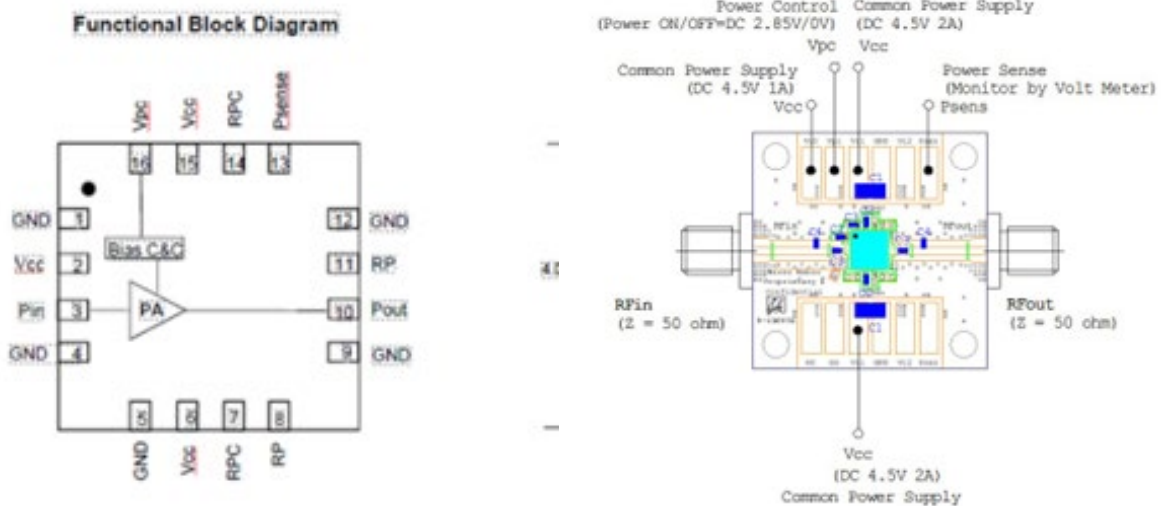


Figure 3-16 RF PA Block Diagram and Physical Structure

4 PHY Layer Enhancements for Improving RAN Power Efficiency – Final Research Results and Benchmarking

As mentioned earlier, the RAN has the largest energy consumption in a mobile network. The energy consumption of the RAN can account for up to 70-80% of the overall network consumption. Therefore, BeGREEN primarily targets optimization of the RAN, i.e., PHY layer functions in order to improve its EE or uses PHY layer functionalities to optimize the RAN.

In this section, two main approaches are pursued. The first approach is to directly reduce the power consumption of the power amplifier at the RU, since it is responsible for the largest power consumption. The second approach, discussed in this section is to use different PHY layer functionalities and technologies to improve the RAN power efficiency. To achieve this, BeGREEN pursues a few different technologies like relays, RIS and ISAC.

4.1 ISAC-assisted PHY layer improvements results and benchmarking

The main idea of using the ISAC system for PHY layer improvements and network optimization is to obtain the positions of the potential users, i.e. UEs, and the obstacles, and based on this data to optimize the PHY layer functions and the network. The main goal of these optimizations is the power consumption reduction.

In BeGREEN, an ISAC system was developed. The purpose of the developed system is to test the sensing functionalities of such a system under realistic conditions and in different scenarios to obtain the KPIs of interest.

In this section we present the sensing data obtained using the sensing functionality of the developed ISAC system and evaluate the KPIs for detecting users and their positions. Additionally, its resolution is tested, to evaluate the capability for distinguishing between different users.

4.1.1 Angular and ranging resolution of the sensing system

Resolution of a sensing system is the ability of the system to distinguish (resolve) between two objects which are placed on minimum angle or distance. Namely, due to the limited bandwidth of the system, the signal reflected from an object, will have a limited rise and fall times at the receiver. As a result, the signal representation of a physical object will depict it as wider than its actual size. This will cause overlapping of the signals from two nearby objects, making them unresolvable as separate objects by the ISAC system. The same applies for the angular resolution, due to the limited antenna beam width. Namely, if multiple objects of interest are placed within a single beam width, they will be detected as one. The capability of the ISAC system to resolve the two objects in the angular dimension is called angular resolution.

The developed sensing system is showing the detected object in a so-called heat map. This representation is shown in Figure 4-1. This representation shows the angle and the distance from which the signals are reflected. The color represents the intensity of the signal coming from the given angle and distance. The blue color represents no reflection and the yellow color represents a strong arriving signal. Additionally, different algorithms can be deployed to detect the objects and to obtain the corresponding polar coordinates of the detected objects.

The concept of the angular and the range resolution is shown in Figure 4-2. Namely, the angle and the distance under which the objects are seen are larger than the actual physical dimensions.

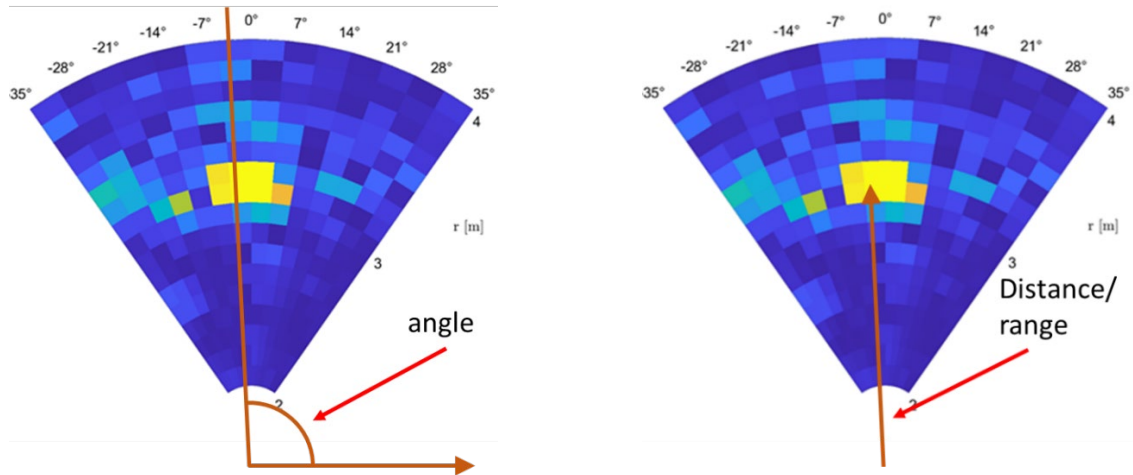


Figure 4-1 Angle and distance of the detected object in polar coordinates

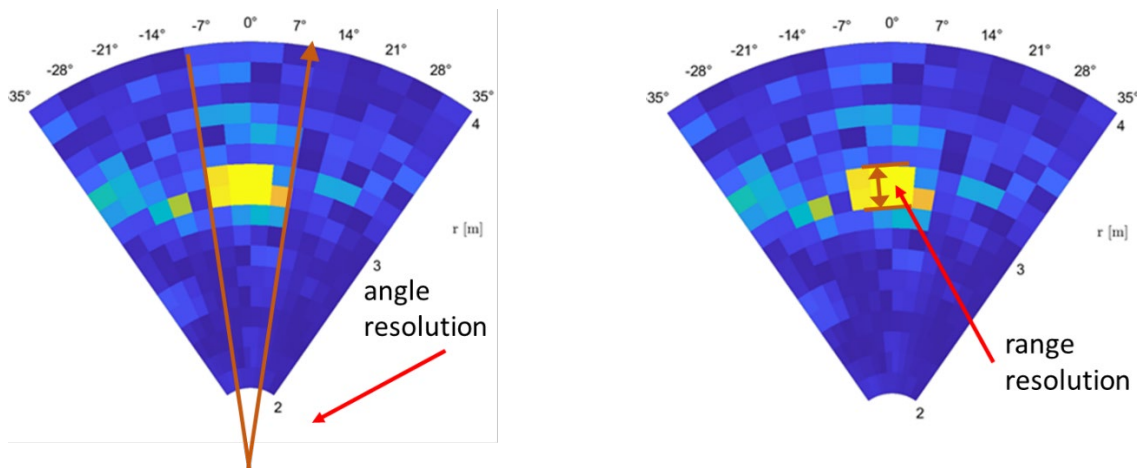


Figure 4-2 Angular and range resolution

The approach pursued in BeGREEN is to use this sensing information to obtain the potential user position as well as their density. Namely, it is possible to distinguish between the static background and the moving objects. The idea is to relate the moving objects to potential users and to use these estimates to assign network resources optimally. Additionally, machine learning (ML) and artificial intelligence (AI) algorithms can be used to detect different classes of objects based on their sensed properties.

4.1.2 Referent sensing tests in an anechoic chamber

In a real and complex environment, it is relatively hard to characterize a sensing system due to the multiple reflections and the multipath propagation of the radio waves. Therefore, the tests at IHP were performed in an anechoic chamber.

A sketch of the anechoic chamber and the installed ISAC setup is shown in Figure 4-3. The anechoic chamber itself is 4×7 meters. The ISAC system is positioned at a distance of 2 meters from the right-hand side wall. The person, or the persons, are moving in front of the ISAC system, between the left-hand side wall and the ISAC system.

A detailed description of the ISAC system is given in deliverable D3.2 [4]. Namely, it is built of $4 \times$ USRP N321 software defined radios (SDRs). These radios share the same local oscillator and have a common timing source. The channel bandwidth used is 200 MHz. A single transmit antenna with antenna gain of 7 dBi is used. The receiving antenna is a patch array antenna consisting of 8 patches. Each patch has an antenna gain of around 3 dBi and a beam width of approximately 60 degrees.

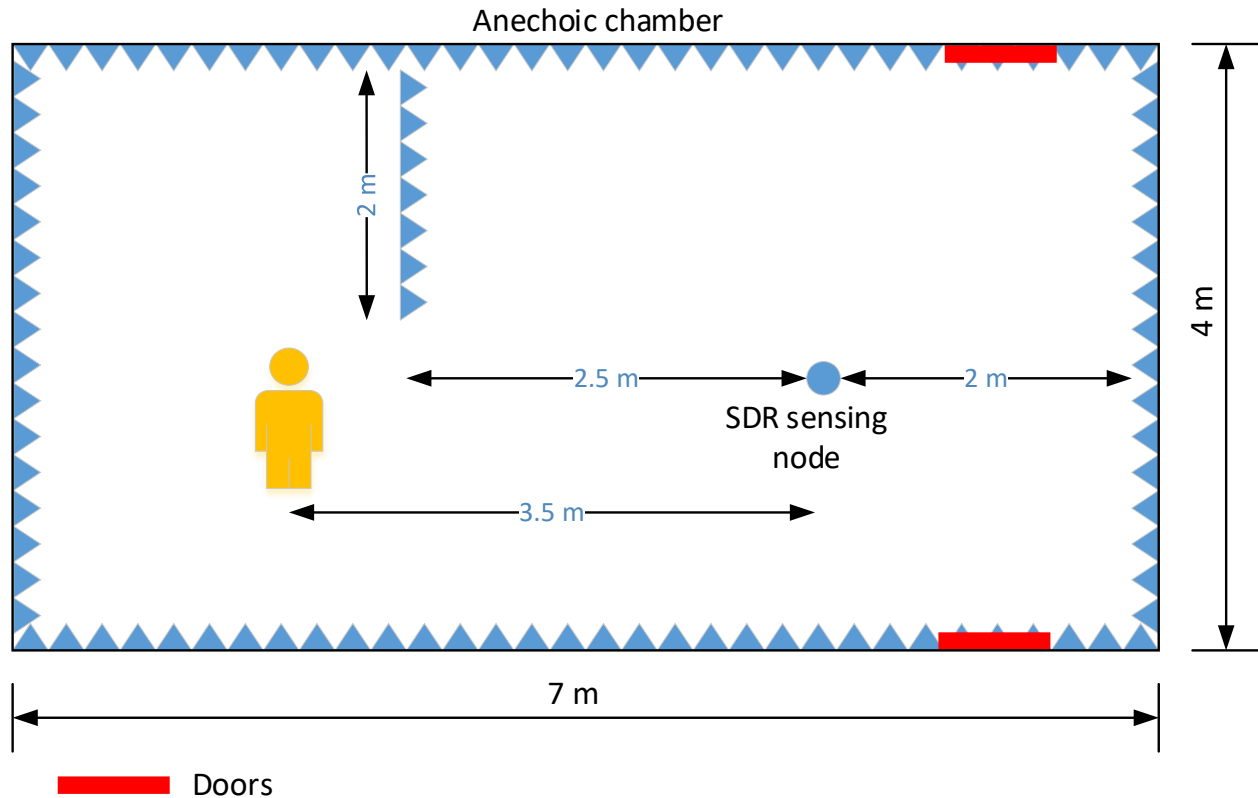


Figure 4-3 Anechoic chamber at IHP with installed ISAC system

In Figure 4-4, a photo of the setup in an anechoic chamber with a person standing in front of it is shown. Before performing any sensing with persons in the anechoic chamber, the system was calibrated and background modelling was performed. The calibration is needed to correct for the different local oscillator (LO) phase at each receiving channel at the SDRs. The background modelling is needed to remove any static clutter, i.e. the static background which is not of interest. This allows for better visualization of the dynamic objects that appear on the scene.

The first performed test is with a single person that appears in front of the ISAC system in the anechoic chamber. With this test it is evaluated what is the expected angular and range resolution of the sensing system.

The heat map created for this scenario is shown in Figure 4-5. As can be noticed, the estimated distance and angle are corresponding to the real situation. The estimated range resolution is about 0.8-0.9 meters. The angular resolution, on the other hand, is about 15 degrees. This is as expected since the beam width of the antenna array is approximately 15 degrees, and the subject has a width of approx. 0.5 meters. The angular resolution might slightly improve for larger distances but would not be better than 7.5 degrees for the used antenna array having 8 antennas.



Figure 4-4 Photo of the anechoic chamber ISAC setup with a person in front of the sensing system

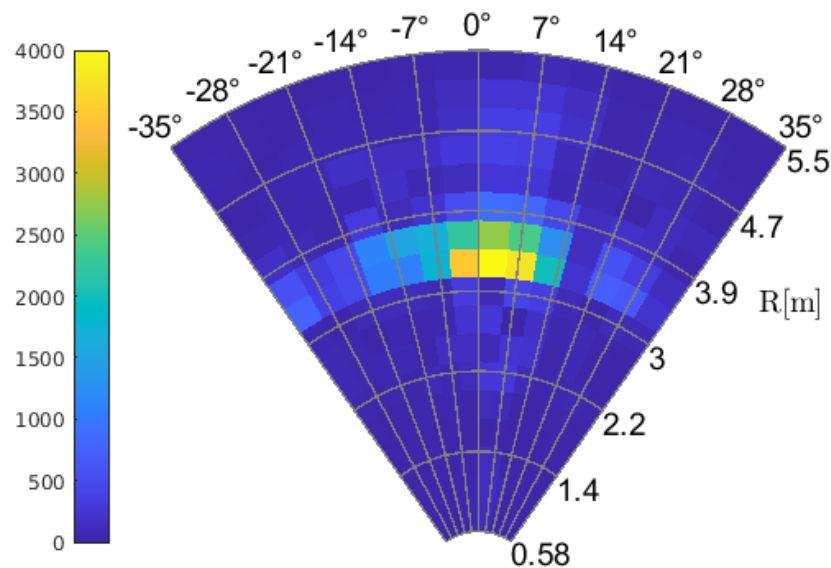


Figure 4-5 Sensing heat map in polar coordinates showing a single person

The second experiment was performed with two persons. Namely, the objective of this experiment was to test the ability of the sensing system to resolve different persons on a relatively small distance. As can be noticed the two persons are on a distance slightly larger than 1.3 meters. In Figure 4-7 a photo of this scenario is shown. The photo is taken from the second person.

