Deliverable 3.2

Initial Developments and Evaluation of the Proposed Enhancements and Optimization Strategies

August 2024

Co-funded by
the European Union

6GSNS

Contractual Date of Delivery:	**May 31, 2024**
Actual Date of Delivery:	**August 6, 2024**
Editor(s):	**Ory Eger (PW)**
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)** **Revaz Berozashvili, German Castellanos (ACC)** **Vladica Sark, Jesús Gutiérrez, Mert Özates (IHP)** **J. Xavier Salvat, Jose A. Ayala, Marco Rossanese, Placido Mursia (NEC)** **Esteban Municio (I2CAT)** **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	16.11.2023	Vladica Sark (IHP)	Initial document created
0.2	07.03.2024	Govindarajan Mohandoss (ARM)	Appendix
0.3	21.03.2024	Anna Umbert (UPC)	Added interference analysis in relay-enhanced scenarios (Section 4.4)
0.4	30.04.2024	Ory Eger (PW)	Added Section 2.1 and 2.2
0.5	16.05.2024	Revaz Berozashvili (ACC)	Added RU on OFF description and CU HW acceleration analysis (Section 3.1)
0.6	20.05.2024	Israel Koffman (REL) Eli Shasha (REL) Baruch Globen (REL)	Added Section 3.2
0.61	05.06.2024	Ory Eger (PW)	Revised Section 2.1 and 2.2
0.7	10.06.2024	J. Xavier Salvat (NEC)	Completed Section 4.2
0.71	15.07.2024	Anna Umbert (UPC)	Chapter 4 revision
0.8	20.07.2024	Vladica Sark (IHP) Jesús Gutiérrez (IHP)	Section 4.1 and 4.3 completed
0.81	29.07.2024	Jesús Gutiérrez (IHP)	Revised Section 3.2
0.9	29.07.2024	Jesús Gutiérrez (IHP) Mir Ghoraishi (GIGASYS)	Final revision and proof reading
1.00	06.08.2024	Simon Pryor (ACC)	Submission to the participant portal

Table of Contents

List of Figures

List of Tables

List of Acronyms

3GPP	3rd Generation Partnership Project
5GC	5G Core
5GNR	5G New Radio
AI	Artificial Intelligence
AoA	Angle-of-Arrival
API	Application Programming Interface
AWGN	Additive White Gaussian Noise
B5G	Beyond 5G
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
CU-CP	Centralized Unit – Control Plane
CU-UP	Centralized Unit – User Plane
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 Isotopically Radiated Power
ET	Envelope Tracking
FFT	Fast Fourier Transform
FMCW	Frequency-Modulated Continuous Wave
FP	Fractional Program
GPU	Graphics Processing Unit
HRIS	Hybrid Reconfigurable Intelligent Surfaces
HW	Hardware
IFFT	Inverse Fast Fourier Transform
IOS	Internet-of-Surfaces
IP	Internet Protocol
ISM	Industrial Scientific and Medical
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

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
near-RT	near Real-Time
NFV	Network Function Virtualisation
NLoS	Non-Line-of-Sight
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
PDCP	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
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
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
UE	User Equipment

UL	Uplink
ULA	Uniform Linear Array
UPA	Uniform Planar Array
UPS	User Plane Stack
UWB	Ultra-Wideband
WR	White Rabbit
XDP	eXpress Data Path
WP	Work Package

Executive Summary

BeGREEN D3.2 presents the initial developments and evaluation of the proposed physical (PHY) layer enhancements and optimisation strategies. While BeGREEN D3.1 described State-of-the-Art (SotA) on PHY mechanisms energy consumption and provided suggestions on innovative solutions and enhancements for this field [1], This document gets into the details of the specific solutions, indicates their potential for improving energy efficiency in cellular networks and presents some performance and power measurement results.

A wide range of solutions and modules are discussed, each with its unique angle on the O-RAN/3GPP framework for 5G deployments, as well as forward looking to 6G. On the distributed unit (DU) and central unit (CU) accelerations, it is shown advanced architectures can be employed to reduce power consumption of modules with high computational complexity. Initial results of the comparative analysis of CU performance using x86 and Arm platforms indicates that while x86 shows better energy efficiency in scenarios with a limited number of simultaneous users, the Arm architecture excels in large-scale, high-throughput environments, achieving greater overall power savings. Specifically, as the data throughput increases, the energy savings with Arm become more pronounced compared to x86, as this is attributed to Arm's efficient handling of parallelized tasks.

On terms of DU, for heavy processes algorithms, such as Low-Density Parity Check (LDPC) decoding, in the DU, it is observed that the Arm consumes less power, but is perform less processing than x86 for the same period. For Arm implementations with a larger vector length, much better ratios could be observed.

The radio unit (RU) consumes most energy in the radio access network (RAN) and in the cellular network, and thus the document explores several power saving techniques through simulation. It is observed that artificial intelligence (AI) based Digital-Pre-Distortion and Envelope Tracking can lead to more than 60% RF power amplifier (PA) energy saving. In addition, the PA blanking technique introduced in BeGREEN D3.1 [1] that powers-off the RU PA when there is no data in the downlink stream, was developed and implemented on a commercial RU and a PA energy consumption reduction higher of more than 40% is achieved.

The document introduced integrated sensing and communication (ISAC) to detect active users, and to assign resources based on the user distribution. To be used for extending the coverage of the RAN in urban areas, using Reconfigurable Intelligent Surface (RIS) is proposed and a practical autoconfiguring RIS integration is introduced, where its functionality and contribution in energy efficiency improvement is being evaluated.

In the end, several strategies for interference management to improve the network energy efficiency are proposed and evaluated. The power consumption model for three different spectrum allocation strategies is developed. These differ in the way how the spectrum is shared and managed between the base station and the relay. The impact of these strategies on the system's power consumption is currently being studied with simulations, and the results are expected in the upcoming BeGREEN D3.3.

1 Introduction

BeGREEN main goal is developing new strategies to reduce energy consumption in the components of radio access network (RAN). Specifically, Work Package 3 (WP3), focuses on developing enablers to significantly reduce the power consumption of the radio, i.e., stack layers 1 to 3 (L1/L2/L3) in 5G and beyond 5G (B5G) systems. These are the most resource and energy hungry functions in the RAN. For instance, amplifying the transmitted signals and ensuring its high fidelity has been reported as one of the top contributors to energy consumption in mobile networks [1]. Furthermore, performing complex baseband processing algorithms (e.g. decoding and channel estimation), required for handling different channel conditions, can also increase energy consumption [2].

BeGREEN D3.1 presented a survey of State-of-the-Art (SotA) energy consumption in L1/L2/L3. The performance of current solutions, namely: (i) hardware (HW) acceleration for Distributed Unit (DU) and Centralized Unit (CU) processing; (ii) Radio Unit (RU) energy saving methods; and (iii) xApps/rApps for controlling the RU (in the energy consumption context); were analysed. The use of Integrated Sensing and Communication (ISAC), reconfigurable intelligent surfaces (RISs), and relays, for improving network energy efficiency were also introduced. And in the end, basic concepts, ideas, directions, and architectures for energy and power saving, to be developed BeGREEN WP3, were discussed.

This document, BeGREEN D3.2, describes the progress on the development of proposed enablers. It aims at providing an initial description for these developments and presenting the initial results. This document drafts each solution's potential and advantage for energy and power consumption reduction.

The document is organized as follows:

- Chapter 1 provides a short introduction and presents the technical objectives covered by WP3.
- Chapter 2, discussed and presented the achievable reduction in power consumption when running the Low-Density Parity Check (LDPC) decoder and CU-UP (User Plane) algorithms on an Arm-based platform compared to an x86-based platform. Perform such comparison when accelerating LDPC decoder and other modules using a Graphics Processing Unit (GPU) will be reported in the next deliverable.
- Chapter 3 describes RU power consumption reduction methods which use RU on/off schemes, Power Amplifier (PA) blanking schemes, Peak to Average Power Ratio (PAPR) reduction methods, Digital Pre-Distortion (DPD) and Envelope Tracking (ET).
- Chapter 4 presents the developments for enhancing coverage in radio cells using ISAC, RIS, and relays. The waveforms and algorithms for ISAC are presented and assisted beam training and network coverage optimization are discussed. In addition, the system model, optimization methods and channel estimation for Hybrid Reconfigurable Intelligent Surfaces (HRISs). Finally, relay power consumption models for different spectrum allocation strategies are elaborated.
- Chapter 5 is summary and conclusions.

2 DU and CU Acceleration Initial Developments and Results

As part of the overall effort to reduce power consumption, it is suggested to use new architectures and acceleration methods for implementing the DU L1 processing. A description of SotA acceleration methods can be found in BeGREEN D3.1 [1].

One known approach is using a dedicated HW accelerator such as *Intel® vRAN ACC100* or *NXP Layerscape® Access LA12xx*[1]. These accelerators are usually power efficient, but they are limited to run some specific algorithms, thus being not flexible. A different approach is implementing these algorithms over x86 based Commercial Off-The-Shelf (COTS) servers, which offer flexibility at the price of higher power consumption. Finally, an approach for implementing the whole L1 processing chain on a GPU is suggested by NVIDIA, as part of their Aerial Software Development Kit (SDK)[2], where the parallel computing capabilities of the GPU can be utilized for the complex algorithmic operations. However, GPUs are not well designed for general purpose operation. Thus, implementing the whole L1 over a GPU is inefficient and cumbersome.

The CU HW acceleration in Arm servers represents a transformative approach in telecommunications infrastructure for energy efficiency management. By leveraging the inherent efficiencies and capabilities of Arm architectures, hardware acceleration modules are specifically designed to offload complex computational tasks from the main processor, to significantly enhance the processing speed and energy efficiency of the CU. This is crucial for the CU's role in handling the higher layers of the protocol stack, including the SDAP, PDCP, and RRC layers, where rapid data processing and decision-making are paramount. Hardware acceleration enables the CU to manage these tasks more efficiently, ensuring seamless data flow and robust network management. Furthermore, the use of Arm servers with integrated hardware acceleration facilitates a more scalable and flexible network deployment, allowing telecom operators to dynamically adjust to varying network demands while minimizing power consumption and operational costs. This strategic adoption not only underscores the technological advancements in network equipment design but also aligns with the industry's push towards more sustainable and efficient network infrastructures, heralding a new era of high-performance, energy-conscious telecommunications networks.

2.1 DU acceleration

BeGREEN offers a different approach than those described in the previous paragraph. This is, using Arm architecture to perform most of the DU processing, where only the most computation complex algorithms will be offloaded to a GPU. This offers the “best of all worlds”. In general, using Arm has potential in power saving due to its various power optimized architecture aspects. On top of that, with the suggested approach the GPU can be utilized for performing the most complex algorithms and does not need to perform the mainline L1 operations that are not well suited for its architecture and will be done in the Arm cores. Flexibility is obtained where both the Arm and the GPU cores are programmable. Finally, as these are not dedicated HW accelerators, for periods when the network load is low (for example at night) these cores can be used for other purposes, such as deep neural network training.

Two algorithms with significant level of computation complexity were identified in [1]. One is the LDPC decoder, and the other is the sphere decoder (SD). To recap, the complexity of the LDPC decoder stems from the need to perform the “message passing” algorithms for a significant number of bit nodes, check nodes and several iterations. The complexity of SD stems from the need to calculate a significant number of Euclidian distances and calculating the Log Likelihood Ratios (LLRs) based on these distances.

[1] https://networkbuilders.intel.com/solutionslibrary/intel-vran-dedicated-accelerator-acc100-product-brief
[2] https://developer.nvidia.com/aerial

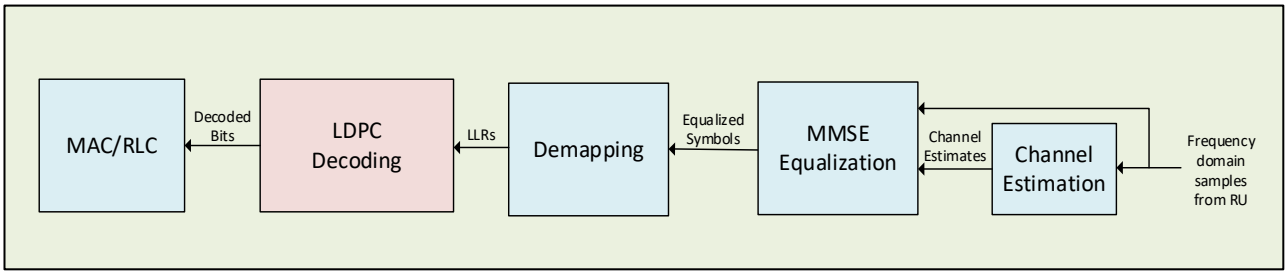


Figure 2-1 DU receiver architecture

The target is to reach at least eight layers (i.e. spatial multiplexing of 8 concurrent streams) in the uplink (UL), which is the standard industry number of layers for massive Multiple-Input Multiple-Output (mMIMO) deployments as described in the O-RAN IoT profiles[3]. This amounts to a very large number of code blocks, which means substantially increasing the complexity of the LDPC decoder resulting in increased power consumption. Therefore, potential power reduction by implementing them over the power efficient Arm architectures (rather than on x86 architectures) and accelerating them using a GPU are being examined. The LDPC decoder, suggested architectures the implementation methodology and some results on performance and power consumption are presented in this chapter.

2.1.1 DU architecture

The Arm DU receiver architecture is shown in Figure 2-1. This receiver flow is usually implemented on HW or an x86 CPUs, and in BeGREEN plan is to implement it on an Arm CPU as well as a combination of Arm and GPU. It is described in BeGREEN D3.1 Section 2.2.1 [1]. To recap, the received samples for the DU are used for estimating the fading channel. Then, these samples and the estimated channel are used for equalizing the samples. The equalized samples are demapped into LLRs per bit. The LDPC decoder uses the LLRs as inputs and performs an iterative process where, in each iteration, more and more erroneous LLRs are flipped, i.e., corrected. Successful decoding is achieved when all erroneous bits have been corrected before the maximal number of iterations is reached.

In BeGREEN the focus is on the cases where LDPC is being run on a dedicated HW accelerator. It is assumed that the DU is fully SW based, to have the flexibility and scalability needed for supporting different kinds of deployments and network loads. For example, a CPU being used for DU algorithms during the day, when the network is loaded, can be utilized for other tasks like training a Machine Learning (ML) model for the BeGREEN "Intelligence Plane" (or any other ML model) during the night when the network is underutilized.

2.1.2 Comparison of DU algorithm implementation over x86 and Arm

In BeGREEN the DU algorithms will be implemented on both the x86 platform, the Arm platform, and then on a GPU. For a code to run efficiently, with reduced power consumption, it needs to be optimised to the specific target CPU. Hence, dedicated compiler intrinsic functions are to be used for each CPU that would properly utilize its architecture. Without them, the code could be functional, but would be highly inefficient. Specifically, for the DU algorithms being developed in BeGREEN, the Google Highway[4] is used. Google Highway is an open-source Single Instruction Multiple Data (SIMD) abstraction library. It enables the SIMD code to be written using a generic code that is abstracted from the underlying platform. At the compilation time, generic SIMD instructions are replaced by highly optimised, functionally equivalent, CPU specific intrinsic calls of the maximum supported vector size. Google Highway has been developed since 2019, it is well maintained and primarily used in image and video processing applications. It supports a large range of

[3] VZW-2021.07.02-WG4t-C-IOT-profile-v30

[4] https://google.github.io/highway/en/master/quick_reference.html

CPU architectures including x86, Arm and RISC-V. The major advantage of using the Google Highway abstraction libraries is that a single version of the source code can be coded and maintained, then compiled for all the required platforms, for instance, for x86 and Arm.

BeGREEN's optimizations on the LDPC decoder are: 1) performing partial decoding of only on a subset of columns and rows corresponding to actually received LLRs, 2) using use Cyclic Redundancy Check (CRC) to decide on early termination check instead of the parity check matrix, and 3) optimizing the order in which the processing is done.

The functionality of the LDPC decoder in the x86 and the Arm processors should be bit-exact (meaning that, for the same input, the output of the module will be the same regardless of which processor its running on) for the code or nearly bit-exact, as the same codebase is used for both. Therefore, the decoder performance should be the same for both architectures. As the SIMD vector length is shorted in the Arm processor, the number of cycles it would take to run the LDPC and, hence, the running time, will be higher on the Arm platform. However, not all codes are arithmetic operations that can be parallelized, so the difference in running times will not necessarily be large and will depend on the codeblock size. For a fair comparison, the throughput-per-Watt for each platform is measured.

For example, *Intel ICL-D-1747NTE 10-core* (20 threads) [5] can be compared to *Arm N2 24-core* (24 threads) [6], which are of a similar performance grade, as they are both devices intended for the same Edge or embedded segment and have capacity for comparable number of threads. There are two key differences between these CPUs. The first is that the x86 (SIMD) vector extension has 512 bits per port and Arm has 128 bits per port, each having 2 ports per core. The second is that x86 uses multithreading and Arm does not. Some more information on building energy efficient 5G infrastructure using Arm Neoverse systems is described in the Appendix on page 66.

2.2 DU LDPC Decoder throughput and energy consumption results

A comparison of the LDPC decoder in several scenarios is discussed in this section. First, in this section, the throughput of the LDPC implementation is compared to the required performance defined by 3GPP for these scenarios [3]. Then, the power consumption results between x86 and Arm will be compared.

2.2.1 LDPC throughput analysis

Specifically, the testcases shown in Table 2-1 [3] will be examined and plots for the defined testcases will be presented in Figure 2-2, Figure 2-3, Figure 2-4 and Figure 2-5. These testcases are defined for 100 MHz bandwidth, which means 273 Resource Blocks (RBs) equivalent to 3276 subcarriers.

Table 2-1 3GPP Standard Requirements for PUSCH

	Propagation conditions[7]	Modulation	LDPC Code Rate	Target SNR to reach 70% of the maximum throughput (dB)
1	TDLB100-400	QPSK	193/1024	-2.2
2	TDLC300-100	16QAM	658/1024	10.8
3	TDLA30-10	64QAM	567/1024	13.6
4	TDLA30-10	256QAM	682.5/1024	21.7

[5] https://www.intel.com/content/www/us/en/products/sku/226116/intel-xeon-d1747nte-processor-15m-cache-up-to-3-50-ghz/specifications.html

[6] https://developer.arm.com/Processors/Neoverse%20N2%20Compute%20Subsystem

[7] Propagation conditions are defined in Annex G of [3]

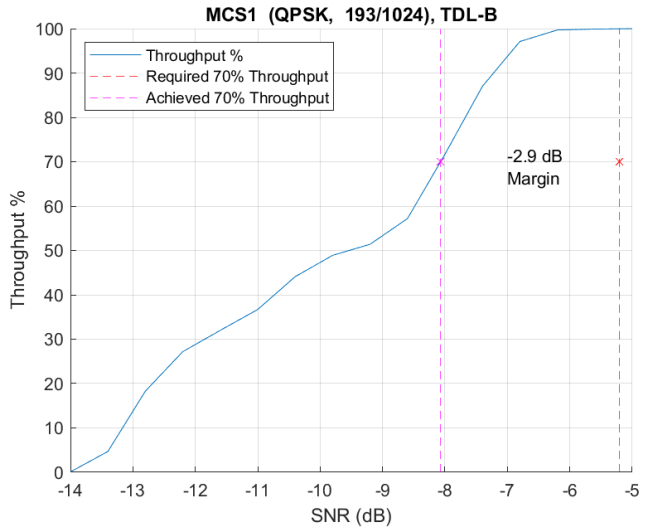


Figure 2-2 PUSCH test 1 performance results

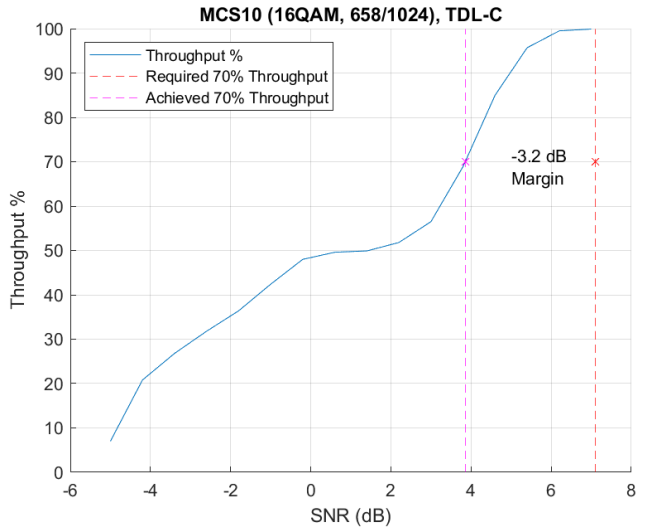


Figure 2-3 PUSCH test 2 performance results

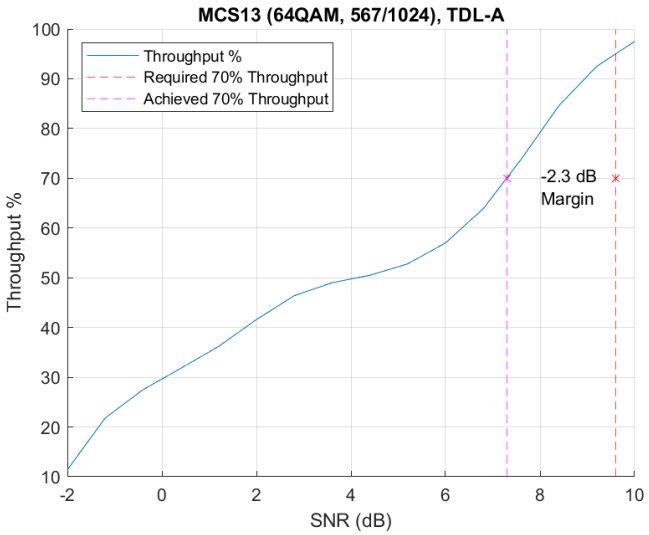


Figure 2-4 PUSCH test 3 performance results

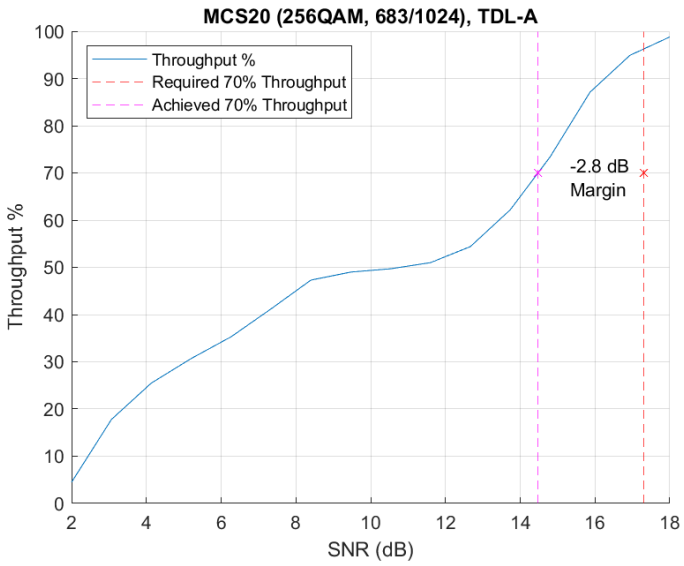


Figure 2-5 PUSCH test 4 performance results

2.2.2 LDPC energy consumption analysis

For the scenarios defined in Table 2-1, the power consumption of the two processors are compared using the setup described in Figure 2-6. This work is defined in BeGREEN D5.1 [5] as the "PoC4 - use case 1 – testing of DU algorithms in simulation mode". The x86 based platform being used is *ADLINK MECS-6120*[8], the Arm based platform being used is the *Marvel Octeon 10 CN10624*[9] and a metered power delivery unit is used to concurrently measure the overall power consumption of each platform. The specific model being is *enlogic EN6812*[10].