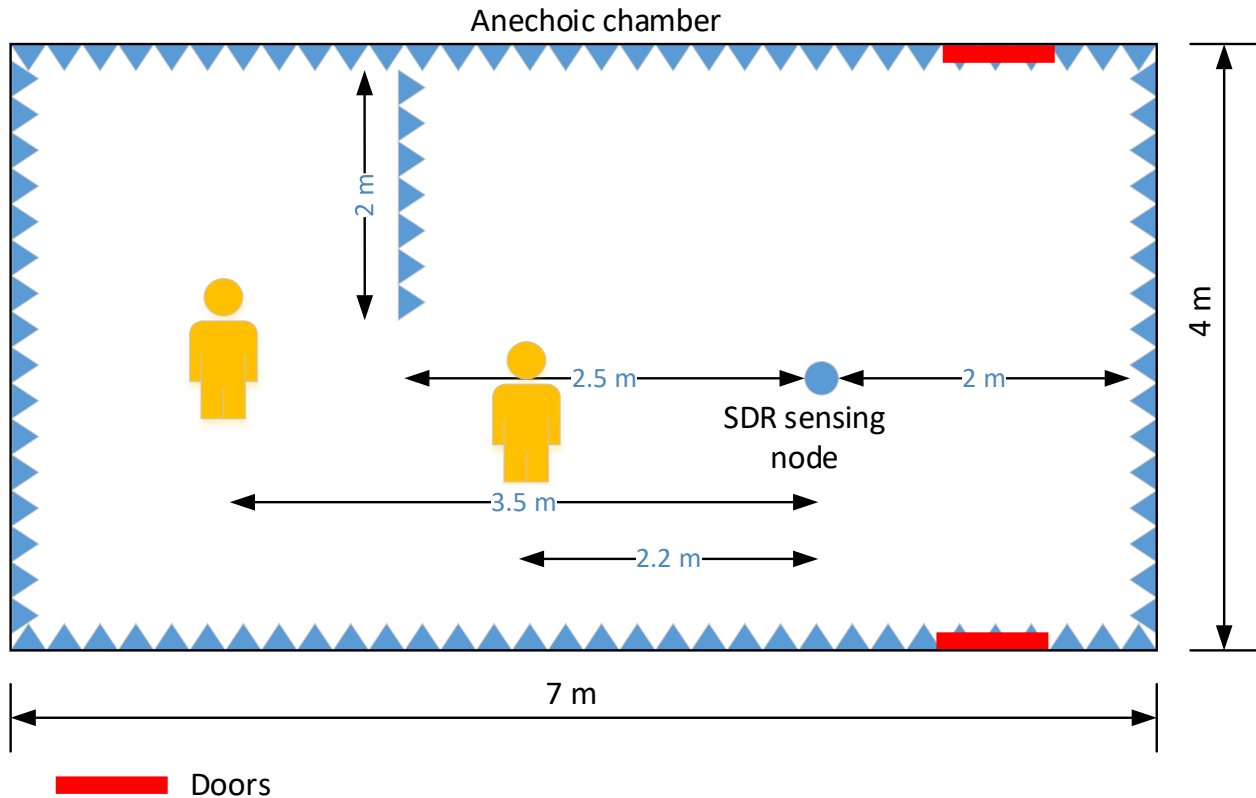


Figure 4-6 Sensing scenario with two persons in front of the sensing system

In Figure 4-8, a photo of the computer showing the positions of the two persons on a heat map is shown. This is a real time representation, and the refresh rate is more than 10 frames per second. The heat map itself is shown in Figure 4-9.

As can be noticed from the performed tests, the system is able to detect persons and to distinguish them from the surrounding environment. The range resolution, with a channel bandwidth of 200 MHz or more is sufficient for scenarios where people should be detected. The angular resolution in this scenario is satisfactory. Nevertheless, if the distances are increased, this angular resolution will not be acceptable. In that case, increasing the number of antennas will be needed if precise detection of people is necessary. In many envisioned use cases this is not the case, therefore, even 8 antennas would give valuable information for estimating user positions.

With these experiments it was shown that the sensing functionality of the radio interface has the potential to be used for improving the beam search and beam tracking functionalities of the physical layer, making them more attractive for use in the next generation radio networks. This should enable using higher antenna gains which will indirectly improve the power efficiency of the network.

The sensing approach demonstrated in this project also shows that it is possible to detect potential users and their positions. This is especially important for optimal and dynamic network resource assignment, which can improve the EE of the whole network.

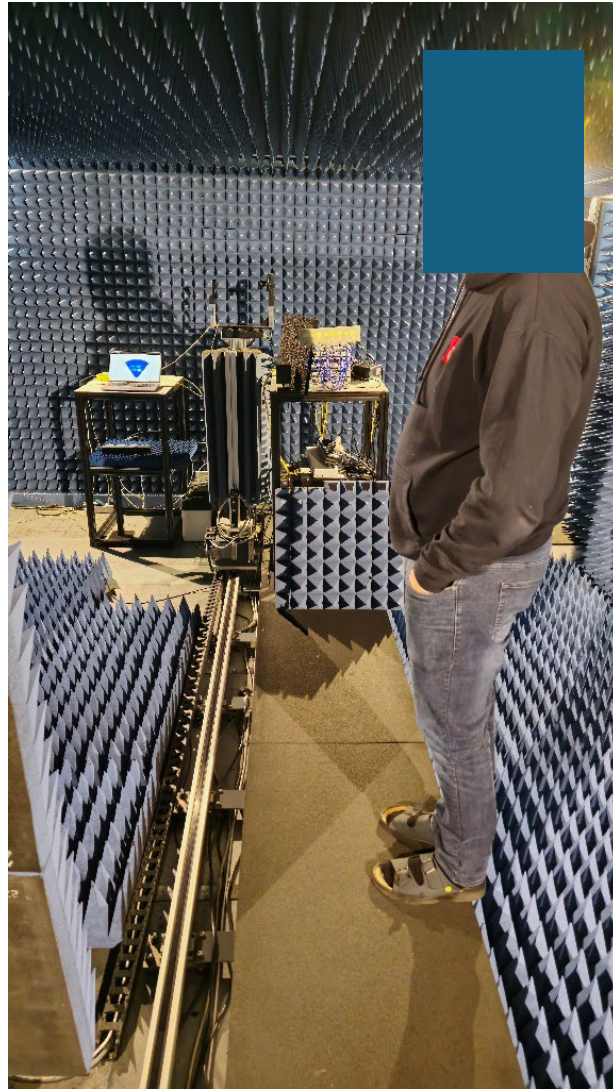


Figure 4-7 Photo of the sensing scenario with two persons, the second person is taking the photo



Figure 4-8 Photo of the computer showing the heat map of the two-person scenario

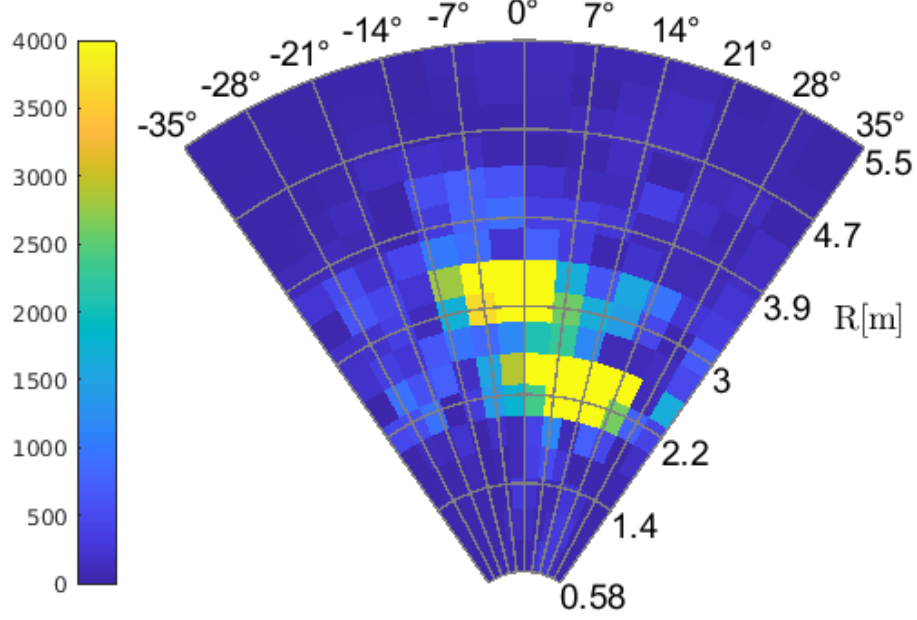


Figure 4-9 Heat map of the two-person scenario

4.1.3 Sensing assisted beam forming

Multiple-input multiple-output (MIMO) radar is an extension of traditional radar utilizing multiple antennas at the transmitter and receiver array, respectively, similar to MIMO communication systems that is used to increase the capacity and the spectral efficiency of the radio links. Unlike the standard phased-array radar transmitting the scaled copies of a transmitting waveform, MIMO radar can transmit mutually orthogonal signals via multiple transmit antennas, offering potential advantages over conventional phased-array systems.

Beamforming is a spatial filtering technique to increase the signal power in desired spatial directions by using the weighted combinations of the antenna inputs and outputs. It is used to increase EE by enabling directive transmissions in MIMO systems. However, in order to make efficient beamforming, the acquisition of the channel state information (CSI) at the transmitter is vital.

In this report, we consider assistance with sensing information to increase the efficiency of the beamforming. We consider a MIMO radar system with multiple users and estimate the Directions of Arrival (DoAs) of the users using the multiple signal classification (MUSIC) algorithm at the receiver side of the MIMO radar. We then use this information to design the beam codebook, namely, we locate more beams on the estimated user areas to decrease the beamwidth and increase EE, while covering the idle areas by only beam. Numerical examples in Section 4.1.3.4 demonstrate the effectiveness of the proposed approach; the success rate and the EE of the beam search is improved compared to the case that there is no sensing parameter estimation.

4.1.3.1 System model

We consider a MIMO radar system at the transmitter, with M antennas at its transmitter and N antennas at its receiver side. The uniform linear array (ULA) model with half-wavelength spacing is adapted. Assuming that there are K single-antenna users, and following the model in [1], the output of the matched filters at the receiver can be written as

$$\mathbf{X}(t) = [\mathbf{a}_r(\phi_1) \otimes \mathbf{a}_t(\theta_1), \mathbf{a}_r(\phi_2) \otimes \mathbf{a}_t(\theta_2), \dots, \mathbf{a}_r(\phi_K) \otimes \mathbf{a}_t(\theta_K)] \mathbf{b}(t) + \mathbf{n}(t)$$

where θ_k and ϕ_k are the degree of departure (DoD) and degree of arrival (DoA) and $\mathbf{a}_t(\theta_k)$ and $\mathbf{a}_r(\phi_k)$ are transmit and receive array of the k -th user, respectively, $\mathbf{b}(t) = [b_1(t), b_2(t) \dots b_K(t)]$ where $b_k(t) =$

$\beta_k e^{j2\pi f_k t}$ with β_k is the complex amplitude and f_k is the Doppler frequency and $\mathbf{n}(t)$ is the additive white Gaussian noise. The transmit and receive array can be written as

$$\begin{aligned}\mathbf{a}_r(\phi) &= [1 e^{-j\pi \sin(\phi)} \dots e^{-j\pi(N-1) \sin(\phi)}] \\ \mathbf{a}_t(\theta) &= [1 e^{-j\pi \sin(\theta)} \dots e^{-j\pi(M-1) \sin(\theta)}]\end{aligned}$$

The covariance matrix of the signal in $\mathbf{X}(t)$ can be estimated by L snapshots, hence $\hat{\mathbf{R}}_x = \frac{1}{L} \sum \mathbf{X}(t_l) \mathbf{X}^H(t_l)$ with a summation over l . After the eigen-decomposition, one can re-write $\hat{\mathbf{R}}_x$ as

$$\mathbf{R}_x = \mathbf{E}_s \mathbf{D}_s \mathbf{E}_s^H + \mathbf{E}_n \mathbf{D}_n \mathbf{E}_n^H$$

where $\mathbf{D}_s$ and $\mathbf{D}_n$ are the diagonal matrices with the largest K eigenvalues and rest of the eigenvalues as their diagonal elements, respectively, and $\mathbf{E}_s$ and $\mathbf{E}_n$ are the matrix of eigenvectors corresponding to them.

At the transmitter side, we apply beamforming for directional transmission. We employ a codebook-based analog beamforming, that is, the beamforming vector to each user is selected from a pre-defined codebook consisting of N_B elements. We utilize a beamsteering codebook [2], namely, the beamforming vectors in the codebook are in the form of array steering vectors, $\mathbf{a}_r(\phi)$ and $\mathbf{a}_t(\theta)$, i.e., they are represented by only an angle.

The elements of the beam codebook are constructed as follows: The search space of angle values is divided into N_B sections and each section has a central angle ϕ_c . Then, the corresponding beamforming vector is formed as $\mathbf{a}(\phi_c)$.

4.1.3.2 Sensing-assisted Beamforming

In this section, the beam search procedures and the design of the beam codebook with the assistance of sensing parameter estimation are presented. To find the beam vector of each user, the transmitter selects the beam codebook element maximizing the desired signal power by testing all possible beam vectors, i.e., an exhaustive search is applied. With the exhaustive search, the optimal transmit beam for the u -th user is selected as

$$\mathbf{f}_{opt,u} = \max_i |\mathbf{H}_u^H \mathbf{f}_i|, i = 1, 2, \dots, N_B$$

where $\mathbf{f}_i$ is the $M \times 1$ beam vector, that is the i -th column of the beam codebook $\mathbf{F}$, and $\mathbf{H}_u$ is the DL transmission channel from transmitter to the u -th user. Note that perfect CSI at the transmitter is assumed in the equation above. In the case of imperfect CSI, $\mathbf{H}_u$ can be replaced with $\hat{\mathbf{H}}_u = \mathbf{H}_u + \mathbf{E}_u$, where $\mathbf{E}_u$ is the vector of channel estimation errors of the u -th user with elements distributed as $CN(0, \sigma_h^2)$, where σ_h^2 is the variance of the channel estimation error.