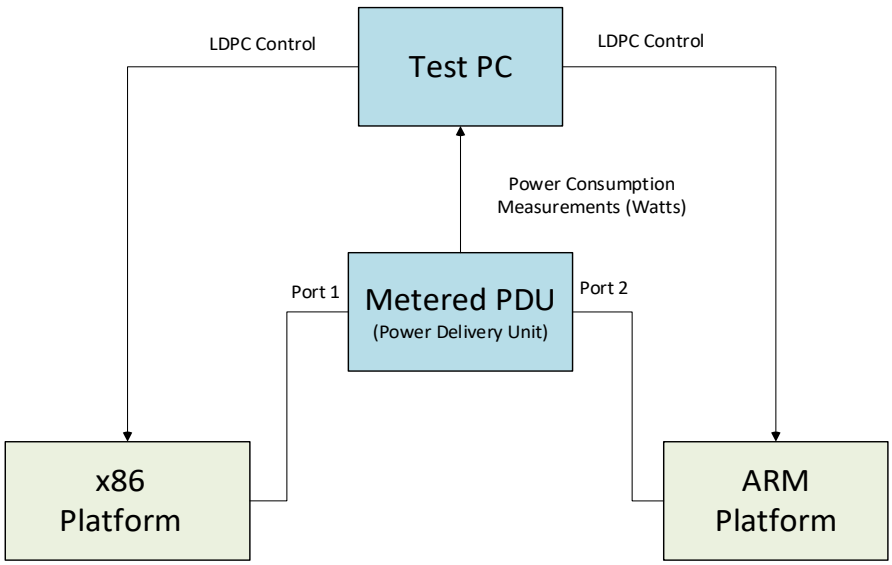


Figure 2-6 DU Power Comparison Setup

[8] https://www.adlinktech.com/Products/Edge_Server/Edge_Server/MECS-6120

[9] https://www.marvell.com/content/dam/marvell/en/public-collateral/embedded-processors/marvell-octeon-10-dpu-platform-product-brief.pdf

[10] https://www.enlogic.com/public/assets/ZipExeFiles/EnlogicAdvantage_Firmware_Release_Notes_v1.0_final.pdf

Table 2-2 LDPC Processing Arm to x86 Consumed Energy Ratio

#	Modulation	LDPC Code Rate	#RBs	x86 Energy kbits/Joule	Arm Neon Energy kbits/Joule	Arm SVE2 Energy kbits/Joule	Arm Neon to x86 Energy Ratio (%)	Arm SVE2 to x86 Energy Ratio (%)
1	QPSK	193/1024	273	2812	2461	2990	114	94
2	16QAM	658/1024	273	8075	7406	8303	109	97
3	64QAM	567/1024	273	6992	6187	7038	113	99
4	256QAM	682.5/1024	273	8854	8142	9016	109	98
5	64QAM	567/1024	10	5054	5796	6693	87	76
6	256QAM	682.5/1024	3	7410	7475	9982	99	74
7	QPSK	193/1024	10	6593	7567	9775	87	67
8	QPSK	193/1024	1	228	391	1035	58	22

The energy consumption results for each of the testcases from Table 2-1 are shown in Table 2-2. In addition, to get a wider perspective on the power consumption tests, with smaller allocations were done, which are common for actual cellular deployments where multiple UEs share the same bandwidth.

The Arm is configured to use two different SIMD instruction sets. The first is Neon, and the second is the more advanced SVE2, which is a superset of Scalable Vector Extension (SVE) and Neon[11]. Results show that SVE2 power consumption per bit per Hz is always lower than x86, while for Neon there are scenarios in which the x86 consumption is lower. Results are shown in Table 2-2.

The reason that in the larger allocations (larger number of RBs) the x86 power consumption is better than the Arm's is because of the large codeblock sizes. For large codeblock sizes the SIMD vector utilization is higher, and due to the fact that Arm has a 128-bit vector length and x86 has a 512-bit vector length, the Arm consumes less power, but is does less processing than x86 for the same period. For Arm implementations with a larger vector length, much better ratios could be observed.

2.3 CU architecture for acceleration

The CU in the landscape of telecommunications infrastructure plays a pivotal role, especially within the framework of O-RAN architectures. At its core, the CU is divided into two critical components as shown in Figure 2-7: the CU-Control Plane (CU-CP) and one or more CU-User Planes (CU-UPs), as specified by 3GPP in [6]. The CU-CP is tasked with hosting the Radio Resource Control (RRC) and the control plane portion of the Packet Data Convergence Protocol (PDCP), which are essential for managing and directing the flow of data across the network. On the other hand, the CU-UP handles the user plane part of the PDCP and the Service Data Adaptation Protocol (SDAP), facilitating the efficient transmission and processing of user data. These CU-UPs are connected to the CU-CP through the E1 interface, a point-to-point connection designed for the seamless management of bearer contexts within the CU-UP by the CU-CP, as outlined by 3GPP in [7].

Given the significantly higher volumes of UP traffic compared to CP traffic, the CU-UP handles with the most of the network's load. This distinction between CU-CP and CU-UP is deliberate, allowing for the distribution of workload across multiple CU-UPs to optimize overall network efficiency. This design philosophy emphasizes the strategic decision to prioritise the porting of the CU from x86 to Arm architectures, particularly for the CU-UP, due to its intensive CPU and power consumption. Additionally, porting the CU-CP to Arm offers the convenience of running the entire CU on Arm, further enhancing the system's efficiency and compatibility.

[11] https://developer.arm.com/documentation/102340/0100/Introducing-SVE2

In Figure 2-7, the containerized implementation of Accelleran CU-UP and CU-CP is presented, showing that each instance of the CU-CP and the multiple possible instances of the CU-UP are containerized inside pods deployed by heml instances. The CU components establish connectivity via Container Network Interface (CNI0), which is a Kubernetes interface to make Kubernetes components talk to each other.

2.3.1 CU porting to Arm

The transition of the CU to Arm architecture proves the adaptability and forward-thinking approach in network equipment design. Given that the CU is predominantly developed in C++, the shift from x86 to Arm is facilitated by the inherent portability of the C++ code. This process involves the introduction of a cross-compilation toolchain in the build environment, replacing the native x86 compiler with one that is compatible with Arm. Adjustments were made to the build scripts to accommodate this new compilation process, including the integration of a *sysroot* for Arm.

A *sysroot* is a directory that mimics the target operating system's file structure and contains all necessary headers, libraries, and dependencies. This ensures the cross-compiler has access to the required files, facilitating accurate and efficient cross-compilation. Moreover, x86-specific optimizations in the code are disabled for testing purposes (in particular NIA1 ciphering), leveraging the fallback implementation in pure C++ to maintain functionality. The culmination of this effort is reflected in the distribution of the CU as Docker images, which are now built to support both x86 and Arm architectures. This not only demonstrates the technical feasibility of porting to Arm but also highlights the strategic importance of such a move in the context of next-generation networks.

The integration of x86 and Arm architectures within Docker[12] and Kubernetes[13] ecosystems facilitates the deployment of containers and services across diverse computing environments, leveraging the strengths of both architectures for optimized performance and resource utilization.

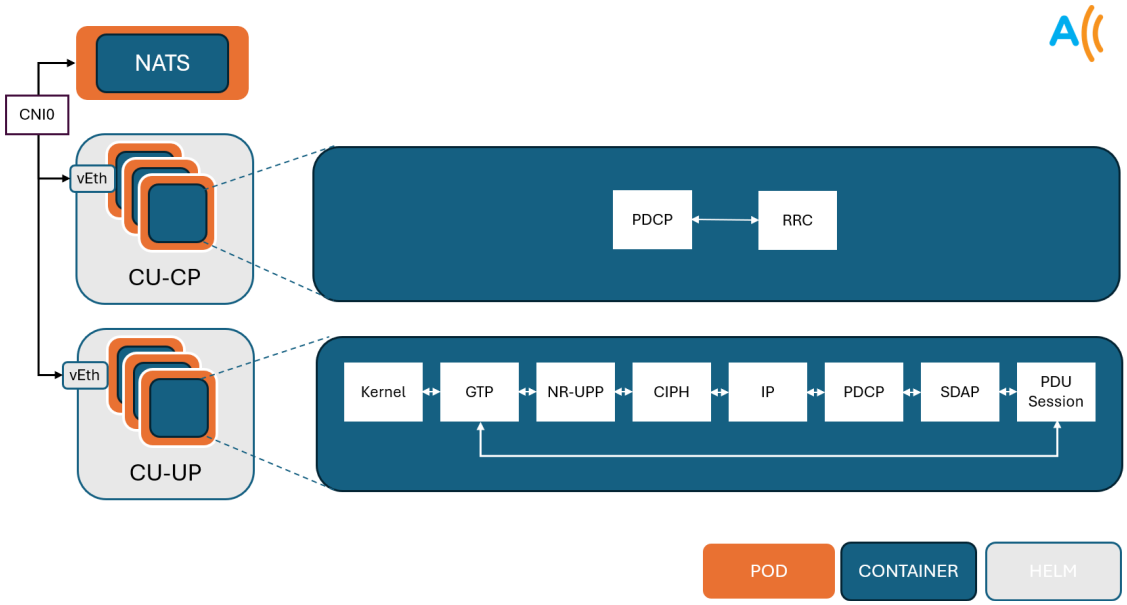


Figure 2-7 Internal CU architecture for HW acceleration

[12] Docker is a platform that automates the deployment of applications in lightweight, portable containers that run consistently across different environments. https://www.docker.com/

[13] Kubernetes, also known as K8s, is an open-source system for automating deployment, scaling, and management of containerized applications. https://kubernetes.io/

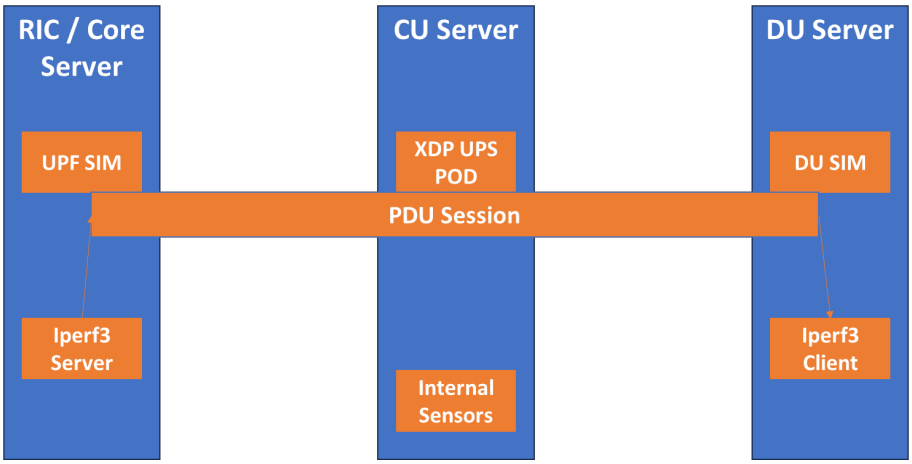
Docker's support for multi-architecture images, achieved through tools like Buildx[14] and manifest lists, enables seamless application packaging and distribution for both x86 and Arm platforms. Concurrently, Kubernetes enhances orchestration by scheduling workloads on compatible nodes, utilizing labels and selectors to match pods with the appropriate architecture, thereby ensuring efficient resource allocation and application scalability. This approach not only underscores the technical agility in managing heterogeneous computing environments but also aligns with strategic objectives for high-density, energy-efficient, and performant computing solutions in both cloud and edge contexts.

Through the redesign and adaptation of the CU for Arm architecture, a more energy-efficient, flexible, and scalable network infrastructure is provided. It exemplifies how BeGREEN can provide innovative technologies that enhance the capacity and reliability of networks, fulfilling the ever-growing demand for faster and more robust communication systems supported by O-RAN architectures.

2.3.2 Initial results for CU acceleration

The laboratory setup for CU HW acceleration is shown in Figure 2-8. An implementation architecture is presented. From bottom up, the first x86 server contains the RAN emulator where RU, DU and UEs are evaluated. Then the x86 server of interest hosting the CU is implemented by its own. On top of it, there is the rest of the O-RAN implementation, meaning the Radio Intelligent Controller (RIC), and the 5G core (5GC), which includes the UPF simulation integration for the tests. Above those, the Arm server hosting the arm instance of the CU. And finally, at the top, is the optical 10 Gbps router to interconnect the down below servers. To compare Arm vs x86 server in terms of energy consumption, 3 different scenarios were tested:

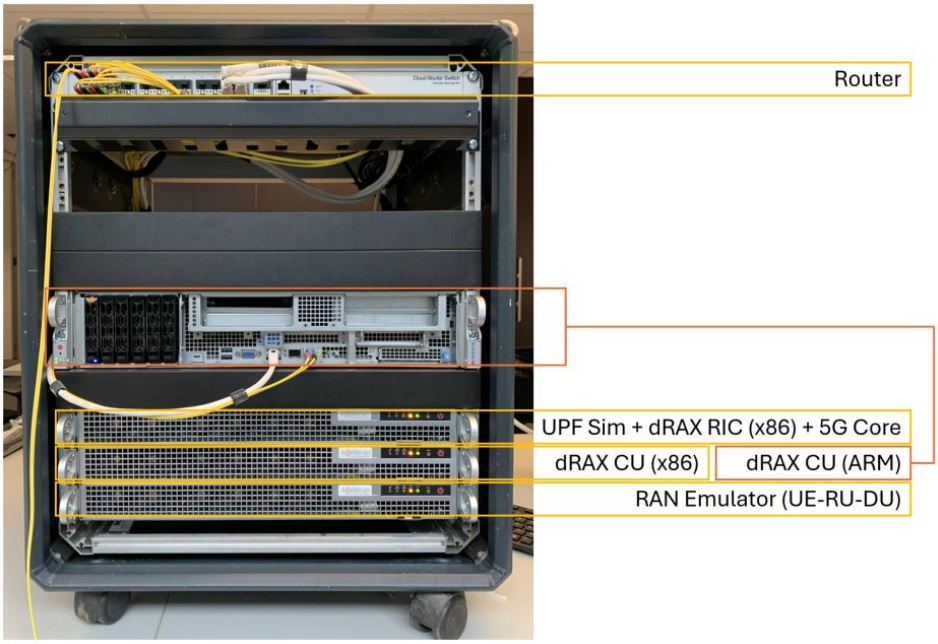
- Scenario 1: UE - max DL UDP throughput: In this scenario a single UE is set to transmit maximum throughput with UDP traffic to evaluate the CU-UP performance with the maximum CPU clock speed fixed.
- Scenario 2: 6 UEs Stable Traffic: In this scenario each UE is allocated to one user plane stack (UPS) instance inside the CU-UP for data traffic handling. Each of UE does 10 Mbps DL UDP throughput. Also, this is tested with maximum and flexible CPU frequency on Arm for the CU-UP. UPS here is the Kubernetes pod (microservice), which handles the user plane GTP traffic within PDU sessions
- Scenario 3: Idle on-air[15] cell consumption: In this scenario the power consumption of a Cell active but without traffic is considered.



a)

[14] https://github.com/docker/buildx

[15] Cell is active, broadcasting signals but no UE to receive them.



b)

Figure 2-8 CU Acceleration lab setup a) block diagram, b) rack implementation.

The configuration for both servers (x86 and Arm) is described as follows. The x86 AMD server is configured as a typical product line server, running Ubuntu 22.04 with no special settings and using a generic kernel (5.15.0-100-generic #110-Ubuntu SMP Wed Feb 7 13:27:48 UTC 2024 x86_64 x86_64 x86_64 GNU/Linux). It is equipped with 128GB of RAM and hyper-threading enabled, providing 16 cores and 32 threads. On the other hand, the ARM server has an 80-core configuration reduced to 32 cores and 32 threads, with the remaining cores disabled. It has a 512GB RAM configuration reduced to 256GB, as reducing further was not possible. Huge pages are configured with the following settings: default_hugepagesz=1G, hugepagesz=1G, hugepages=64, tsc=r, affinity=0, rcu_nocbs=1-69, rcu_nocb_poll, nosoftlockup, cma=256m, and kpti=off. For some tests, the CPU frequency is fixed to the maximum.

The power measurement method involved reading the overall power consumption from internal sensors inside the tested CU servers every 5 seconds over a 10-minute period. After collecting these readings, the average power consumption for the measured period was calculated. However, there was a limitation related to the NIC capabilities and stability. The number of XDP UPS pods was constrained by the NIC's performance and the system's stability, which was compromised when using more than 6 XDP pods.

Empty Servers consumption in Watt are shown in Table 2-3**.** 1UE test results x86 vs Arm are presented in Figure 2-9 and Figure 2-10. Results show that the power consumption of x86 is always better compared to Arm. In addition, the energy efficiency expressed in bit/joules is also depicted, showing that the x86 is more energy efficient for a single UE case scenario.

Table 2-3 Empty Server (only OS) Power Consumption

	RIC/CORE Server	x86 CU Server	Arm CU Server	DU Server
Consumption [W] (max/min/avg)	54/53/53	64/53/54	110/112/114	60/54/55

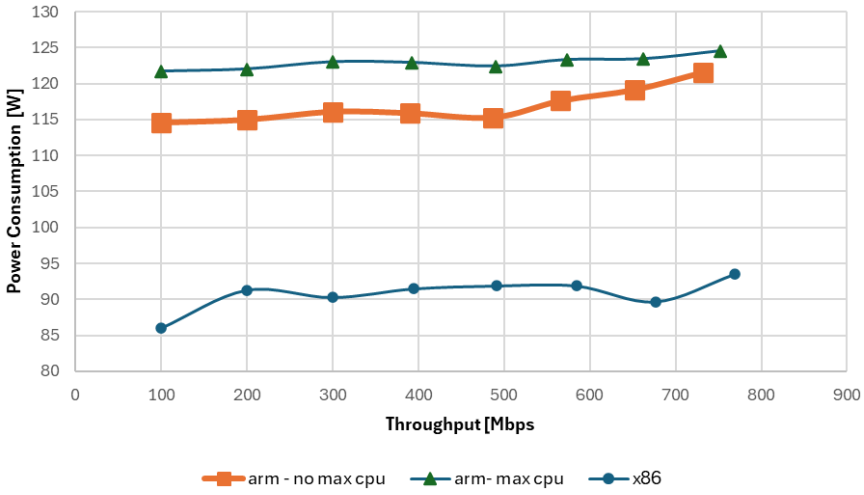


Figure 2-9 Power consumption results for 1 UE

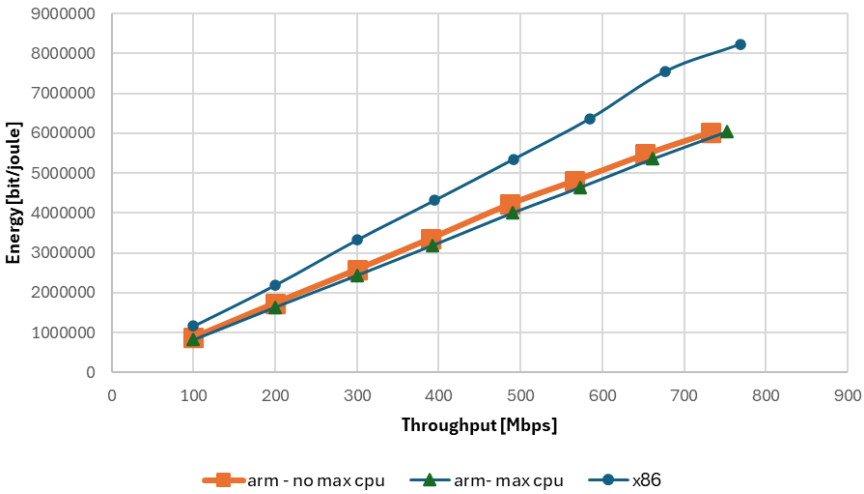


Figure 2-10 Energy efficiency results for 1 UE

Test results for 6 UEs comparing x86 vs Arm are presented in Figure 2-11 and Figure 2-12, respectively. The power consumption of Arm is quite stable and with traffic slightly increasing, but x86 with traffic increasing dramatically. Unfortunately, due to NIC card limitation no more than 6 UEs could be configured.

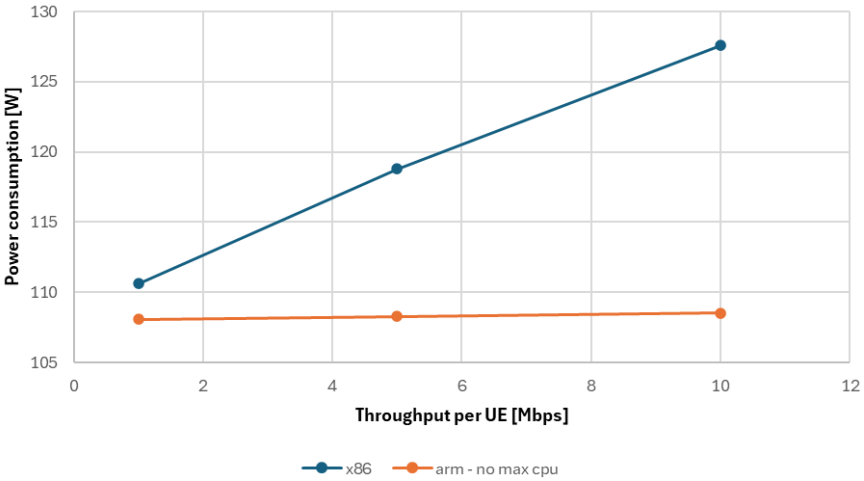


Figure 2-11 Power consumption results for 6 UEs

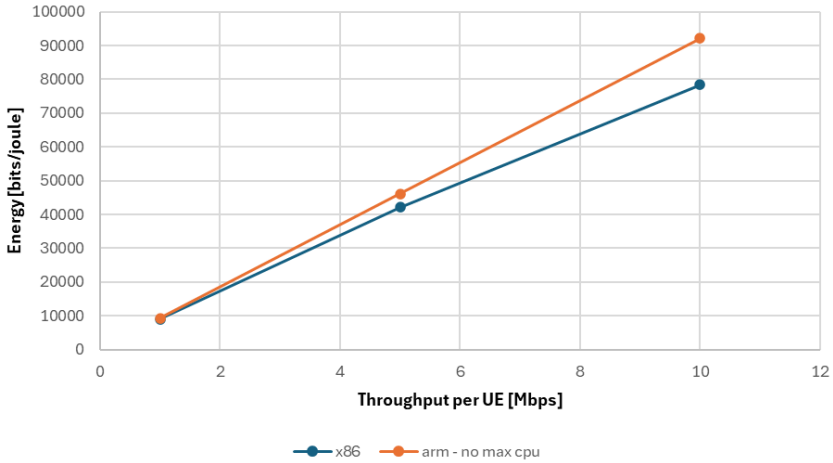


Figure 2-12 Energy efficiency results for 6 UEs

From the performed tests, it can be concluded that for small networks with a limited number of simultaneous users, x86 is the best choice in terms of energy savings, achieving about 22% better energy efficiency compared to Arm, as it shown in Figure 2-13. Conversely, for traditional MNOs with large networks and many simultaneous users, Arm performs better than x86 in terms of energy savings, achieving approximately 10% more energy efficiency compared to x86 in this scenario, as can be seen in Figure 2-14.