The sensitivity of the beam codebook affects the accuracy of the beam selection. Namely, if the angle elements in the angle space $[-\frac{\pi}{2}, \frac{\pi}{2}]$ are more quantized, the probability of selecting a more accurate beam increases, which will increase EE. If there is no sensing parameter estimation, the DoD and DoAs of the users are unknown to the transmitter, and the beams can only be located based on the distribution of the DoAs. For instance, if the DoA of the users is uniformly distributed, the angle space is uniformly divided into the sections, and each beam is located in a section (it is called the uniform codebook in the rest of the report).

4.1.3.3 Sensing-assisted Beam Codebook Design

In order to exploit the sensing information, we first estimate the DoD and DoAs of the users by the MUSIC algorithm. For this purpose, we construct the spatial spectrum function as [1]

$$f_{MUSIC}(\phi, \theta) = \frac{1}{[\mathbf{a}_r(\phi) \otimes \mathbf{a}_t(\theta)]^H \mathbf{E}_n \mathbf{E}_n^H [\mathbf{a}_r(\phi) \otimes \mathbf{a}_t(\theta)]}$$

The angles of the users can be estimated by taking the K peaks of the spectrum function above. Note that our MIMO radar is monostatic (the transmit and receive antenna arrays are co-located), hence, $\phi_k = \theta_k$ and a 1-D search is sufficient.

We propose to utilize the estimated DoA of the users to design the beam codebook, which results in a higher sensitivity if the angle estimation is successful. To do that, for each estimated user angle ϕ_k , $\frac{N_B-2}{K}$ codebook elements are located in the angle area $[\theta_k - \alpha, \theta_k + \alpha]$, where α is a small integer in degrees that can be set adaptively based on the number of beam codebook elements. If $\frac{N_B-2}{K}$ is not an integer, it is rounded to the nearest smaller integer in order to prevent exceeding the number of beam elements. The areas that do not fall into the vicinity of any user are covered by only one beam. In this way, the precision of the beam search is increased since the beams are more concentrated in the areas of the users.

Let us illustrate this with a numerical example: Suppose that there are $N_B = 32$ beam codebook elements and $K = 3$ users. Assuming a uniform user distribution in the angle space, in the case of no sensing assistance, the sensitivity of the beam codebook becomes 5.63 degrees. On the other hand, in sensing-assisted beam codebook design, with $\alpha = 5$ degrees, if the DoAs of the users are correctly estimated, 10 beam elements are located in an interval of 10 degrees, hence, the sensitivity of the beam codebook becomes 1 degree.

4.1.3.4 Numerical Results

We evaluate the performance of the beam codebook design method with Monte Carlo simulations in MATLAB. In our simulations, we take $M = N = 8$, $L = 100$, assume that there $K = 3$ users distributed in angle space $[-\frac{\pi}{2}, \frac{\pi}{2}]$, and the amplitude is a complex zero-mean unit variance Rayleigh random variable unless otherwise specified.

In Figure 4-10, we evaluate the success rate of beamforming using sensing-assisted adaptive codebook and the uniform beam codebook with perfect and imperfect CSI assuming that the users are uniformly distributed in the angle space for different codebook sizes. We set the system SNR to 10 dB. We define the success rate as follows: if the DoA of a user is estimated with an error of less than 1 degree, the beam training is assumed to be successful. The results in Figure 4-10 shows that sensing-assisted beamforming is superior to the beamforming with a uniformly quantized codebook even with perfect CSI at the transmitter if the number of beam codebook elements is less than 128. The success rate of sensing-assisted scenario has a flat behavior since its performance is limited by the performance of the MUSIC algorithm, which has an error rate about 0.06 at 10 dB for the given set of parameters. For the imperfect CSI case, the success rate of uniform codebook reaches a maximum of 0.73 for $\sigma_h^2 = -10$ dB and 0.3 for $\sigma_h^2 = 0$ dB.

We assess the success rate of the proposed solution for an SNR range of $[-15, 20]$ and compare it to that of uniform codebook case for $N_B = 64$ in Figure 4-11. Note that SNR does not affect the success rate of the uniform codebook. The results in Figure 4-11 illustrate that for the given set of parameters sensing assisted beamforming is superior when SNR is -10 dB or higher.

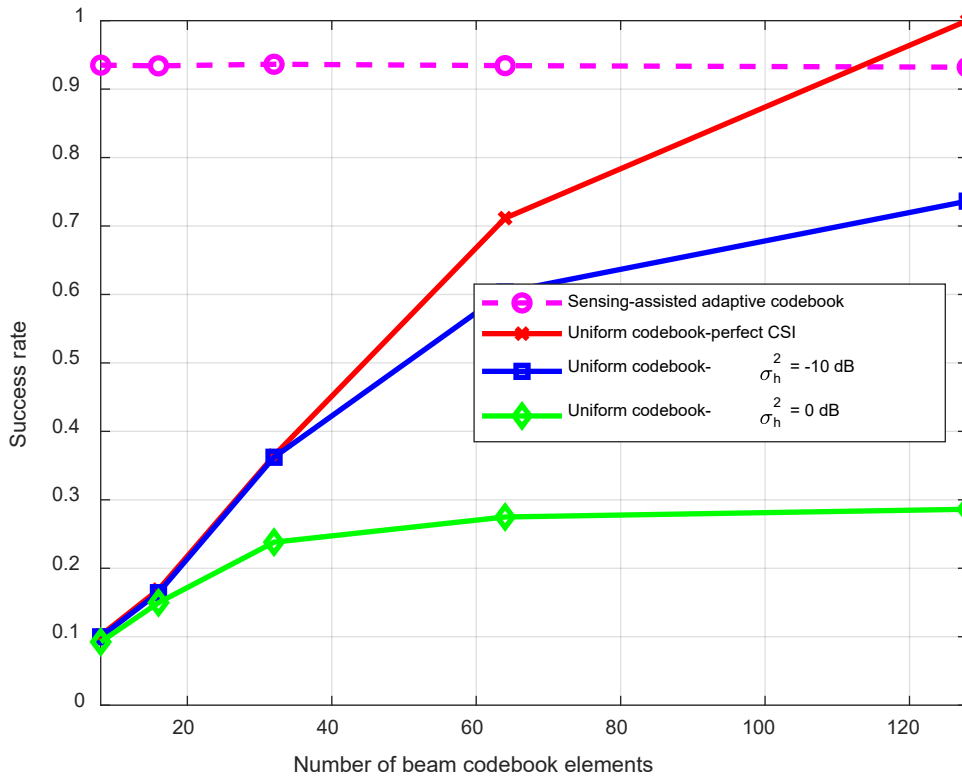


Figure 4-10 Success rate of sensing-assisted beamforming and beamforming without sensing information for different codebook sizes

In Figure 4-10 and Figure 4-11, we assume a uniform distribution of DoAs. In Figure 4-12, we assume that the DoAs of the users are non-uniformly distributed, where θ_k is a uniform random variable in $[-30, 30]$ degrees, with a probability of 80 %, and uniformly distributed in $[-90, -30]$ and $[30, 90]$ degrees, with a probability of 10% each. In this case, one can locate the beams based on the distribution of the angles, i.e., there will be more beams in the area where more users are expected. This approach is superior to the uniform codebook for most of the beam codebook sizes as shown in Figure 4-12, where we calculate the success rate of the proposed solution and these codebook design approaches for SNR = 10 dB. However, sensing-assisted codebook design still offers a higher success rate since the exact angles are estimated with a high probability.

In another line of investigation, we conduct bit error rate (BER) simulations using the selected beams. We assume an uncoded BPSK transmission, the users are separated by 20 degrees in the angle space and set $N_B = 16$. The other parameters are the same as the previous simulations. The results in Figure 4-13 demonstrate that our proposed sensing-assisted beamforming offers a similar performance with the uniform codebook even with the perfect CSI at the transmitter. However, as the perfect CSI is not a realistic assumption, we also consider imperfect CSI case with different channel estimation error variances. The results show that sensing-assisted beamforming outperforms beamforming with uniform codebook with imperfect CSI by about 10 dB for a BER of 10^{-2} when $\sigma_h^2 = -10$ dB, and about 13 dB for a BER of 0.05 when $\sigma_h^2 = 0$ dB. This means that sensing-assisted beamforming leads to important energy savings for practical settings.

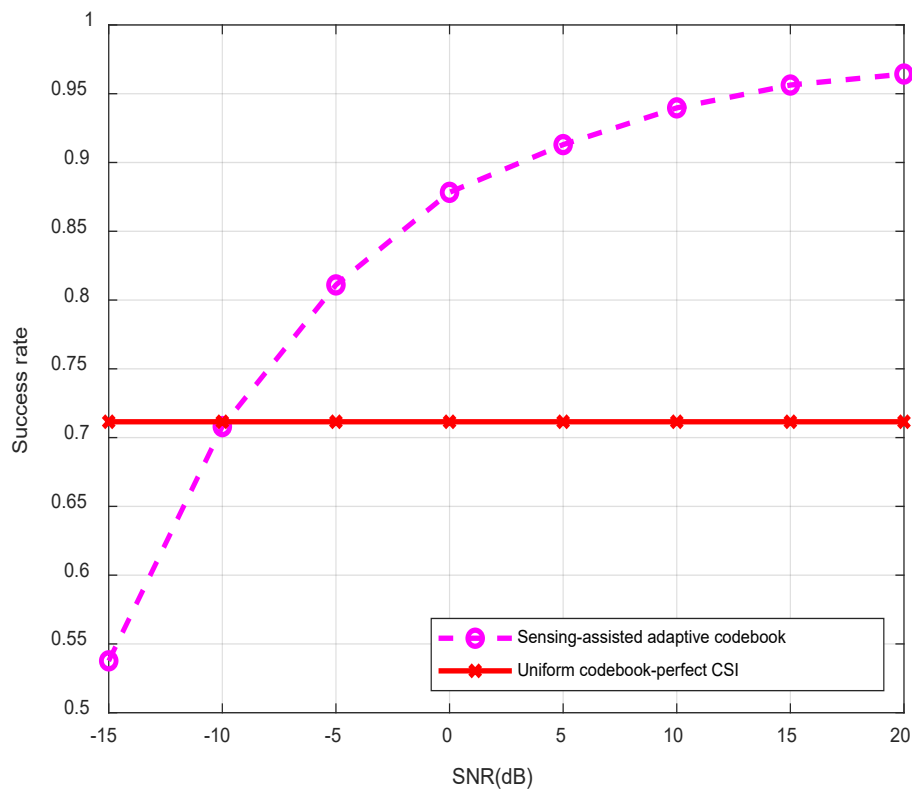


Figure 4-11 Success rate versus SNR for sensing-assisted beamforming and uniform codebook

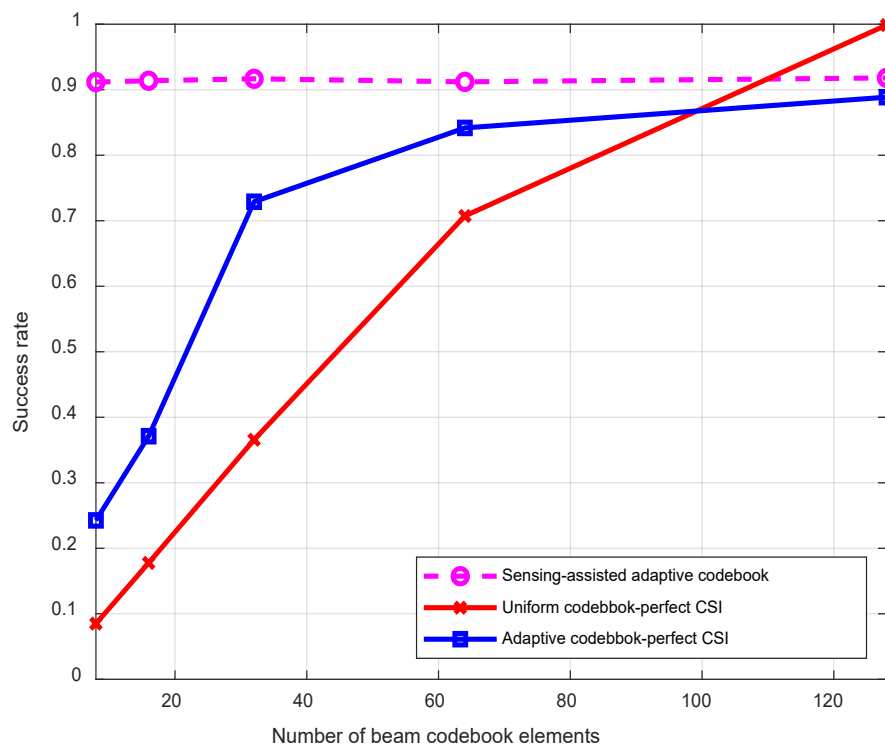


Figure 4-12 Success rate versus codebook size for non-uniform user distribution

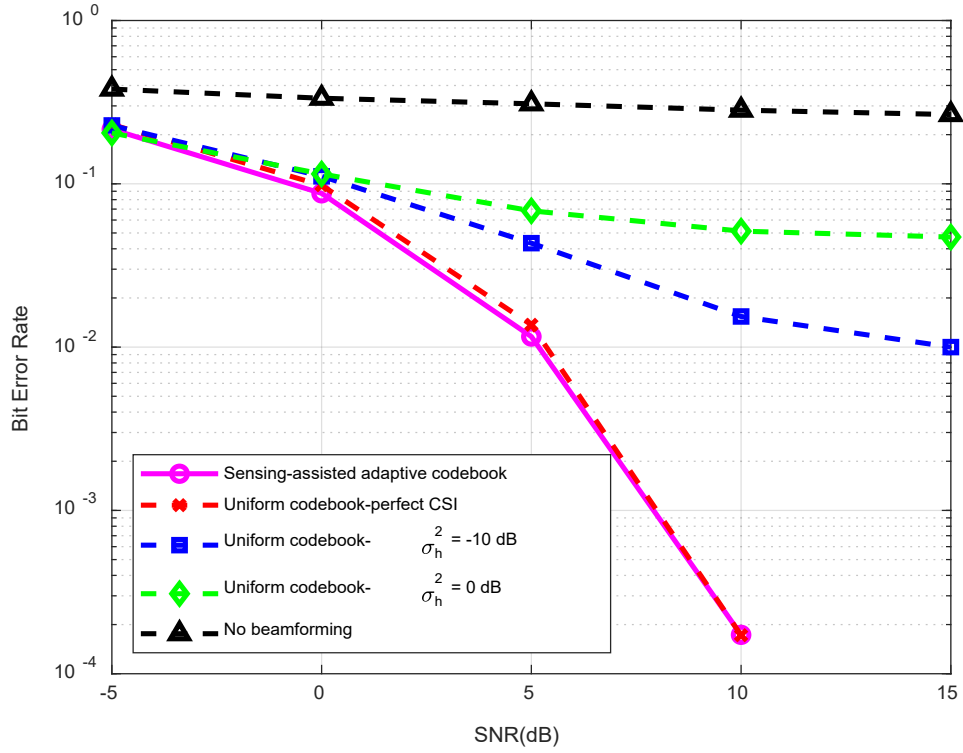


Figure 4-13 Bit error rate comparison of proposed sensing assisted beamforming, uniform codebook and no beamforming case

We also compare root mean square error (RMSE) of the beam angle estimation for our proposed sensing-assisted adaptive codebook and uniform codebook with perfect CSI. For this purpose, we calculate the RMSE between the actual DoA values and the selected beam angle. We set the number of beam codebook elements to 16 and use the default system parameters for number of antennas and users. Note that RMSE of uniform codebook case is not affected by the SNR. The results in Figure 4-14 show that sensing-assisted beamforming offers a lower RMSE compared to uniform codebook even if there is perfect CSI.