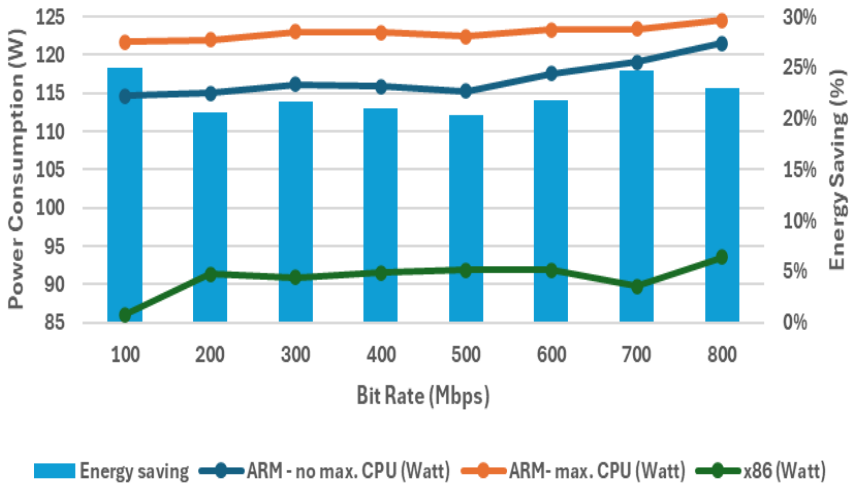


Figure 2-13 Energy saving in case of 1 UE scenario

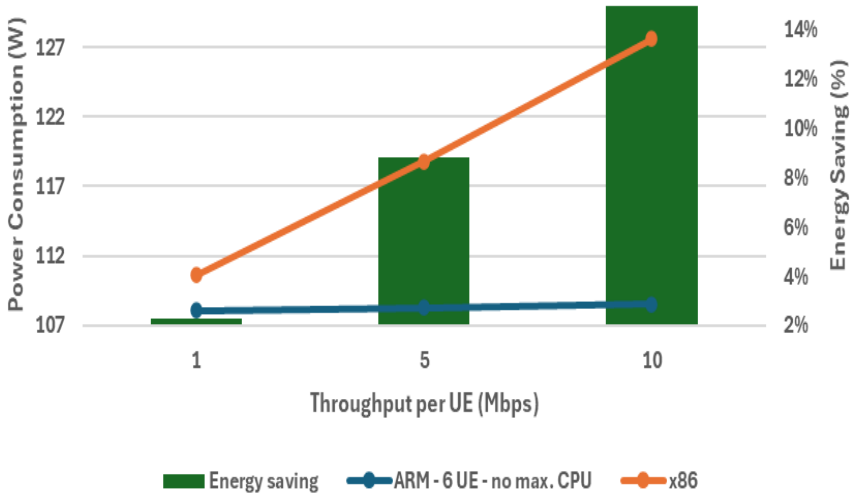


Figure 2-14 Energy saving in case of 6 UEs scenario

2.4 Collaborative Insights

Looking at the DU power consumption results presented in 2.2 and the CU power consumption results shown in Section 2.3, it can be observed that there are cases where x86 consumes approximately the same power as the Arm, there are cases where the x86 consumes less power and there cases where the Arm consumes less power. For the CU algorithms, it can be observed that as the throughput gets larger, the power saving when using Arm compared to Intel increases. For the DU algorithm the results are the other way around. For the smaller code blocks and allocation sizes the best power saving is achieved when using Arm. The reason for that is that the LDPC decoder algorithm being examined in the DU processing uses more parallelization as codeblock size gets bigger, and that the for the CPUs being compare the Arm has a SIMD register bit width which is 4 times smaller. In real deployments UEs codeblock size will vary, and it is expected to have UEs with small codeblock sizes, where bigger gain in using Arm compared to x86 will be achieved. In any case, the aggregated power reduction when using Arm covers a variety of cases.

3 RU Energy Consumption Optimization Approaches

The RU in a cellular network, particularly in 5G and B5G, is a major power consumer. Estimates typically suggest that RUs can account for 74-78 % of the total power consumption of a mobile network base station (BS) [8]. This large range depends on various factors such as the specific technologies in use, network configuration, and the environment in which the BS operates.

In 5G networks, the complexity and power demands increase due to technologies like mMIMO and due to the use of higher frequency bands (such as mmWave). The RUs not only process and transmit signals to and from user devices but also handle real-time tasks such as beamforming, which are energy intensive.

While other components of the network architecture, such as the Baseband Unit (BBU) and core network elements, also consume power, they typically do so at lower levels compared to the RUs. Therefore, network operators and technology developers continuously focus on optimizing the energy efficiency of RUs. This includes advancements in Power Amplifier (PA) technology, more efficient cooling solutions, and intelligent software algorithms that optimize signal processing and transmission to reduce power usage.

In the following, two different approaches investigated that can substantially reduce the RU energy consumption:

- Applications that optimize the use of the RUs in the Network by turning them on and off in low traffic situations (Section 3.1).
- New RU internal modules such as PA Blanking, AI based Digital Pre-Distortion (DPD) and Envelope Tracking (ET), which are able to reduce the RU energy consumption (section 3.2).

3.1 Near-RT and non-RT applications for RU management (xApps/rApps)

In the current O-RAN architecture, control of RUs is limited and considered insufficient to implement the objectives of energy-efficient cellular networks. The present Open Fronthaul (OFH) management plane (OFH M-Plane) allows the Service and Management Orchestrator (SMO) to control RUs, but the RU is not directly manageable by the Near-Real-Time RIC (Near-RT RIC).

Near-RT applications (xApps) are applications running on the RIC designed to operate with minimal delay. These applications play a critical role in the O-RAN architecture, enhancing the capabilities of the RAN with swift decision-making capabilities. Non-RT applications (rApps), in contrast, do operate with longer processing times and are not constrained by immediate response requirements. They are instrumental in strategic, long-term planning and optimisation of the RAN.

In BeGREEN, the implementation of the RIC will include different deployments and flavours for the Near-RT RIC and the Non-RT RIC. This will allow the BeGREEN team to assess the performance of the A1 interfaces and the different distributed roles of the Intelligence Plane framework. In particular, the implementation of the xApp in BeGREEN will integrate accelerated CU metrics into the Near-RT RIC and will create a fast interface with the Intelligence Plane for easy integration with the Intelligent SDK (see Figure 3-1). Here, an Intelligent xApp will control the RU utilisation, power transmission and cell status to reduce its energy consumption [9].

The plan is to develop an xApp for on/off the RU, while in in the framework of T4.1 [10] an xApp and an rApp for RU on/off will be implemented by ACC and I2CAT. Since the RU on/off is based on TeraVM emulator[16], the implementation will be different from the actual (used in real network) as the way O-RAN sees the RU controlling mechanism (shown in Figure 3-2) is not implementable on a real commercial RU.

[16] Viavi Solutions TeraVM, https://www.viavisolutions.com/en-us/products/teravmt

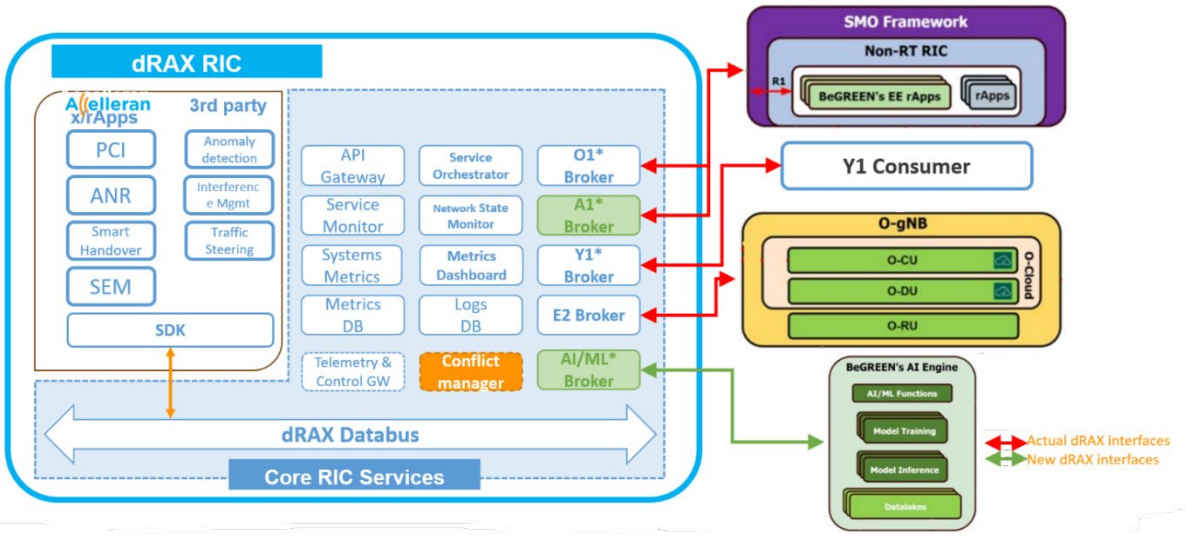


Figure 3-1 RU on/off implementation architecture

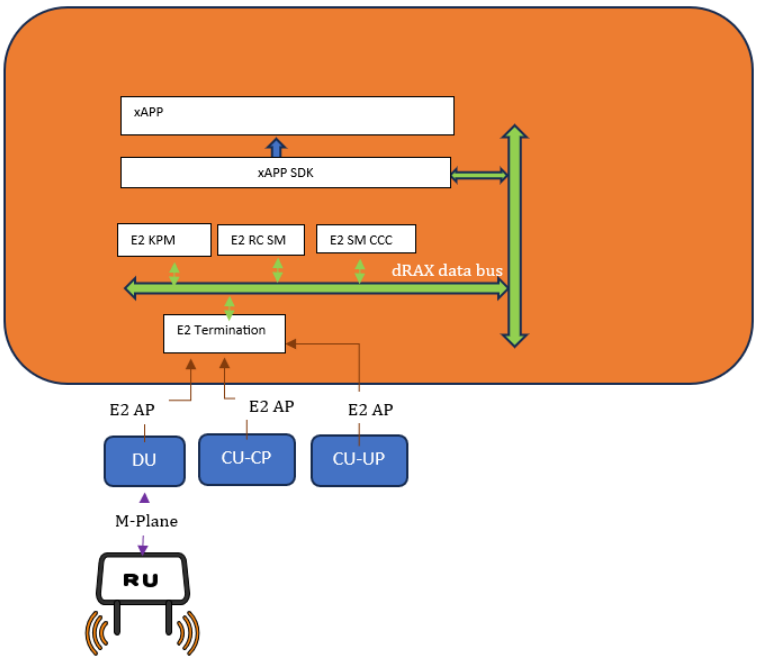


Figure 3-2 O-RAN vision on RU control from RIC/xApp

Near-RT RIC provides E2 termination entity that supports encoding and decoding of service model specific messages. These messages will be transferred to the E2 Node using O-RAN E2 AP procedures. E2 SM CCC service model supports two configurable states 'toBeEnergySaving' and 'toBeNotEnergySaving' states per cell. xApps in the RIC can control the energy saving state of the cell over E2 interface connection to the DU. The xApps monitor the load on the cells continuously and move the cells to Energy Saving mode based on ML algorithm. When the cell is moved to 'toBeEnergySaving' state, the DU initiates shutdown of the cell and switches off the RU using M-Plane interface over NETCONF. When the load on the cell is likely to increase based on ML prediction algorithm, the xApp switches on the cell in the DU by moving it to 'toBeNotEnergySaving' state. The DU turns on the cell over M-Plane interface using NETCONF when this happens.