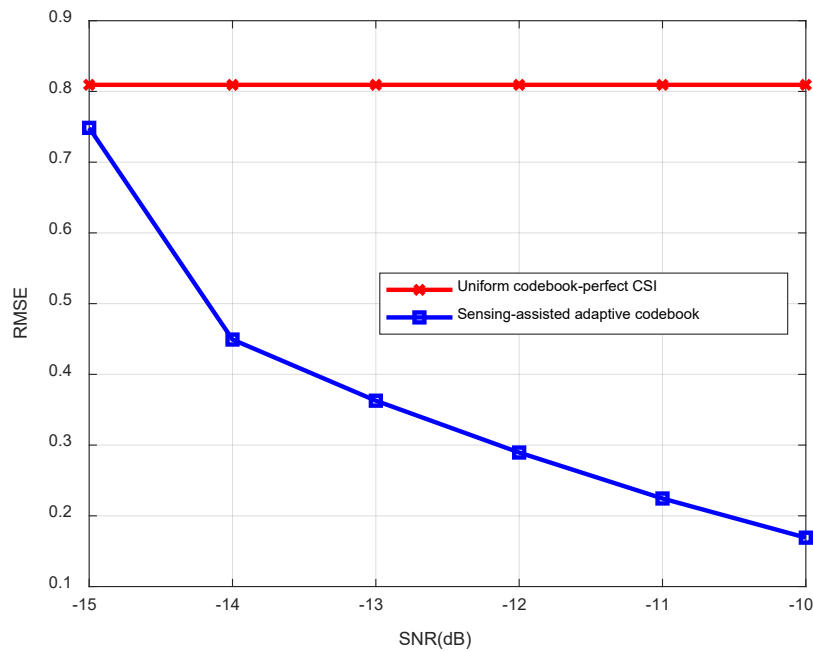


Figure 4-14 RMSE comparison of sensing assisted beamforming and uniform codebook

4.2 Self-configuring RIS final research results and benchmarking

BeGREEN D2.2 built a mathematical model for optimizing the Hybrid Reconfigurable Intelligent Surface (HRIS) configuration. From a practical standpoint, however, the optimal solution $\mathbf{v}_{BU}$ depends on the array response vectors from the HRIS towards the BS and the UEs. To implement the obtained solution, the array response vectors, i.e., the BS-RIS and RIS-UEs AoAs, need to be estimated, but this is not possible at the HRIS due to the absence of RF chains and of a control channel. In this section, we propose a codebook-based approach for estimating the necessary AoAs and then computing $\mathbf{v}_{BU}$ in a distributed manner and locally at the HRIS, i.e., our proposed Metasurface Absorption and Reflection for Intelligent Surface Applications (MARISA).

MARISA optimizes the HRISs based on an appropriately designed codebook, which allows for the estimation of the BS-RIS and RIS-UEs AoAs in a distributed manner. The use of codebooks is a known approach in RIS-assisted communications, e.g., [7], [8] and it is usually implemented by assuming that the electronic circuits of the RIS can realize a finite number of phase responses (e.g., through PIN diodes [9]). Therefore, our proposed MARISA is compatible with conventional implementations of RISs, but it does not need a control channel.

Let us consider a codebook $\mathcal{C} = \{\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_L\}$, whose codewords $\mathbf{c}_l \in \mathbb{C}^{N \times 1}$ are unit-norm beamforming vectors that correspond to a discrete set of possible phase shift matrices $\Theta_l = \text{diag}(\mathbf{c}_l^H)$. In particular, each codeword $\mathbf{c}_l$ is constituted by discrete-valued entries that mimic a sort of phase quantization. The discrete values of the codewords are assumed to belong to the following set

$$\mathcal{Q} = \left\{ \frac{2\pi}{2^Q} m : m = 0, \dots, 2^Q - 1, m \in \mathbb{N} \right\}$$

where 2^Q is the possible number of discrete values.

In MARISA, the HRIS operates in two possible modes: probing and communication. In the probing mode, the HRIS estimates the AoAs that correspond to the BS and to the UEs. In the communication mode, the HRIS assists the transmission of data between the BS and the UEs, while still being capable of detecting new UEs joining the network or previously inactive UEs. Therefore, we relax the assumption of simultaneous UEs transmissions, i.e., the UEs may transmit at different times. Probing and communication phases are detailed in the following.

Probing phase. Without loss of generality, we assume that each codeword of the codebook is, to a certain extent, spatially directive, i.e., the resulting HRIS configuration maximizes the absorbed power only in correspondence of a (narrow) solid angle. This is relatively simple to realize by enforcing, e.g., some constraints on the design of the codewords in terms of half-power beamwidth of the corresponding radiation pattern of the HRIS. Therefore, by iteratively sweeping across all the codewords $\mathbf{c}_l \in \mathcal{C}$, the HRIS can scan, with a given spatial resolution, the three-dimensional (3D) space and can detect network devices (the BS and the UEs) by using pilot signals emitted only by those devices. During this probing phase, the HRIS collects a set of power measurements, or equivalently a power profile, $\mathcal{P} = \{\rho_1, \dots, \rho_L\}$ where each element $\rho_l \in \mathbb{R}$ is the power level sensed (measured) by the HRIS when using the codeword $\mathbf{c}_l$. As a result, the array response vectors can be estimated from $\mathcal{P}$. In practice, this boils down to detecting the peaks in $\mathcal{P}$ and identifying the corresponding angular directions. By construction, in fact, the HRIS detects a power peak only if there is at least one transmitter in the direction synthesized by the HRIS beampattern (i.e., the considered codeword). The finer the angular selectivity of the HRIS, the longer the probing phase. Therefore, a suitable compromise needs to be considered. An example of power profile as a function of the steering angle of the HRIS is reported in Figure 4-15.

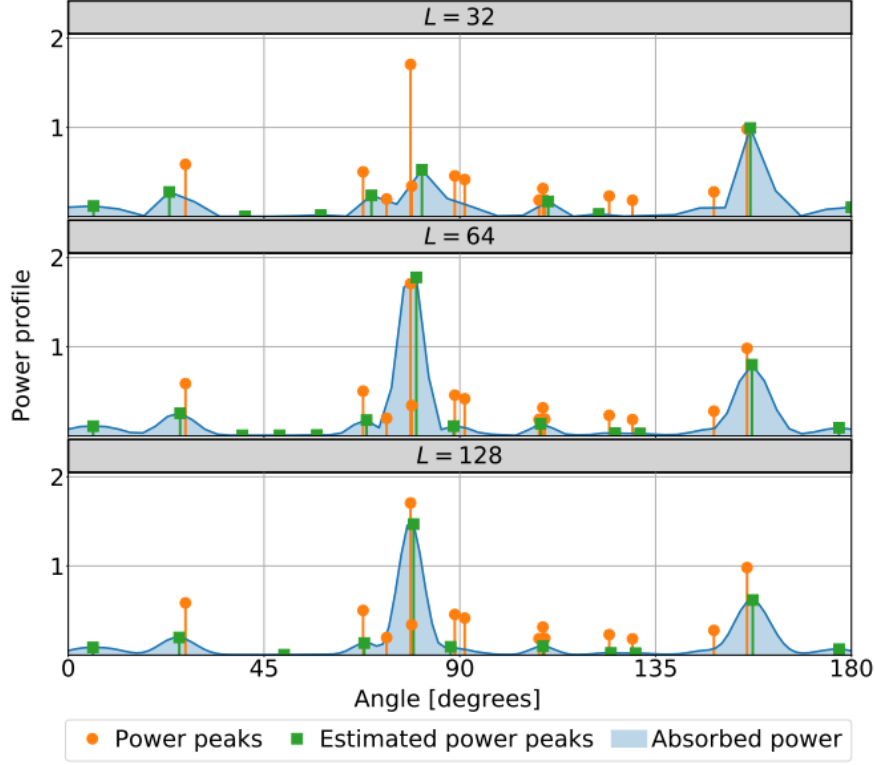


Figure 4-15 Example of power profile and corresponding estimated peak for different codebook sizes L in a multi-UE scenario with $K = 14$ UEs and $N = 32$ HRIS elements.

In particular, we assume that ρ_l is a power peak in $\mathcal{P}$ if it is greater than a given threshold $\tau \in \mathbb{R}^+$. Let $\mathcal{I} = \{i < L : \rho_i \in \mathcal{P} > \tau\}$ be the set of indexes l corresponding to the power peaks. Then, depending on which devices transmit their pilot signals, $\mathbf{v}_B$ and $\mathbf{v}_U$ can be estimated as

$$\mathbf{v}_B = \sum_{i \in \mathcal{I}} \delta_i \mathbf{c}_i, \quad \mathbf{v}_U = \sum_{i \in \mathcal{I}} \delta_i \mathbf{c}_i$$

where $\delta_i \in \{1, \rho_i\}$ is a weight parameter that allows performing a hard ($\delta_i = 1$) or a soft ($\delta_i = \rho_i$) combining of the power peaks in $\mathcal{I}$ based on the actual measured power ρ_i .

The end-to-end HRIS optimal configuration $\mathbf{v}_{BU}$ is first computed from the equations presented in the previous deliverable D3.2 and is then projected onto the feasible set of discrete phase shifts, which eventually yields the desired $\bar{\mathbf{v}}_{BU}$. The proposed probing phase is summarized in Algorithm 1.

Algorithm 1 MARISA – Probing phase

- 1: Data: $\mathcal{C}$, $\tau \in \mathbb{R}^+$
 - 2: Perform a beam sweeping setting $\Theta_l = \text{diag}(\mathbf{c}_l)$, $\forall \mathbf{c}_l \in \mathcal{C}$
 - 3: Measure the corresponding power profile $\mathcal{P}$
 - 4: Obtain $\mathcal{I} = \{i < L : \rho_i \in \mathcal{P} > \tau\}$
 - 5: **if** the BS transmits the pilot signals **then**
 - 6: Compute $\mathbf{v}_B = \sum_{i \in \mathcal{I}} \delta_i \mathbf{c}_i^H$
 - 7: **else if** the UEs transmit the pilot signals **then**
 - 8: Compute $\mathbf{v}_U = \sum_{i \in \mathcal{I}} \delta_i \mathbf{c}_i^H$
 - 9: **end if**
 - 10: Obtain $\mathbf{v}_{BU} = \mathbf{v}_B \circ \mathbf{v}_U^*$ and $\bar{\mathbf{v}}_{BU}$ after quantization
-

Communication phase. Upon completion of the probing phase, the HRIS enters the communication phase, which is aimed to assist the reliable transmission of data between the BS and the active UEs, as well as to probe the 3D space in order to discover new UEs. The communication phase is summarized in Algorithm 2. Once $\bar{\mathbf{v}}_{BU}$ is obtained, the HRIS can easily aid the BS and the active UEs to communicate with each other. More challenging is the execution of the probing phase simultaneously with the communication phase.

Therefore, hereafter we focus our attention on it.

Algorithm 2 MARISA – Communication phase

- 1: Data: $\mathcal{C}, \tau \in \mathbb{R}^+$
 - 2: Obtain $\hat{\mathcal{I}} = \{l \in \mathbb{N}^+ : l \leq L\} \setminus \mathcal{I}$
 - 3: Define $\hat{\mathcal{C}} = \{\text{diag}(\hat{\mathbf{v}}_{BU}^H) + \mathbf{c}_j : \mathbf{c}_j \in \mathcal{C}, j \in \hat{\mathcal{I}}\}$,
 - 4: Execute the probing phase in Algorithm 1
-

To this end, we introduce the set of indices $\hat{\mathcal{I}} = \{l \in \mathbb{N}^+ : l \leq L\} \setminus \mathcal{I}$ that do not correspond to any power peaks estimated during the probing phase. Then, we construct the following codebook for simultaneous probing and communication

$$\hat{\mathcal{C}} = \{\text{diag}(\hat{\mathbf{v}}_{BU}^H) + \mathbf{c}_j : \mathbf{c}_j \in \mathcal{C}, j \in \hat{\mathcal{I}}\}$$

whose codewords, after normalization and phase quantization, allow the HRIS to scan the 3D space, looking for new UEs, while keeping unaffected the beam steering that corresponds to the UEs being already served. It is worth mentioning that the probing phase for discovering new UEs does not replace the probing phase, executed on a regular basis, in Algorithm 1. This is because it is necessary to check whether the estimated directions of the HRIS still point towards the active UEs.

Performance Evaluation

To prove the feasibility of MARISA, we evaluate it in different scenarios and compare it against the SoA benchmark scheme, recently reported in [10], which relies upon a control channel to perform a centralized optimization. The simulation setup and the parameters are given in Table . All results are averaged over 100 simulation instances.

Table 4-1 Simulation setup and parameters

Parameter	Value	Parameter	Value	Parameter	Value
M	4	N_x, N_z	8, 4	f_c	28 GHz
$\mathbf{b}$	(−25, 25, 6) m	$\mathbf{r}$	(0, 0, 6) m	A	$50 \times 50 \text{ m}^2$
P	20 dBm	$\beta_{\text{LoS}}, \beta_{\text{NLoS}}$	2, 4	σ_n^2	−80 dBm
d_0, γ_0	1	η	0.8	L	32
λ_B	0.3 m^{-2}	h_B	1.8 m	r_B	0.6 m

The network area A is a square, and the BS and the HRIS (or the RIS) are located in the midpoints of two of its adjacent edges. The UEs are uniformly distributed in the network area, i.e., $\mathbf{u}_k \sim \mathcal{U}(A)$. To show the robustness of MARISA in realistic propagation scenarios, we relax the assumption of LoS propagation conditions and account for the non-line-of-sight (NLoS) paths as well. In particular, we consider the stochastic geometry-based model in [11], which relates the geometric properties of the communication path in terms of the path length l and height of the communicating devices (i.e., u_z, b_z , and r_z) in the presence of physical obstacles that may obstruct the links, which are referred to as blockers. The blockers are modeled as cylinders of height h_B , diameter r_B , and are distributed according to a Poisson point process (PPP) with intensity λ_B . Therefore, for each path in the network area, the probability of NLoS propagation is

$$p(l) = 1 - e^{-2\lambda_B r_B (\sqrt{l^2 - (b_z - u_z)^2} \frac{h_B - u_z}{h_B - u_z} + r_B)}$$

The pathloss exponent for the LoS or NLoS paths are denoted by β_{LoS} or β_{NLoS} , respectively. It is worth mentioning that the performance of centralized deployments and MARISA depend on the overhead for channel estimation and reporting [12], and the overhead of the probing phase [13], respectively. These two solutions are very different from each other, and a fair comparison of the associated overhead is postponed to a future research work.

Comparison with the Centralized Deployment

We analyze the viability of self-configuring an HRIS by solving the optimization problem presented in the

previous deliverable with perfect knowledge of the aggregate HRIS-UE channel h_{Σ} and of the response vector of the HRIS towards the BS $\mathbf{a}_R(\mathbf{b})$. We refer to this design as the Oracle (**O**) scheme, since the channels are known already and do not need to be estimated.

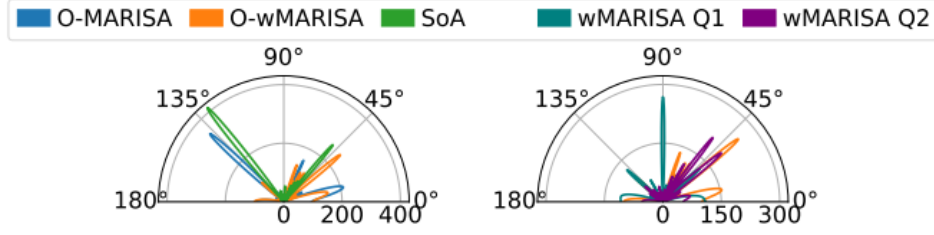


Figure 4-16 Radiation pattern at the HRIS along the azimuth directions obtained with O-MARISA, O-wMARISA, and SoA (left), and radiation patterns obtained with Q bits of phase quantization (right).

Moreover, two solutions are analysed that assume real-valued (continuous) phase shifts: i) O-MARISA, which calculates $h_{\Sigma} = \sum_k \frac{h_k}{\|h_k\|}$, and ii) O-weighted MARISA (O-wMARISA), which calculates $h_{\Sigma} = \sum_k h_k$. Specifically, O-MARISA estimates $\mathbf{v}_B$ and $\mathbf{v}_U$ only based on the direction of the paths that are assumed to have a unit gain, while O-wMARISA utilizes the direction and the gain of the paths. Figure 4-16 (left) shows a comparison of the HRIS configuration obtained by O-MARISA, O-wMARISA, and the SoA centralized solution in [10], which jointly optimizes the BS precoder and the RIS phase shifts by means of a control channel. While the SoA provides a very directive beampattern with few enhanced directions, both versions of O-MARISA result in a wider range of directions at the expense of a smaller gain due to the presence of multiple secondary lobes. Despite the different beampatterns, the sum-rates obtained by MARISA and the centralized benchmark, as shown in Figure 4-17, are very similar. In particular, O-MARISA and the SoA provide a sum-rate that does not increase with the number of UEs, which hints to an interference-constrained scenario. Notably, O-wMARISA delivers better performance thanks to the weighting mechanism that strengthens the reflected paths with higher power gains. This behavior is further confirmed in Figure 4-18, where the average sum-rate is analyzed against the number of HRIS elements N and UEs K . We see that, in this case, only O-wMARISA has a non-decreasing behavior regardless of the interference-constrained nature of the scenario, largely outperforming the SoA centralized solution in [10].

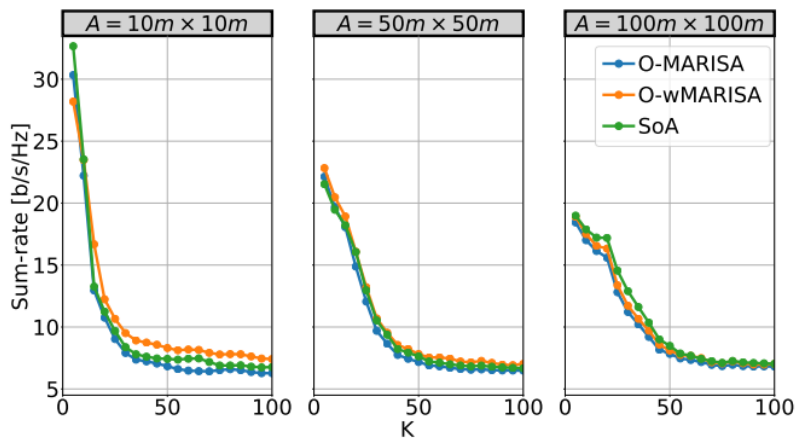


Figure 4-17 Average sum-rate in a multi-UE scenario obtained by solving Problem 3 with perfect CSI and by SoA [4] against the number of UEs K for different network areas and when the number of HRIS elements is $N=32$.

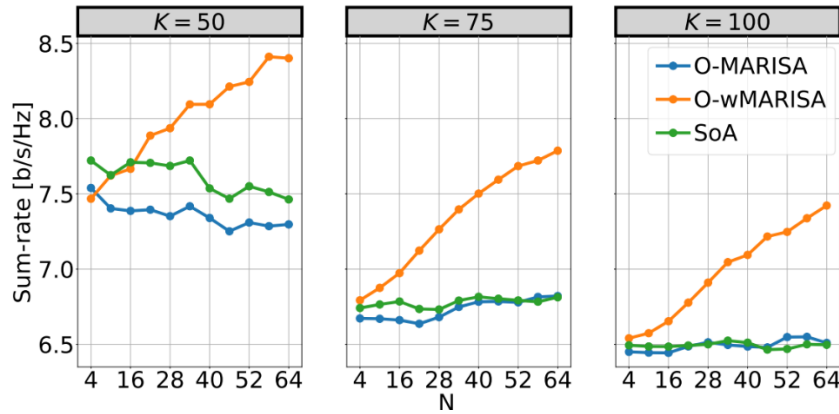


Figure 4-18 Average sum-rate in a multi-UE scenario obtained by solving Problem 3 with perfect CSI and by SoA [4] against the number of HRIS elements N , the number of UEs K . The network area is $A=50$ "m" $\times$ 50 "m"

4.3 Integration concept for ISAC and RIS into O-RAN

To integrate these ISAC/RIS based EE solutions, the O-RAN enabled BeGREEN architecture is augmented with several RIS/ISAC specific components. Figure 4-19 shows part of the BeGREEN architecture which is related with RIS and ISAC components.

For the case of RIS, and as introduced in BeGREEN deliverable D2.2 Section 4.3.2 [16], in common RIS-enabled use cases such as coverage extension (EE, signal enhancement, deployment costs, etc.), RIS may be controlled by the Near-RT RIC for relatively fast operations, such as RIS per-element configuration, and by the Non-RT RIC for slower operations, such as logical split of the RIS or change of static configurations. Therefore, the BeGREEN architecture envisages the O1+ and E2+ interfaces. As discussed in BeGREEN deliverable D4.2 Section 2.2.3 [14], O1+ interface is based on NETCONF and implements the specific YANG model that is defined by each RIS vendor. The interface will be managed by the O1+ RIS Functions module, and as any other standard NETCONF client such as netopeer [17], it will include common NETCONF commands. These commands are mainly to manage the connection (e.g., connect, lock, unlock, close-session, kill-session, etc.), get the RIS current status (e.g., get, get-data, get-config, etc.), modify existing configurations (e.g., edit-config merge, edit-config replace, etc.) and register to notifications to receive specific information (e.g., subscribe) [18]. The O1+ termination at the RIS actuator will receive these commands triggered by a non-RT RIS rApp and apply the YANG configuration. These YAML configuration files will follow the RIS vendor specific configuration model and, e.g., for the case of RIS logical splitting, will contain information about which antenna elements correspond to which logical RIS ID.

On the other hand, the E2+ interface will mimic the E2 interface but for RIS specific operations without the need of the involvement of gNBs. As also mentioned in BeGREEN deliverable D4.2 Section 2.2.3 [14], the E2+ interface is similar to the E2 interface and has just two new RIS specific E2 Service Models (E2SM), i.e., E2SM Smart Surfaces Monitoring (E2SM-SSM) and E2SM Smart Surfaces Control (E2SM-SSC). E2SM-SSM is very similar to E2SM Key Performance Measurements (E2SM-KPM) [19], and reports RIS specific information such as the state (configurations, loaded codebooks, current phase-shifts, RIS energy consumption, etc.). Tentatively, it can also support reporting sensed information such as received signal at the RIS (e.g., RSSI) or direction of the main signal source (e.g., azimuth and elevation) in those RISs that support it. As in E2SM-KPM, this information is just sent through the E2AP REPORT service. This service is accordingly associated to a RIS controller xApp that has previously subscribed to the service. Therefore, such RIS xApp would be able to both receive RIS specific metrics from E2+ using the E2SM-SSM (e.g., RIS state) and available standard gNB metrics from E2 using the E2SM-KPM (e.g., L1M-UL-SRS-RSRP). Moreover, E2SM-SSC triggers modifications in the configuration of the RIS at near-RT intervals. As other control-based E2SM such as E2SM-RC (RAN Control), it will leverage the E2AP REPORT, INSERT, CONTROL, POLICY and QUERY services similarly.

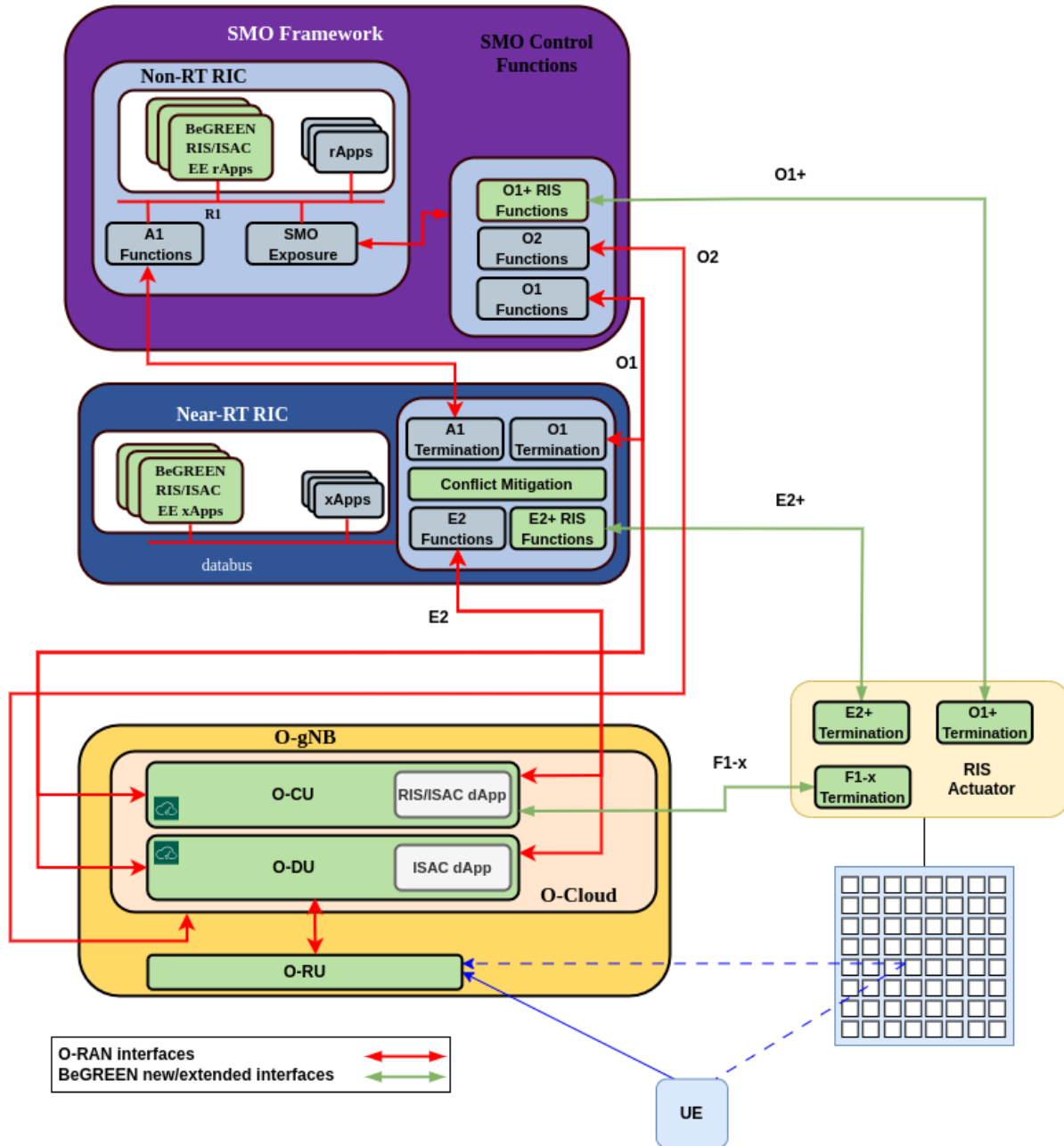


Figure 4-19 O-RAN based BeGREEN architecture focusing on the RIS/ISAC components

The most common procedures will be related with load a pre-calculated configuration from a codebook, configure the phase-shift per antenna element, and manage (load, copy, delete, etc.) codebooks. However, more complex operations, such as rotate N steps the offset of a configuration while keeping the azimuth and elevation, should be also supported. Such operations that abstract high level operations may be implemented by RIS actuators or dApps (see below) and, therefore, the E2SM-SSC need to support both end-point types. E2SM-SSC and E2SM-SSM will not necessarily always use all gNB specific Information Elements (IE) since the E2 termination points are smart surfaces that are not always bound to a single cell. While RIS serving 1 single cell may inherit IEs from the parent cell, this does not apply in multi-cell RISs. Therefore, in such cases, corresponding IEs such as Cell Global ID, Group ID, Core CP, QoS ID, Serving Cell PCI, Serving Cell ARFCN, Cell RNTI, etc. will be optional and padded with dummy values. Finally, as E2+ is just a simplified instance of E2 with two E2SMs that require fewer mandatory IEs, underlaying E2AP running over SCTP will be kept the same.

The architecture also provides an interface to locally interconnect gNBs to RIS at cell sites, the F1-x interface.

This interface was introduced in RISE-6G [20] to cover use cases with strict time deadlines such as RIS-gNB joint beamforming or use cases that trigger large amount of signalling traffic to control the RIS, such as frequent phase-shift modifications per antenna element. BeGREEN proposes to implement this F1-x in a dApp at the O-CU. dApps are distributed applications that enable RAN intelligence at the gNBs for real-time use cases [21]. Such RIS-specific dApp may use the standard E2 interface to communicate with the RIS controller xApps at the near-RT RIC and act as intelligent relay for the RIS actuator. To do so, a RIS dApp can use both E2SM-SSM and E2SM-SSC through the standard E2 interface as if they were two more service models. Unlike E2+, the E2 interface is bound to particular gNBs, and therefore, gNB-specific IEs such as Serving Cell PCI, Serving Cell ARFCN, etc. may be used as in a standard manner in both E2SM-SSM and E2SM-SSC. For instance, following this approach, a RIS controller xApp may use the standard E2 interface for both standard gNB metrics (e.g., L1M-UL-SRS-RSRP) and RIS specific metrics using the E2SM-SSM implemented at the RIS dApp. The implementation of F1-x will depend largely on the RIS dApp and the RIS hardware and may run over both wired and wireless channels with a varying range of capacity and communication latency.