3.2 RU power consumption reduction

Several RU power consumption reduction methods are described in this section. The RU PA Blanking method that was implemented and demonstrated in a lab test (Section 3.2.1) and the AI based DPD and ET methods

that were verified by extensive simulations (Section 3.2.2.4).

3.2.1 RU PA Blanking Module

The "Power Amplifier Blanking" method is described. PA blanking stops the PA energy consumption for symbols without active data allocations. This is done by disconnecting the DC bias to the RU PAs when there is no data allocated in the Orthogonal Frequency-Division Multiple Access (OFDMA) symbols that needed to be transmitted in the downlink to the UEs in the RU coverage area. When resources of a downlink symbol have no allocated data, the related RF PA will be blanked (momentarily shut down) for the symbol duration. Figure 3-3 a) depicts this method. On the left side of the figure, the signal without blanking is presented and on the right side an Orthogonal Frequency-Division Multiplexing (OFDM) signal with blanked symbols is displayed.

This method can be further improved by maximizing the data empty OFDMA symbols. This can be achieved by configuring the scheduler in the DU to allocate the data to be transmitted in the DL in the frequency domain rather than the time domain.

RU energy saving by symbol blanking was demonstrated at EuCNC 2024, whose testbed configuration is shown in Figure 3-3.

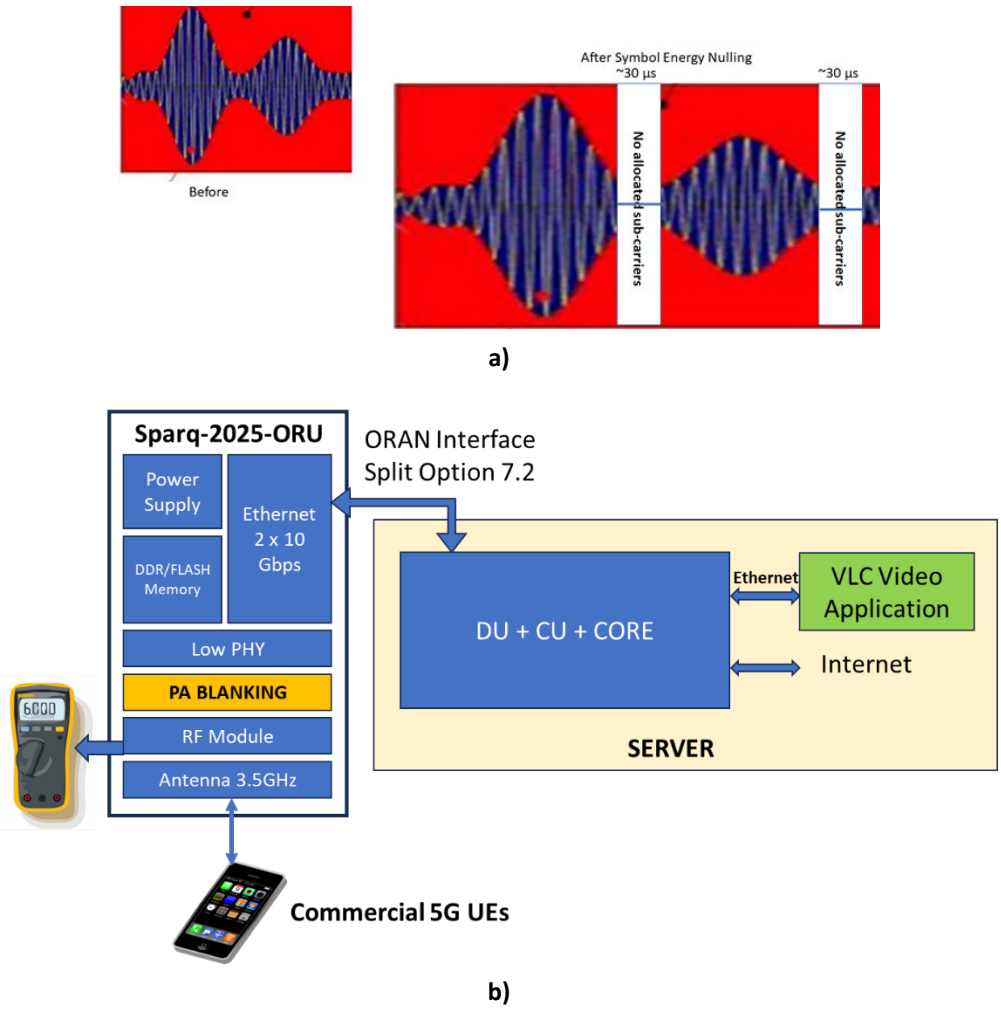


Figure 3-3 a) PA Blanking Method, b) RU energy saving by symbol blanking demo for EuCNC 2024

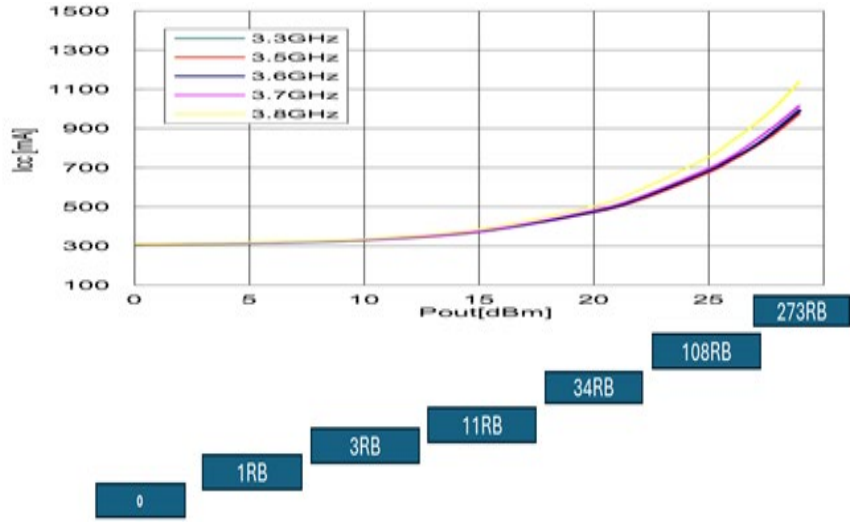


Figure 3-4 PA power consumption related to the traffic load

To perform the blanking, a PA blanking module (depicted in yellow in Figure 3-3) is integrated in the RU. This data will be streamed by the DU scheduler and will be transferred to the RU over the O-RAN 7.2 interface [17]. Figure 3-4 shows the PA power consumption curve related to the traffic load in RBs that was taken from a typical 4.5-5 Volt RF PA at 3.5 GHz.

From the above graph the relation between the traffic and the PA current feed is observed. To decrease the energy consumption, the DU scheduler can concentrate the allocated RBs over several OFDMA symbols while leaving the rest of the slot symbols unloaded and by carrying out power blanking reduce the energy consumed over a slot by 5 times as indicated is the following examples.

Example-1: 100 RB allocation

The example compares a case with 100 RBs allocations over a 10 symbols slot (TX part), the traffic load can be allocated over 10 symbols with 10 RBs per symbol or over a single symbol with 100RBs per this single symbol.

For the case with PA blanking, for 10RBs per symbol the consumed power will be 10 x 0.35 x V = 3.5 x V (where 10 is the number of RBs, 0.35 is the PA current for 10 RBs taken from Figure 3-4 and V is the PA voltage) and for 100 RBs allocated in one symbol and rest blanked the power is 1 x 0.7 x V= 0.7 x V (where 1 is the number of RBs, 0.7 is the PA current for 100 RBs taken from Figure 3-4 and V is the PA voltage). This ratio between 3.5 x V and 0.7 x V represents the 5 times power reduction achieved by the PA Blanking algorithm when the 100 RBs are allocated in a single symbol instead of 10 symbols.

Without PA blanking, for 10 RBs per symbol the consumed power will be 10 x 0.35 V = 3.5 V and for 100 RBs allocated in one symbol the power is 1 x 0.7 x V + 9 x 0.3 x V= 3.4 x V (where 0.3 is the PA current for the 9 symbols with 0 RBs taken from Figure 3-4). The ratio between 3.5 x V and 3.4 x V represents a very limited 0.3% power saving when the 100 RBs are allocated in a single symbol instead of 10 symbols when the PA Blanking module is not activated.

Example-2: 273 RB allocation

In this example an OFDMA slot with 273 RBs is used (the maximum number of RBs defined in the 5G 3GPP standard for Sub 6 GHz). This load can be allocated over a single symbol with 273 RBs or over 10 symbols with 27 RB each. Using the same calculation process as the one used in Example-1 above, it is concluded that the consumed energy when using power blanking is also about 5 times lower when the data is allocated in one symbol compared to a uniform load distribution allocated over 10 symbols. Without blanking this ratio is much lower (only 1.35).

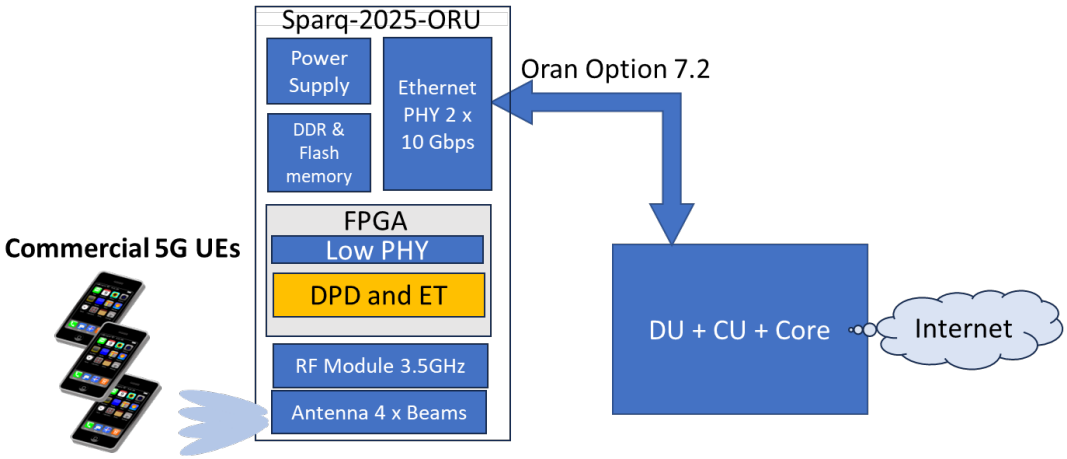


Figure 3-5 AI based Digital Pre-Distortion and Envelop Tracking System

There is some (less than 5%) loss in this mechanism efficiency due to early turn off and on to match the PA response time (less than 0.5 microseconds).

The above examples indicate the importance of the power blanking mechanism in the RU RF frontend.