Finally, BeGREEN architecture may also integrate ISAC and RIS-enabled ISAC solutions. Common ISAC use cases such as people detection [22] can easily be integrated with ISAC rApps/xApps that consume TS28.552 metrics (e.g., L1M.SS-RSRP, L1M.SS-RSRPNrNbr, DRB.AirIfDelayDL, etc.) and E2SM-KPM metrics (e.g., L1M-UL-SRS-RSRP) using the standard E2 interface. Other use cases such as user location based on AoA+ToF [23 and [24] in a multi-antenna gNB can leverage ISAC dApps deployed at the O-DU, where access to high bandwidth I/Q samples is available and RT computations are possible without violating timing deadlines or saturating the monitoring interfaces. Regarding RIS-enabled ISAC solutions, e.g., RIS-aided localization in MISO 1 RIS 1 BS scenarios [25], the BeGREEN architecture enables the joint use of RIS/ISAC rApps/xApps/dApps, E2 metric reporting procedures (E2SM-SSC and E2SM-SSM) and RIS specific interfaces to control/monitor the RIS (O1+, E2+ and F1-x). D4.3 will provide an example of a RIS-enabled ISAC solution integrated into the BeGREEN architecture.

4.4 Frequency assignment strategies in relay-enhanced scenarios

4.4.1 Frequency assignment strategies

The deployment of relays to improve a B5G network involves the decision of the radio resources, i.e. frequency bandwidth, that the relay will use for serving its users. A key aspect in this decision is the interference that can be produced in the different links involving the BS, the relay and the served UE, because an increase of interference can lead to an increase in power consumption. While this interference can be totally cancelled if the relay and the BS operate in different bandwidths, this will result in a poorer utilisation of the available spectrum. Thus, a trade-off has to be found between making an efficient use of the spectrum and at the same time sufficiently mitigating the interference to reduce the power consumption.

To deal with the above trade-off, a strategy that exploits the availability of different beams in a BS with massive MIMO was described in Section 4.4.2 of BeGREEN D3.2 [1]. In this strategy, referred here as *Spatial Beam (SB)-based* and illustrated in Figure 4-20, the relay and the BS operate with the same frequency carrier than the BS, but their signals are separated by a combination of both spatial and frequency multiplexing. In particular, the beams of the BS that point in the direction where the relay is located, defined as “Relay’s Incidence Area” (RIA) use a subset B_p of the total bandwidth of the carrier B , while the remaining bandwidth $B_R=B-B_p$ is used by the relay to communicate with its serving UEs. In contrast, the beams of the BS that point in other directions different than the RIA use the total bandwidth B .

The power consumption model of the SB-based strategy was formulated in Section 4.4.4.2 of BeGREEN D3.2. Based on this model, this deliverable presents a performance evaluation of this strategy and compares it against two benchmark strategies. The first benchmark is the *Out-band* strategy, in which the BS and the relay use a different portion of the spectrum.

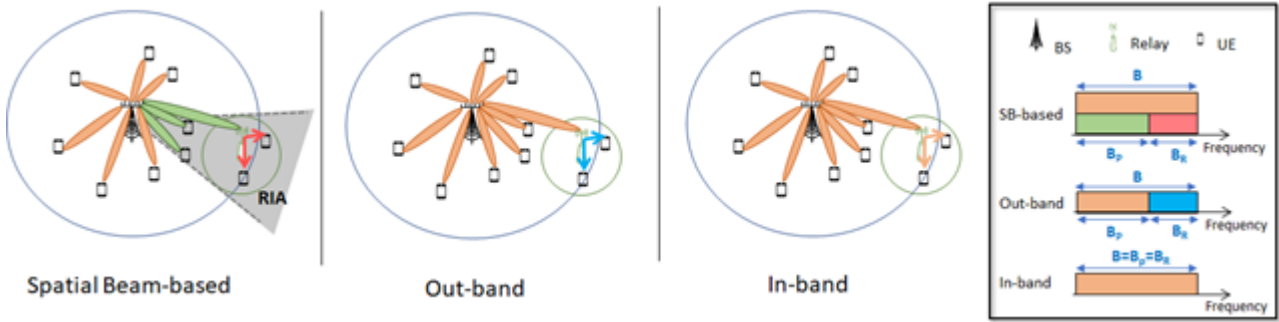


Figure 4-20 Frequency assignment strategies in a single-cell scenario

As seen in Figure 4-20 for a carrier of bandwidth B , the BS uses only the portion of bandwidth B_p in all its beams while the relay uses only the bandwidth B_r . The power consumption model of the out-band strategy was formulated in Section 4.4.4.1 of BeGREEN D3.2.

The second benchmark is the *In-band* strategy, in which the BS and the relay operate using the same spectrum, i.e. the total bandwidth B , as seen in Figure 4-20. A power consumption model of this in-band strategy was presented in Section 4.4.4.3 of BeGREEN deliverable D3.2 assuming a full-duplex transmission at the relay, i.e. allowing that the relay and the BS transmit at the same time. However, initial analyses showed that this full-duplex transmission was not feasible due to excessive self-interference at the relay. Then, in order to avoid this problem, half-duplex transmission is considered with the in-band strategy. This means that the transmission and reception at the relay node are carried out at different time instants. To do this, the time domain should be appropriately split to accommodate the UL and DL transmissions and receptions in 3 different links, namely the BS access link between the BS and its served UEs, the relay backhaul link between the BS and the relay, and the relay access link between the relay and its served UEs. Then, considering that both the relay and the BS operate with frequency division duplex (FDD) the transmission and reception is organized in two slots as shown in Figure 4-21. During the first slot, the BS transmits in the DL frequency the information of the BS access link and the relay backhaul link (these links are separated by using different beams), while the relay receives in the UL frequency the information of the relay access link and in the DL frequency the information of the relay backhaul link (see Figure 4-22). In contrast, the opposite occurs during the second slot. In this case, the relay transmits in the DL frequency the information of the relay access link and in the UL frequency the information of the relay backhaul, while the BS receives in the UL frequency the information of both the BS access and the relay backhaul links.

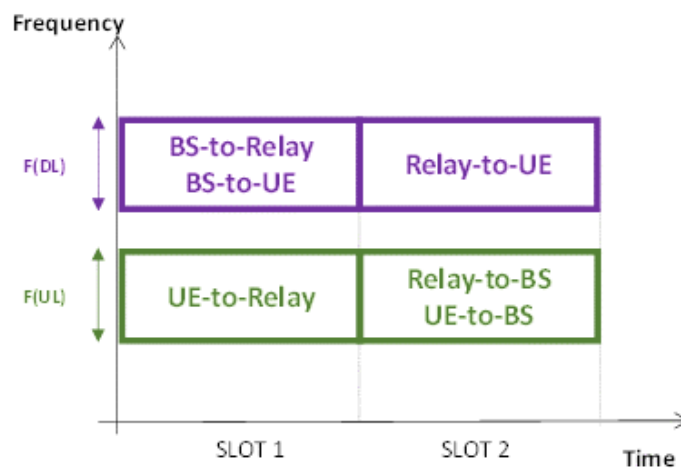


Figure 4-21 Time-frequency resource usage for the in-band strategy with half-duplex

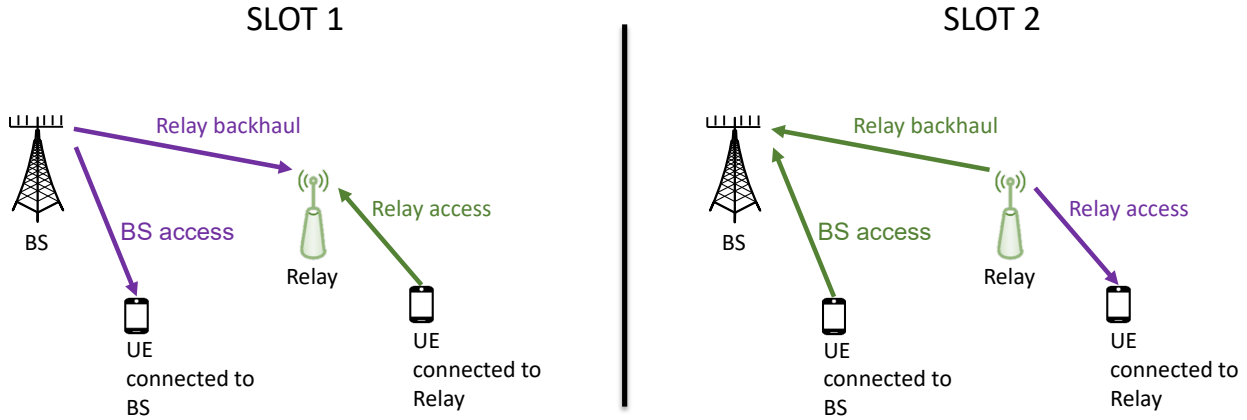


Figure 4-22 Uplink and downlink use in the two slots for the in-band strategy with half-duplex

With this approach, the SINR obtained by the users connected to the BS is affected by the interference that results from the non-orthogonality of the channels of the different users connected to the BS (relay included), and from the beamforming gain uncertainty. As for the UEs served by the Relay, they do not suffer any interference. Therefore, the model formulation in this case is the same as for out-band strategy in BeGREEN D3.2, Section 4.4.4.1 [1]. The only difference is that the efficiency factor ε in equation (4-37) of [1] must consider that the links only operate half the time and, therefore, the efficiency factor is half that of the out-band case.

The scenario considered to analyze the performance of the three strategies, the parameters considered, and the results obtained are presented in the next subsections.

4.4.2 Scenario under study

The considered scenario consists of a 5G NR BS that uses massive MIMO with M antennas, and a relay, connected to the BS, located within its coverage area, which covers a shadow area generated by a large obstacle (like a building, or wall), as shown in Figure 4-23. Communication in the DL direction is considered.

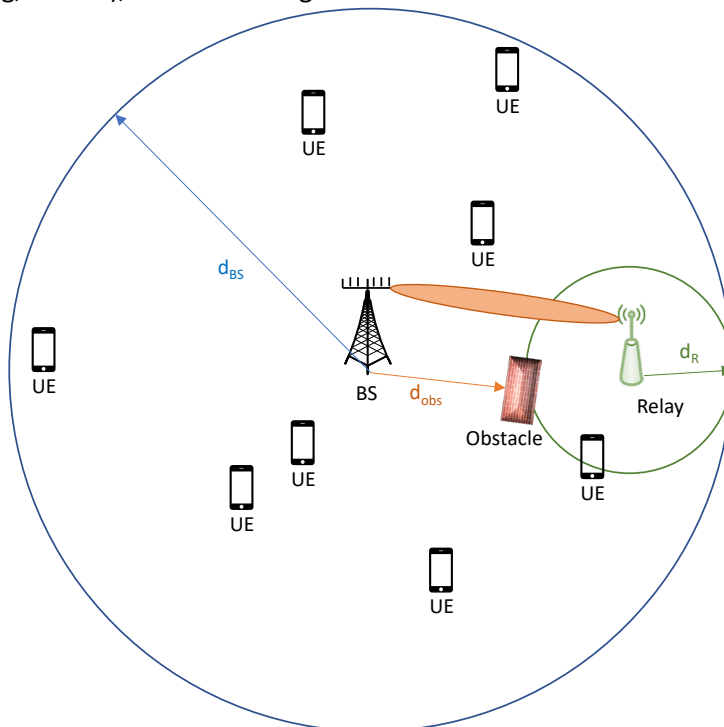


Figure 4-23 Scenario under study

The relay does not use massive MIMO and is just modeled as one of the other UEs. There are K single-antenna UE devices (including the relay, UE_k) numbered as $k=1, \dots, K$, with $M \gg K$. The relay has a fixed position and the rest of UEs are uniformly distributed throughout the coverage area and associated with the BS or the relay depending on which path has lower propagation losses. The number of UEs connected through the relay must be $K_R \geq 1$. Each UE connected to directly the BS or through the relay requires a certain bit rate denoted as R_k for $k=1, \dots, K-1$. The bit rate of the relay is given by the aggregate of all UEs served by it.

The configuration parameters include the frequencies at which the BS and the relay operate (f_{BS}, f_R), the number of antennas at the BS (M), the antenna gains of the BS, Relay and UE (G_B, G_R, G_U), the channel estimation quality (c_{CSI}), the fraction of bandwidth assigned to relay users (α_k), the bit rate required by the UEs (R_k). The model to calculate the pathloss with the BS (L_{BS}) and the pathloss with the relay (L_{Relay}), and the size of the BS coverage area (d_{BS}) and the relay (d_R) are also set. The values for the noise power spectral density (N_0), the efficiency factor (ε), and the parameters of the obstacle are also defined. Table 4-2 shows the values used in this study, related to the BS, the relay, the UEs, and the scenario.

Table 4-2 Configuration Parameters

BS		Relay		UE		Scenario	
f_{BS}	2.6 GHz (n7)	f_R	2.1 GHz (n1)*	G_U	3 dB	N_0	-168 dBm/Hz
M	64	α_k	$1/K_R$	R_k	50 Mbps	ε^{**}	0.8
G_B	10 dB	G_R	3 dB	L_{Relay}	UMi	L_{obs}	40 dB
d_{BS}	1000 m	d_R	200 m	L_{BS}	UMa	d_{obs}	510 m
c_{CSI}	1					w_{obs}	200 m

* Only in out-band strategy. For the other two strategies: $f_R = f_{BS}$

** Efficiency factor in full duplex case. For half-duplex it is 0.4.

To conduct a fair comparison, the total bandwidth used B is the same for all three strategies, and the allocation to the BS and relay in each strategy is shown in Table 4-3 according to the splits shown in Figure 4-20.

Table 4-3 Spectrum Distribution

Strategy	B	B_P	B_R
Out-band	50 MHz	30 MHz	20 MHz
In-band	50 MHz	50 MHz	50 MHz
Spatial-Beam-based	50 MHz	30 MHz	20 MHz

Table 4-4 Combinations of Power Consumption Parameters

Combination	BS			Relay		
	a_{BS}	$P_{0,BS}$	Ref.	a_R	$P_{0,R}$	Ref.
1	28.4	156.38 W	[3]	20.4	13.91 W	[2]
2	2.57	12.85 W	[2]			
3	4.7	130 W	[4]			
4	2.8	84 W	[4]			
5	28.4	156.38 W	[3]	4	6.8 W	[4]
6	2.57	12.85 W	[2]			
7	4.7	130 W	[4]			
8	2.8	84 W	[4]			

Regarding the total power consumption at the BS and the relay, a model linearly dependent on the transmitted powers $P_{T,BS}$ and $P_{T,R}$ including a constant term per antenna that accounts for the circuit power consumption has been considered, as explained in section 4.4.4 of deliverable D3.2. The scaling factors of the transmitted power at the BS and relay are a_{BS} and a_R respectively. Similarly, $P_{o,BS}$ and $P_{o,R}$ are the circuit power consumption per antenna at the BS and relay, respectively. Several combinations for these parameters, shown in Table 4-4, have been analysed.

4.4.3 Performance assessment

The evaluated KPIs to analyse the performance are:

- **Transmitted power** at BS ($P_{T,BS}$) and Relay ($P_{T,R}$), as formulated in BeGREEN D3.2 [1].
- **Total power consumption** (P_{TOT}), as formulated in BeGREEN D3.2 [1].
- **Energy efficiency** (EE), that is the number of bits transmitted per unit of energy, calculated as: $EE = R_k / P_{TOT}$ (bits/J), where P_{TOT} is the total consumed power.

The first performed analysis involves UEs distributions where at least one UE is connected through the relay. Combination 1 from Table 4-4 is selected for power consumption parameters, and the number of users in the scenario (K) increases from 3 to 32. Specifically, we have analysed two cases, namely with one UE connected through the relay and with two UEs connected through the relay.

Figure 4-24(a) and Figure 4-24(b) show the transmitted power at the BS and relay respectively for the three strategies in the two cases.

In the case of 1 UE connected through the relay, it can be observed that the SB-based strategy can provide the required data rate to all the users in all the analysed situations with a maximum of 32 users in the scenario. On the other hand, out-band strategy requires less power for $K < 13$ users since there is no relay interference, but can only serve up to 20 UEs due to SINR requirements. Note that the SINR requirement of UEs is higher in out-band strategy because the spectrum used to serve them (B_p) is less. Similarly, in-band strategy can only serve up to 14 UEs, and is the one that requires the most power.

In the case of 2 UEs connected through the relay, it can be observed that both BS and relay require more transmitted power, and strategies out-band and in-band reduce the number of users they can serve (out-band reduces from 20 to 17 users, and in-band reduce from 14 to 10 users) while Spatial Beam-based strategy continues to serve all distributions up to $K=32$.

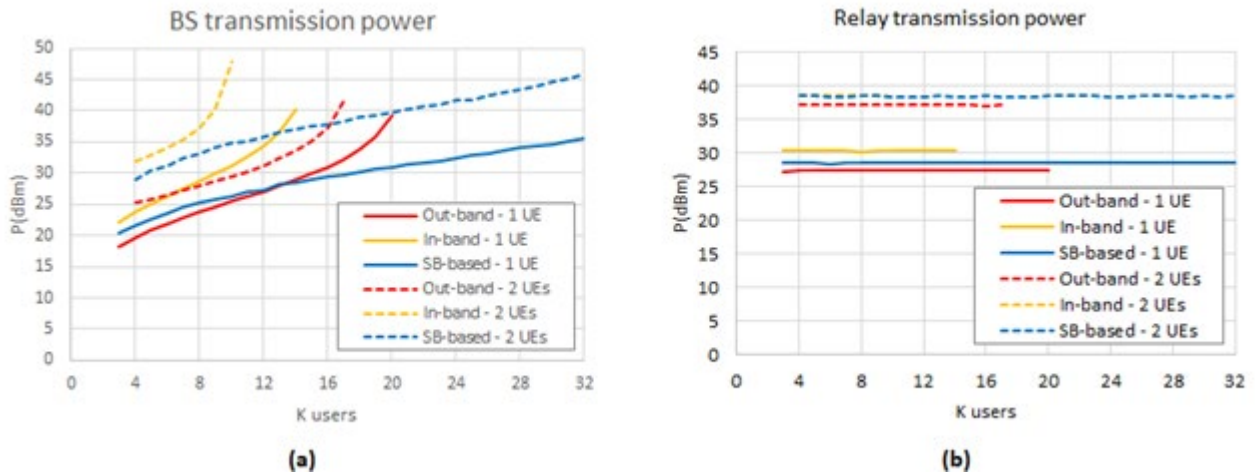


Figure 4-24 Transmitted power at the BS (a) and relay (b) when 1 or 2 UEs are connected through the relay

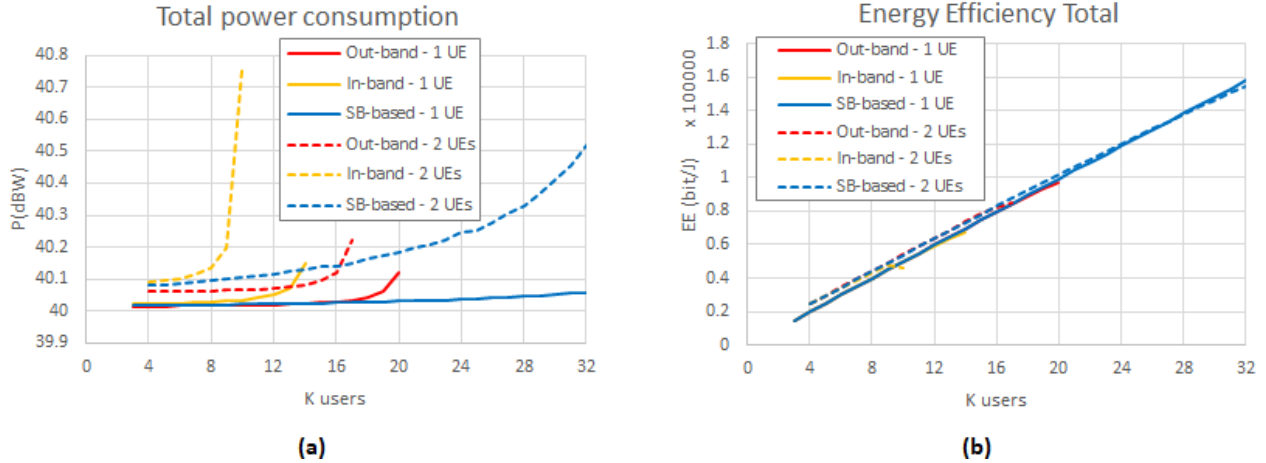


Figure 4-25 Total power consumption (a) and EE (b) when 1 or 2 UEs are connected through the relay

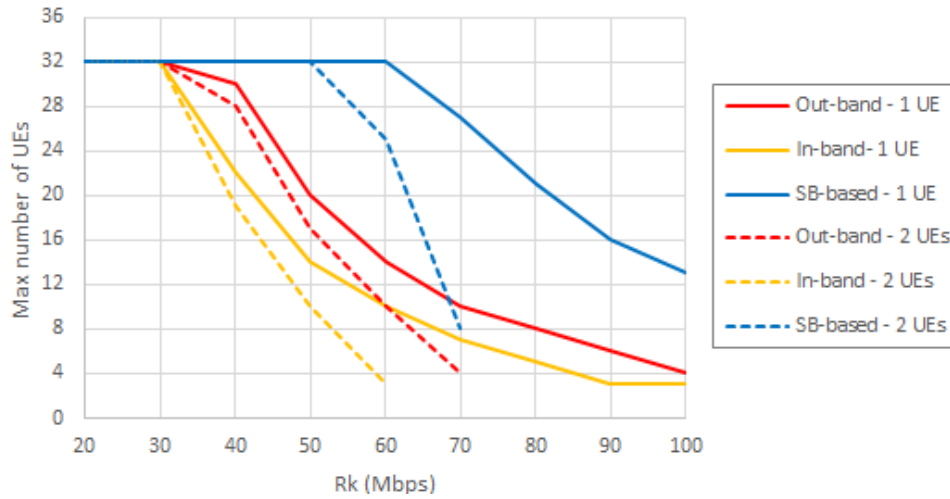


Figure 4-26 Maximum number of supported users as a function of the required bit rate R_k

When computing the total power consumption, see Figure 4-25 (a), in the case of 1 UE, strategies out-band and in-band are quite similar up to 15 users and from here out-band strategy grows to 40.119 dBW for 20 users due to the required total transmitted power of the BS. In-band strategy shows the same behaviour with less users (up to 8) and more power. In the case of 2 UEs, the total power consumption increases and the number of users served in strategies out-band and in-band is lower. Despite the differences mentioned, EE is practically the same in the three strategies and for the two cases, as shown in Figure 4-25 (b).

The results in previous figures reveal that there exists a limit in the maximum number of UEs that can be served by the BS and/or relay. After exceeding this limit, it is not possible to fulfil the SINR requirements. To further study this effect, Figure 4-26 plots the maximum number of UEs that can be served by each strategy as a function of the required bit rate of a UE (R_k) when it varies between 20 and 100 Mbps. It is observed that the SB-based strategy supports the highest number of UEs and therefore provides more capacity to the system. We also observe that when there are 2 UEs connected through the relay the total number of connected users decreases in all strategies with respect to the case with 1 UE connected through the relay. Moreover, in this case the strategies cannot guarantee R_k greater than 60 Mbps in the case of in-band and 70 Mbps in the other two cases.

Figure 4-27 depicts the achieved EE as a function of the required bit rate. It is observed that it increases with the required bit rate up to a maximum and then begins to decrease, widening the differences between strategies.

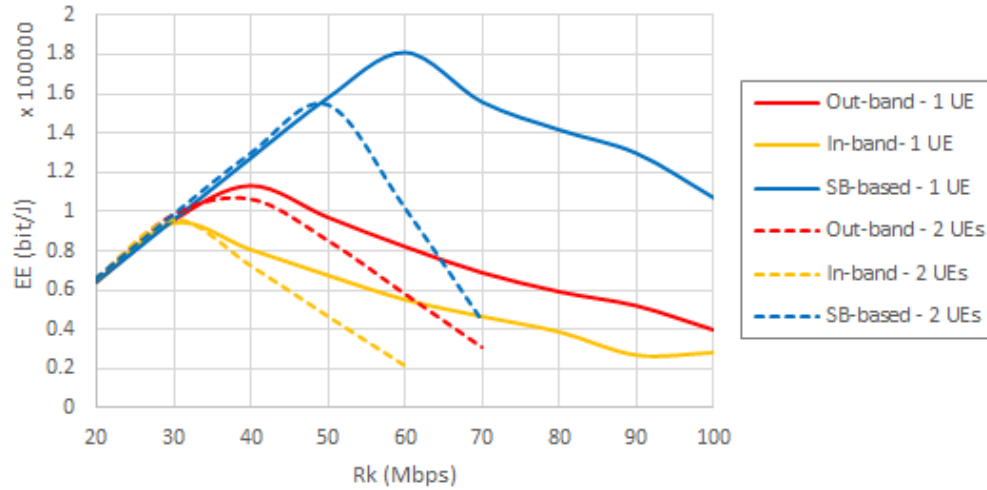


Figure 4-27 EE as a function of the required bit rate R_k

The maximum achieved EE is higher with the SB-based strategy than with the out-band and in-band. Specifically, in the case presented here with 1 UE connected through the relay, the SB-based strategy achieves a maximum EE of 1.81×10^5 bits/J for an R_k of 60 Mbps, while the out-band strategy achieves a maximum of 1.13×10^5 bits/J for an R_k of 40 Mbps, and the in-band strategy achieves a maximum of 0.94×10^5 bits/J for an R_k of 30 Mbps. Then, the SB-based strategy improves the maximum EE in 60% with respect to the out-band strategy when there is 1 UE connected through the relay, and in 45% in the case of 2 UEs. Similarly, in relation to the in-band strategy, the SB-based strategy improves the EE in 92% in the case of 1 UE and in 61% in the case of 2 UEs.

Another study has been to compare the different combinations of power consumption parameters shown in Table 4-4. Results for the total power consumption are shown in Figure 54. It is observed that there are large differences, higher than 10 dB, depending on the values used for these parameters. Combinations 2 and 6 are the ones that require less power since the circuit power consumption at the BS is the lowest. In contrast, combinations 1 and 5 are the ones that require more power for the opposite reason, the circuit power consumption at the BS is the highest. It is also worth noting that the values of the power consumption model do not affect the comparison between strategies, and in all the cases the SB-based strategy outperforms the other two in terms of number of supported UEs and power consumption as the number of UEs increases.

To better see the comparison between the different combinations, Figure 4-29 (a) plots the results with the Spatial Beam-based strategy for all combinations of power consumption values. It is observed that combinations 1, 2, 3 and 4 consume the same as combinations 5, 6, 7 and 8, so we conclude that the relay parameters do not significantly affect the total power consumed, while those of the BS do. It can be checked that combination 1 is the one that consumes the most power (over 40 dBW), followed by combination 3 and combination 4 which consume, respectively, 0.8 dB and 2.7 dB less. And finally, combination 2 is the one that consumes the least power, with a value more than 10 dB below the maximum.

The differences in total power consumed are reflected in EE as shown in Figure 4-29 (b) where the total EE with Spatial Beam-based strategy is plotted for all the combinations of power consumption parameters. As expected, EE is better if combinations 2 or 6 are considered, as they consume less power. For example, for $K=20$ users, combinations 2 and 6 provide an EE increase of 1100% over combinations 1 and 5, 880% over combinations 3 and 7, and 540% over combinations 4 and 8.