3.2.2 AI based RU enhancements for power efficiency optimisation

Previously, in BeGREEN D3.1- Section 4.2 [1], two methods for RU PA power consumption reduction were described: 1) The AI-based DPD, and 2) the AI-based ET method. The block diagram of the location of these two modules on the RU is depicted in **Error! Reference source not found.**.

The following section starts with a review of the techniques and methods researched by BeGREEN to optimize the RU operation and continues with additional simulation results, a power saving example, simulation results and intermediate conclusions. The current research recommendations to further enhancements of the B5G RAN operation concurrent with energy savings are reviewed at the end of the section.

3.2.2.1 PAPR reduction methods

OFDM is broadly used in digital communication systems due to its advantages including high bit rate, immunity to multipath and high spectral efficiency but it suffers a high PAPR at the transmitted signal as it is generated from independent orthogonal subcarriers. The OFDM signal is generated by a set of frequency domain sub-carriers (OFDM symbol), which are assigned and transformed to time domain using Inverse Fast Fourier Transform (IFFT). This results in a time domain signal, which is a summation of orthogonal frequency components. This signal fluctuates in time and exhibits high peak values at random locations.

The complex OFDM signal is modulated into RF signal using Quadrature Amplitude Modulation (QAM), which is followed by an RF PA. The RF PA, like any electronic device, is not perfect. Its transfer function has non linearities, memory (hysteresis) effects and saturation (clipping) at max signal value. The PA nonlinearities result in destroying subcarriers' orthogonality and generates out of band emissions resulting from inter-modulation products [1].

BeGREEN target is to introduce methods to improve OFDM waveform PAPR effects with less energy waste. Enhanced techniques which were studied preliminary in BeGREEN D3.1 [1], are discussed more in depth in the following sections.

3.2.2.2 Digital Pre-Distortion (DPD)

RF PAs have non-linearities that induce inter-modulation products and out of band emission. To improve their performance, PAs can be used to increase the back-off in order to get less clipping and keep operating

at the PA linear region. However, this method increases dramatically the power consumption of the RU, therefore a more effective method to resolve the PA non-linearities is required. DPD is a common technique to cope with signal PAPR it is mainly used to increase the RF PA efficiency (Section 4.2.3.2 in [1]).

3.2.2.3 Power Envelope Tracking

RF PA ET is a method in which the power supply voltage applied to the RF PA is continuously adjusted ensuring the PA is operating at peak power efficiency (see initial details in Section 4.2.3.3 in [1]). For ET, PAPR becomes less of an issue, being the main problem to tackle the average power. A 4K (FFT size) OFDM may induce a 12dB PAPR. Figure 4-14 in [1] presented the PA signal and its envelope. In the following paragraphs the result of an extended study regarding the impact of the envelope amplitude, A_{sat}, on the RF signal quality ia described.

For A_{sat}=1 the envelope follows exactly the signal amplitude. The impact on the SNR and the spectrum out of band emissions are studied. Figure 3-6 shows for reference the signal constellation and spectrum without ET. Figure 3-7 depicts the spectrum out of band emissions for A_{sat}=1.2.

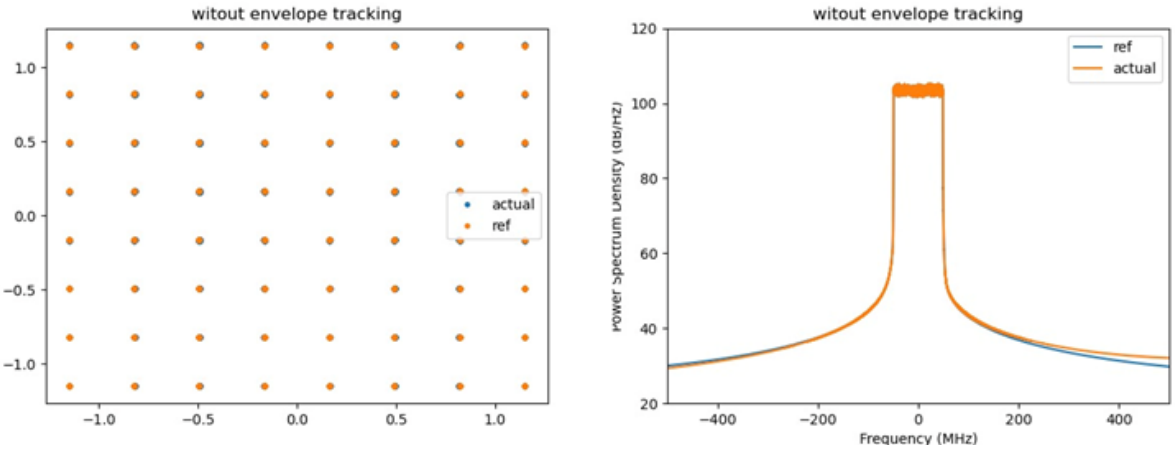


Figure 3-6 Spectrum and Constellation without Envelope Tracking

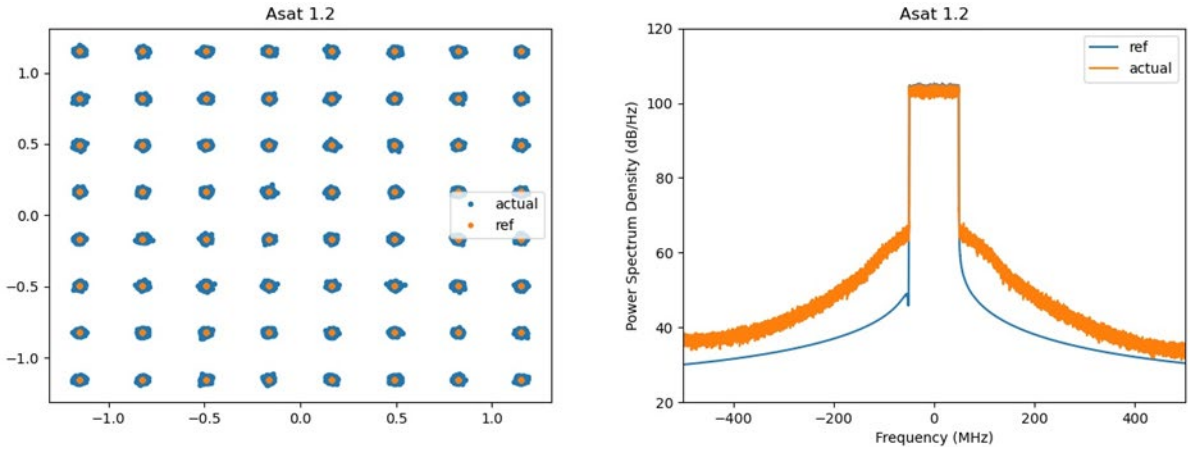


Figure 3-7 Spectrum and Constellation with A_{sat} =1.2 ET

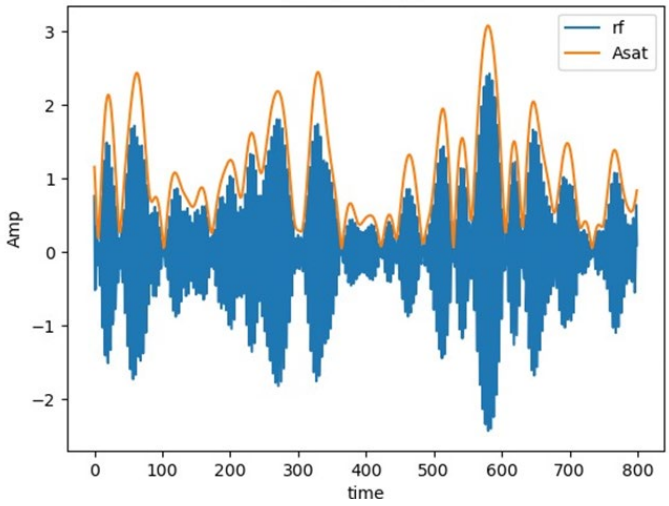


Figure 3-8 Signal envelope at A_{sat} =1.2

It can be observed that the limited out of band emissions and the clean constellations (64 QAM) indicating a high SNR. The waveform (blue) and the signal envelope (orange) are depicted as well in Figure 3-8. It can be observed that there is a power loss due to A_{sat}.

To gain a better efficiency $A_{sat} < 1$ is considered. Care should be taken to maintain high enough SNR and low enough emission to enable proper operation. Thus, the same analysis is made for A_{sat}=1 and 0.9, whose results are shown Figure 3-9 and Figure 3-10. It can be observed that the noised constellation and increase in out of band emissions. The signal envelope for A_{sat} = 0.9 is shown in Figure 3-11.