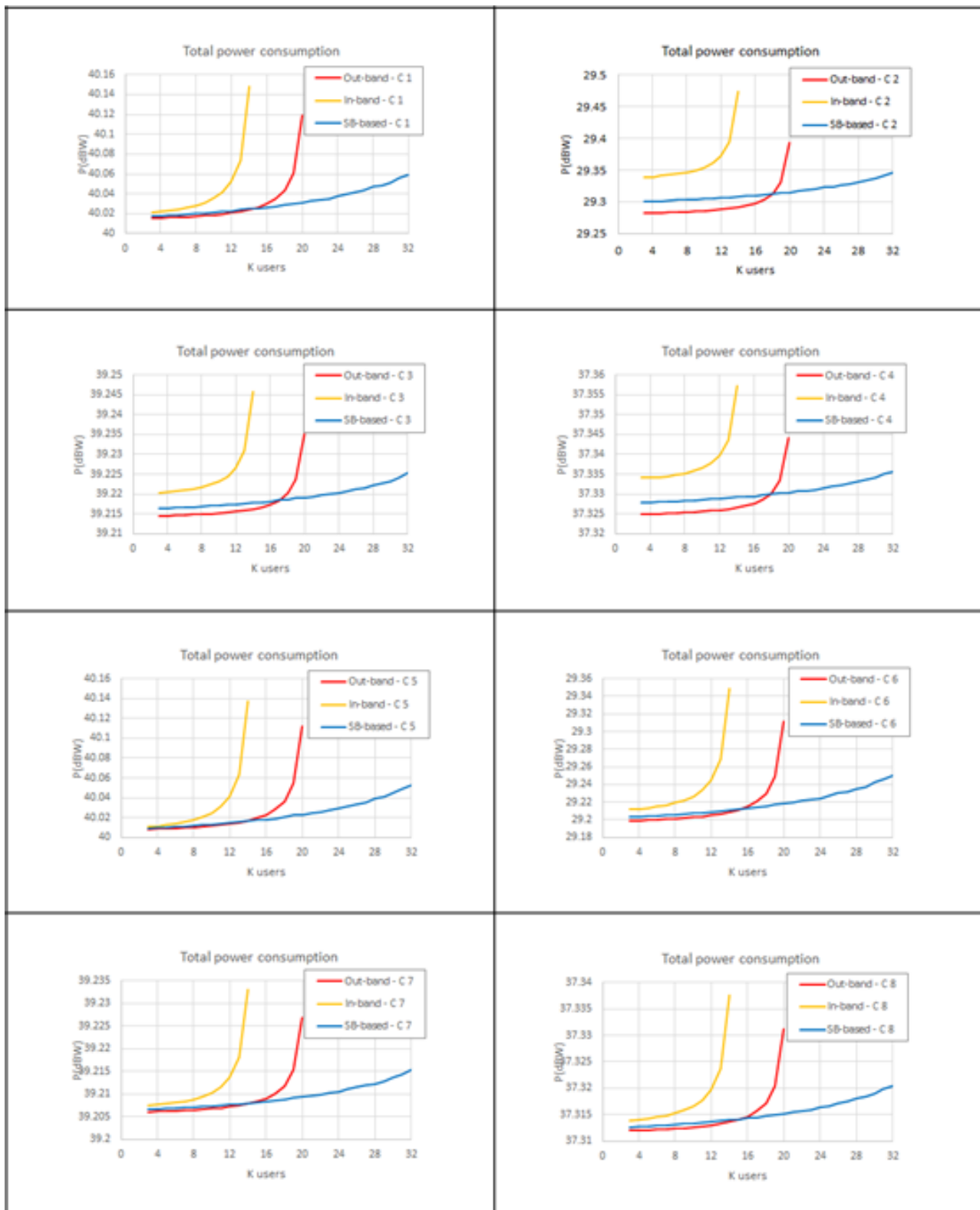
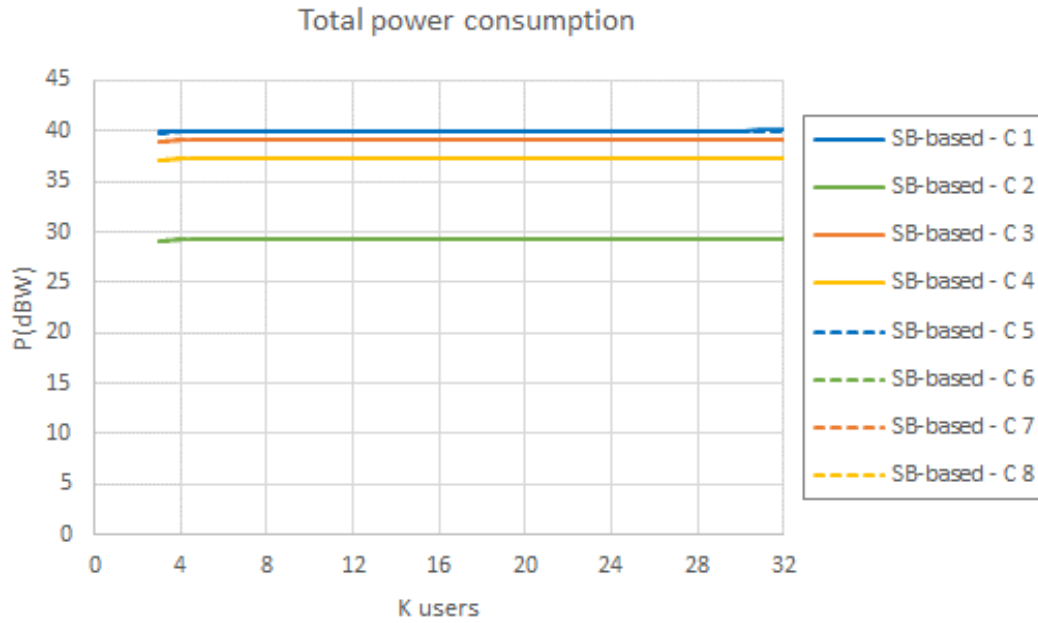
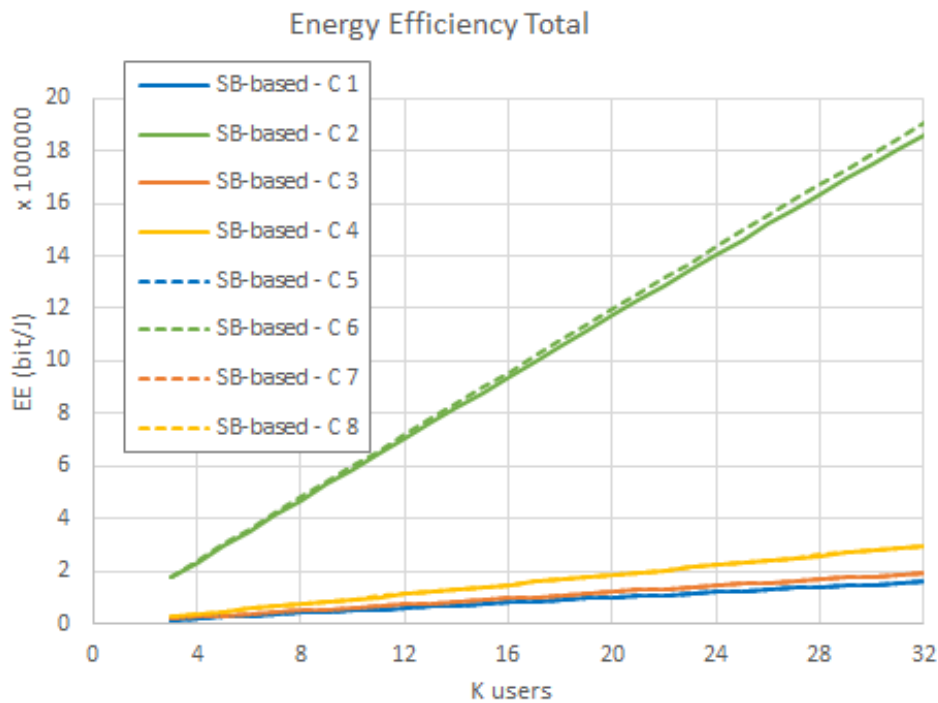


Figure 4-28 Total power consumption for all combinations of power consumption values



a)



b)

Figure 4-29 Comparison with Spatial Beam-based of total power consumption (a), and total EE (b)

4.4.4 Conclusions

In this section, the proposed SB-based strategy to assign frequencies between relay and BS has been studied, comparing it with two benchmark strategies: out-band and in-band. The aim of SB-based strategy is to make an efficient use of the spectrum and at the same time sufficiently mitigating the interference to reduce the power consumption by exploiting both the spatial and frequency domains when separating the signals of the BS and the relay.

Results show that the SB-based strategy provides the highest capacity in terms of number of supported UEs compared to the two benchmark strategies. For example, when the bit rate required by the UEs is $R_k=50$ Mbps, SB-based strategy can allocate 32 UEs, whereas the out-band strategy only supports 20 UEs and the in-band strategy only 14 UEs.

Besides that, the SB-based strategy also improves significantly the EE. Specifically, the presented results have shown that the SB-based strategy improves the maximum EE in 60% and 92% with respect to the out-band and in-band strategies, respectively, when there is 1 UE connected through the relay. In turn, when there are 2 UEs served by the relay, the improvements are 45% and 61%, respectively.

This study has been done with different values of the power consumption model parameters, concluding that, although they have a significant impact on the absolute values of energy consumption, they do not impact substantially the relative comparison between strategies.

5 Summary and Conclusion

The main objective of BeGREEN is to investigate the possible EE improvements of the RAN in future mobile networks. BeGREEN D3.3 presented the final result obtained in the BeGREEN WP3, where PHY layer technologies that can be used for direct or indirect EE improvement of the overall mobile network are investigated and developed. To directly improve the EE of the RAN, two approaches are pursued. First, optimization of the BB processing, i.e. LDPC decoder; and, second, optimization of the PA operation in the RU. The technologies that can indirectly improve the EE of the network, considered in WP3, are ISAC, RIS and radio relays. These technologies mainly provide information or service which need less energy, but allows for the same functionality of the network.

The BB processing of the received waveforms in the DU is highly complex and requires a large amount of energy. The highest energy consumption is in the LDPC decoder. This means that optimizing the LDPC decoder has the highest potential for EE improvement. BeGREEN D3.3 extends the implementation and testing of the LDPC decoder. In addition to the implementation done on x86 and ARM, the LDPC decoder is implemented on a GPU. Several optimization iterations are carried out to be able to efficiently utilize the GPU and use all its available threads. The results showed that the LDPC wireless performance of the GPU was the same as the x86 and the ARM as expected. The power consumption of the GPU was higher when compared to ARM CPU using multiple cores. However, when comparing to using only a single ARM core the GPU was much more energy efficient.

BeGREEN D3.3 presents a comparison of energy saving and power consumption between ARM and x86 servers using the A1 policy service. The results show key differences in their power consumption behaviour, showing that ARM servers are more efficient due to their dynamic power allocation mechanism, which adjusts based on CPU activity and the Energy Savings system, achieving reductions of up to 5%. In contrast, x86 servers exhibit significantly higher overall power consumption, with minimal impact from the Energy Savings system (less than 1%). These findings suggest that ARM servers offer better adaptability for EE under variable workloads.

The RU has the highest power consumption in the RAN. Therefore, optimizing its EE has the highest potential for improvements. This deliverable has summarised the potential power saving of a RU PA blanking module that turns the PAs off when there is no data to be transmitted in the downlink. From the obtained results, it can be concluded that more than 50% of the power consumed by the RU PAs, which are one of the major power users that can be saved in the cellular network.

ISAC is a new technology that offers radar like sensing with the available radios in the RU. The main approaches were pursued to use ISAC in order to improve the EE of the network. The first one is to use ISAC to assist the beam training process to make it more attractive for use in different RUs and wireless technologies. Our analysis showed that this approach is superior to the state-of-the-art approaches, which will solve the issues due to which electronic beam steering technologies are not widely and fully used. The second approach is to use ISAC to detect potential users. This will enable better network resource allocation and will reduce the idling time of the devices in the RAN. The obtained results show excellent results in terms of range and angular resolution, allowing for easy estimation of the number of potential users.

Finally, BeGREEN D3.3 provides contributions on the frequency assignment for relay-enhanced scenarios, proposing the Spatial Beam (SB)-based strategy. It exploits the spatial and frequency domains for separating the signals of the BS and the relay in order to achieve a trade-off between the efficient use of the spectrum and the generated interference. This strategy has been compared against two benchmarks, namely the out-band and in-band strategies. Results have shown that the SB-based strategy provides higher capacity in terms of number of supported users than the two benchmark strategies and it improves the maximum achieved EE in up to 60% and 92% with respect to the out-band and in-band strategies, respectively.

Bibliography

- [1] BeGREEN D3.2, "Initial Developments and Evaluation of the Proposed Enhancements and Optimization Strategies", August 2024. <https://doi.org/10.5281/zenodo.13840393> Available Online: <https://www.sns-BeGREEN.com/deliverables>
- [2] E. Björnson, M. Kountouris and M. Debbah, "Massive MIMO and small cells: Improving energy efficiency by optimal soft-cell coordination," ICT 2013, Casablanca, Morocco, 2013, pp. 1-5, doi: 10.1109/ICTEL.2013.6632074.
- [3] R. Fantini, D. Sabella, M. Caretti, "An E3F based assessment of energy efficiency of Relay Nodes in LTE-Advanced networks", PIMRC, 2011.
- [4] G. Auer et al., "How much energy is needed to run a wireless network?", in IEEE Wireless Communications, vol. 18, no. 5, pp. 40-49, October 2011, doi: 10.1109/MWC.2011.6056691.
- [5] BeGREEN D3.1, "State-of-the-Art on PHY Mechanisms Energy Consumption and Specification of Efficiency Enhancement Solutions", August 2024. Available Online: <https://www.sns-BeGREEN.com/deliverables>
- [6] 3GPP TS 38.212, "NR: Multiplexing and channel coding (Release 18)", September 2023.
- [7] Wu, Qingqing, and Rui Zhang. "Beamforming optimization for wireless network aided by intelligent reflecting surface with discrete phase shifts." IEEE Transactions on Communications 68.3 (2019): 1838-1851.
- [8] He, Jiguang, et al. "Adaptive beamforming design for mmWave RIS-aided joint localization and communication." 2020 IEEE Wireless Communications and Networking Conference Workshops (WCNCW). IEEE, 2020.
- [9] Boles, T., et al. "AlGaAs PIN diode multi-octave, mmW switches." 2011 IEEE International Conference on Microwaves, Communications, Antennas and Electronic Systems (COMCAS 2011). IEEE, 2011.
- [10] Mursia, Placido, et al. "RISMA: Reconfigurable intelligent surfaces enabling beamforming for IoT massive access." IEEE Journal on Selected Areas in Communications 39.4 (2020): 1072-1085.
- [11] Gapeyenko, Margarita, et al. "On the temporal effects of mobile blockers in urban millimeter-wave cellular scenarios." IEEE Transactions on Vehicular Technology 66.11 (2017): 10124-10138.
- [12] Zappone, Alessio, et al. "Overhead-aware design of reconfigurable intelligent surfaces in smart radio environments." IEEE Transactions on Wireless Communications 20.1 (2020): 126-141.
- [13] Rouissi, Ines, et al. "Design of a frequency reconfigurable patch antenna using capacitive loading and varactor diode." 2015 9th European conference on antennas and propagation (EuCAP). IEEE, 2015.
- [14] BeGREEN D4.2, "Initial Evaluation of BeGREEN O-RAN Intelligence Plane, and AI/ML Algorithms for NFV User-Plane and Edge Service Control Energy Efficiency Optimization", August 2024. Available Online: <https://www.sns-BeGREEN.com/deliverables>
- [15] BeGREEN D5.2, "Solution Integration and Validation", February 2025. Will be Available Online: <https://www.sns-BeGREEN.com/deliverables>
- [16] BeGREEN D2.2, "Evolved Architecture and Power Enhancement Mechanisms", July 2024. Available Online: <https://www.sns-BeGREEN.com/deliverables>
- [17] <https://manpages.ubuntu.com/manpages/jammy/man1/netopeer2-cli.1.html>
- [18] <https://datatracker.ietf.org/doc/html/rfc6241>
- [19] O-RAN E2 Service Model (E2SM) KPM 6.0; <https://orandownloadweb.azurewebsites.net/download?id=810>
- [20] RISE D2.6, "RISE network architectures and deployment strategies analysis: final results", https://rise-6g.eu/Documents/LIVRABLES/RISE-6G_WP2_D2.6_FINAL.pdf
- [21] dApps for Real-Time RAN Control: Use Cases and Requirements: <https://mediastorage.o-ran.org/ngrg-rr/nGRG-RR-2024-10-dApp%20use%20cases%20and%20requirements.pdf>

- [22] Devoti, Francesco, Vincenzo Sciancalepore, and Xavier Costa-Perez. "Passive and Privacy-Preserving Human Localization: A Social Distancing Approach Using Commercial Millimeter-Wave Access Points." *IEEE Vehicular Technology Magazine* 17.4 (2022): 101-109.
- [23] Menta, Estifanos Yohannes, et al. "On the performance of AoA-based localization in 5G ultra-dense networks." *IEEE Access* 7 (2019): 33870-33880.
- [24] Liu, Shengheng, et al. "5G NR monostatic positioning with array impairments: Data-and-model-driven framework and experiment results." *Proceedings of the 3rd ACM MobiCom Workshop on Integrated Sensing and Communications Systems*. 2023.
- [25] Kamran Keykhosravi, et al, "Leveraging RIS-enabled smart signal propagation for solving infeasible localization problems: Scenarios, key research directions, and open challenges". *IEEE Vehicular Technology Magazine* 18, 2 (2023), 20–28.