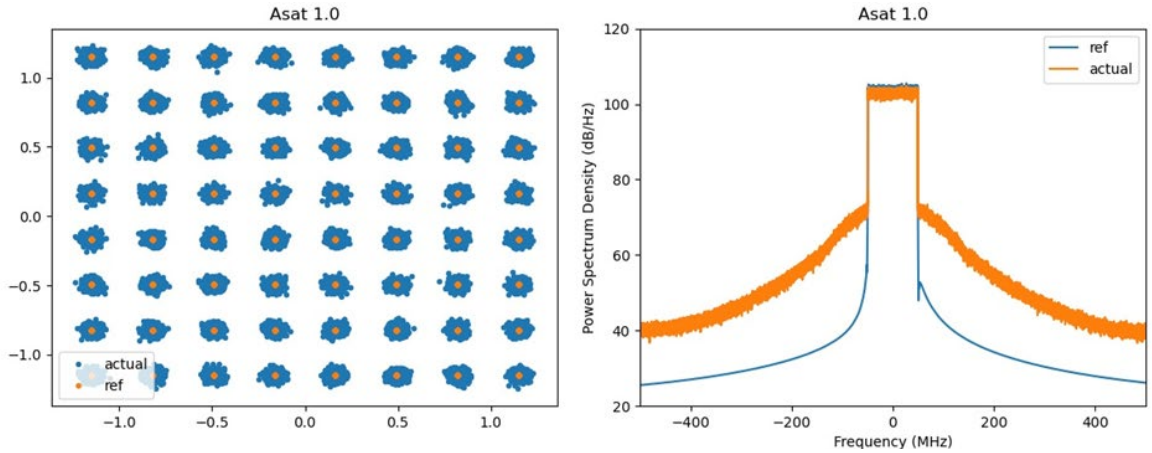


Figure 3-9 Spectrum and constellation with A_{sat} =1.0 Envelope Tracking

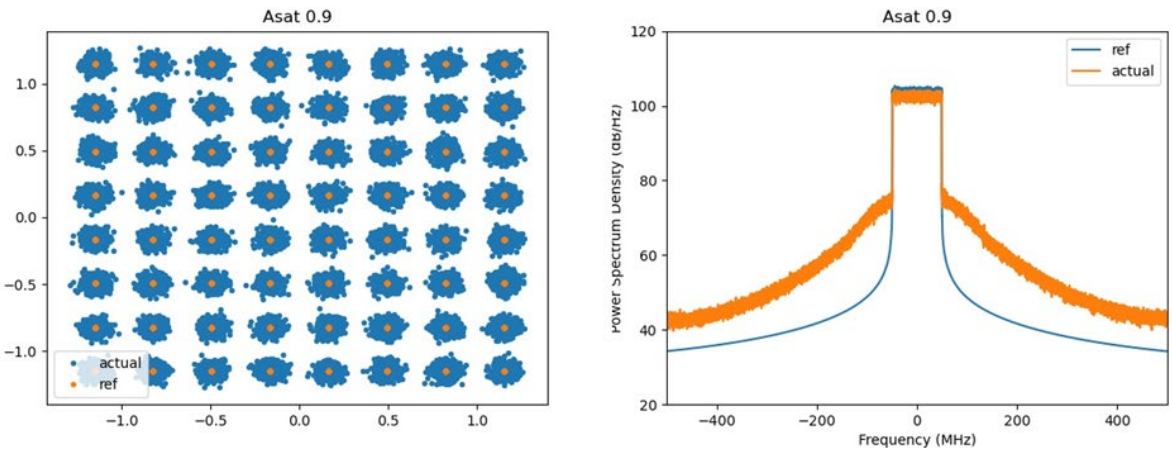


Figure 3-10 Spectrum and constellation with A_{sat} = 0.9 with Envelope Tracking

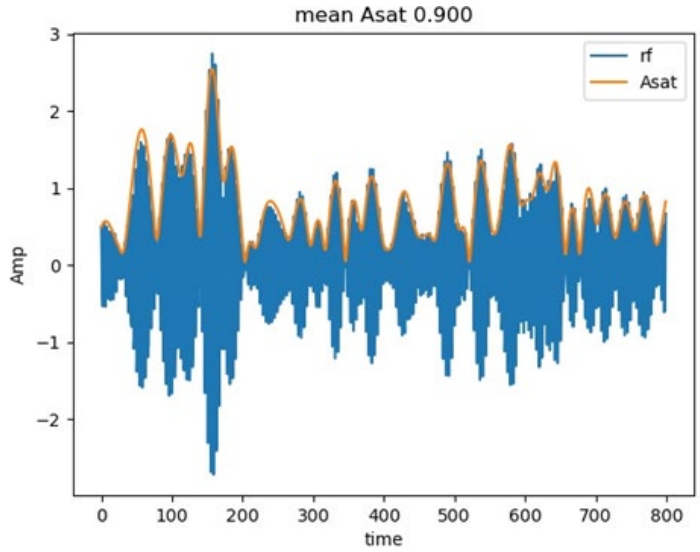


Figure 3-11 Signal envelope at Asat = 0.9 with ET

Clipping can be observed at signal peaks. As can be seen the envelope has effect on the SNR which degrades, but the values remain high enough, 27 dB, to support high order modulation (e.g., 64QAM).

To conclude, ET improves the power efficiency. However, care should be taken to maintain the performance quality. The A_{sat} is a parameter controlling the enveloping process can be optimized according to the required SNR and power efficiency. The following assumptions and design considerations have been considered for obtaining the results:

- The power feed is now a parameter.
- PA parameters may change with envelope.
- The signal envelope changes slower than the samples.
- PA model distortion is even more severe because the signal is always close to the back off power.

3.2.2.4 Calculation of Power Saving examples due to DPD and ET

A power saving example can be presented as follows:

- Assuming a 1-Watt RF Power Output (30 dBm)
- Required PAPR: 10 dB.
- PA efficiency: 30%
- Power Consumption for a standard PA will be: 1 x 10 / 0.3= 33 Watts
- Power Consumption in PA with ET and DPD, assuming the DPD and ET circuits consume 10 Watts: 1/0.3 +10 = 13.3 Watts
- Power Saving 19.7 Watts = 60 % energy savings.

It should be clear that for higher rate PAs (e.g., 10 or 50 Watts), the energy savings will be much higher.

3.2.2.5 Summary and conclusions

From the above it can be concluded that, to cope with OFDM signals' PAPR, DPD is much more effective when augmented with ET. Figure 3-12 depicts this augmented method. The contributions and research results presented in this section have been studied in the context of BeGREEN by means of simulations, which were mainly related to the RU PA energy consumption and its efficiency optimization methods. The issues which were addressed in this research include: 1) PAPR enhancement, 2) AI-based DPD, and 3) AI-based ET.

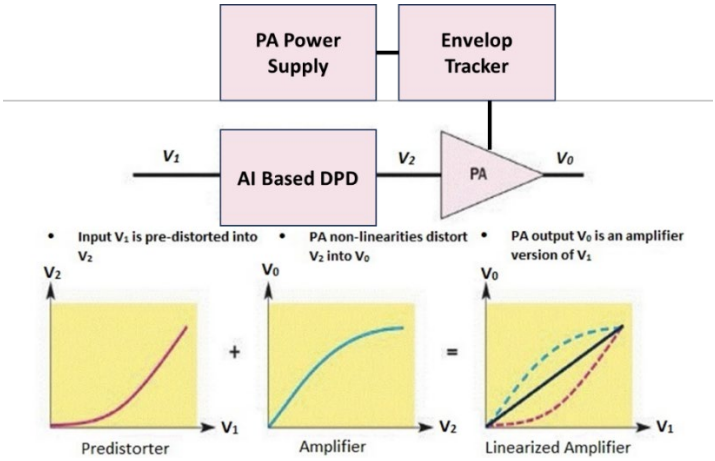


Figure 3-12 DPD augmented with ET

The results indicate that the above techniques can significantly reduce the RU Power consumption and enhance its energy efficiency.

To effectively solve 5G waveform PAPR problems required is a trade-off which depends on various parameters and requirements including:

- In band noise (e.g., QPSK can suffer more in-band noise from QAM265).
- Out of band emission.
- Transmit power.
- Transmitter complexity.
- Processing delay.

The main conclusion is that for enhancing OFDM based waveforms energy efficiency a DPD technique is more effective when combined with ET.

3.2.2.6 Future work

AI enforced techniques to accomplish these techniques were part of this research as presented in the above paragraphs. Examples and plots out of this research and simulations effort were also included.

Further efforts to enhance and improve 5G RAN operation and energy efficiency that RunEL is currently researching in the BeGREEN Project and will be reported in following deliveries are depicted below:

- Capacity/Energy optimization in a cellular cell. To find best work points to balance between energy consumption and capacity.
- Cell configuration. To find how to optimize wireless system capacity as a function of power, spectrum reuse, beamforming and more.
- Possibilities for a power race between transmitters. To explore how to avoid RF power race between neighbouring cells while providing optimum performance and energy efficiency.
- Resources allocation Power/Bandwidth trade-offs to maximize performance.

4 PHY Solutions for Power Efficiency Enhancement

BeGREEN has identified a few different physical (PHY) layer technologies that have a huge potential for improving the overall network energy efficiency of future 6G RANs.

ISAC is a technology that will be present in wireless devices in short term, being its main contribution towards optimal network resource allocation as well as improvement of PHY layer functionalities such as beam training. The additional sensing data obtained using ISAC will be used for additional optimization of different functions to improve the network energy efficiency.

RISs are configurable reflective surfaces used to reflect the incoming radio signal in preferred directions. These devices are identified by BeGREEN as a crucial technology for energy efficiency optimization in the future RANs. Namely, as the carrier frequencies go higher, in order to get larger channel bandwidths, the propagation is becoming more challenging, especially in urban areas. Two main factors contribute to this, the higher free-space path loss as well as worse penetration through walls and buildings. To mitigate these issues, deployment of additional RUs is necessary, thus worsening the energy balance of the network. To avoid this, RISs are seen as a possible solution to enable smaller energy footprint in the network. This is possible as the RISs are almost passive devices needing just a small amount of energy to function.

ISAC and RIS are relatively new concepts that will play a significant role in future 6G networks. The success of RIS applications in various communication scenarios has inspired research to explore the combination of RIS and ISAC technologies [11][12][13].

Deploying relays to enhance a B5G network involves, among other tasks, deciding which radio resources the relay will use. Based on spectrum utilisation, relays can operate in in-band mode when using the same spectrum resources as the cellular system, or in out-band mode, when using different spectrum resources. While the operation in out-band is usually considered as a way to avoid mutual interferences between BSs and relays, 5G/B5G networks offer new elements like beamforming and the management of Physical Resource Blocks (PRBs) that allow to play with the spatial dimension, and to use the available spectrum more efficiently.

An interference analysis in relay-enhanced B5G networks has been considered as activity in BeGREEN WP3, for that three different strategies that differ in the way how the spectrum is shared by the base station (BS) and the relays have been defined, then the required transmitted power at the BS and the relay, to support a given bit rate of the UEs, and the total power consumption, are formulated for each one.

4.1 ISAC assisted physical layer improvements

ISAC will enable a radar like, network wide functionality [14]. This functionality will not disturb the underlying data transmission functionality of the network. In an ideal case, the waveforms used for data transmission should be additionally used for the sensing functionality. Nevertheless, since these waveforms are not initially intended for sensing, they will not be a perfect candidate. Therefore, evaluation and improvement of these waveforms is one of the tasks that is addressed in this document.

The main objective of the sensing function of the network is to detect different objects and persons as well as their features. These features include, but are not limited to: the size of the objects, their speed and direction, type (i.e. human/object), etc. In BeGREEN, different algorithms will be developed to implement these sensing functionalities. These algorithms will be used to create different heat maps representing the detected obstacles, the moving objects or persons, etc.

Receiving the reflections of the transmitted signal from different obstacles would create a complex signal that needs to be analysed further to provide a meaningful information about the surrounding environment. A few different parameters of the surrounding environment and the objects can be estimated. Initially, the

distance between these objects and the transceiver, as well as their position is of primary interest. Furthermore, the movement vector, i.e. direction and velocity can be also estimated. Using different algorithms (e.g. based on ML and AI), detection of the type of obstacle, the size, and the behaviour, etc. is also possible.

Finally, the sensing data will be used to optimize the network energy efficiency based on the surrounding environment, propagation characteristics, user density, etc.

4.1.1 Typical sensing scenario

A typical sensing scenario is shown in Figure 4-1. A transmitter transmits a waveform that will be reflected back by different obstacles. The reflections arriving back at the receiver will be delayed for the ToF, i.e. the time needed to travel to the obstacle and back and will arrive at the receiver at different angles. Therefore, it is necessary, to be able to estimate both the ToA at the receiver, i.e. the ToF as well as the direction of arrival (DoA). With this approach, one can have both the distance and the direction at which the obstacles reside.

Having the distance to the object, as well as the azimuth angle, one can estimate the position of the obstacles in 2-dimensional (2D) space. If the elevation angle is also known, one can estimate the position of the obstacle in 3-dimensional (3D) space. The 2D polar coordinate system, used for estimation of the position of the obstacle is given in Figure 4-2.

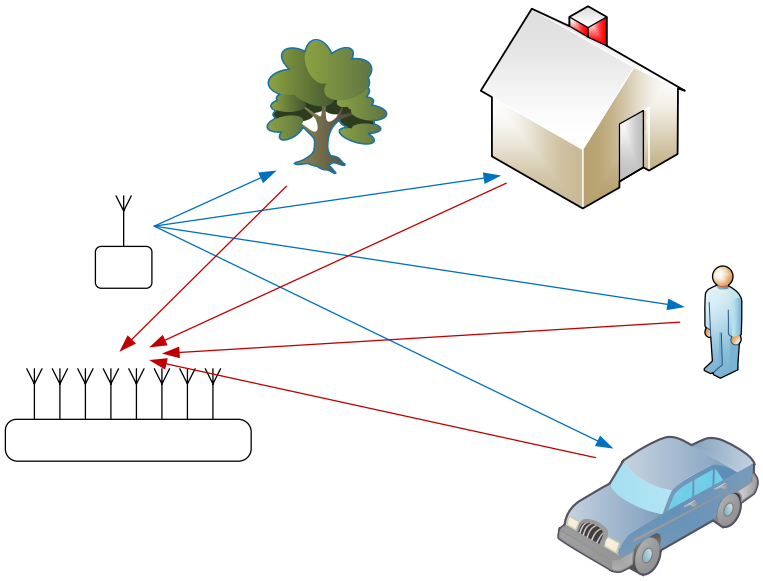


Figure 4-1 Typical sensing scenario

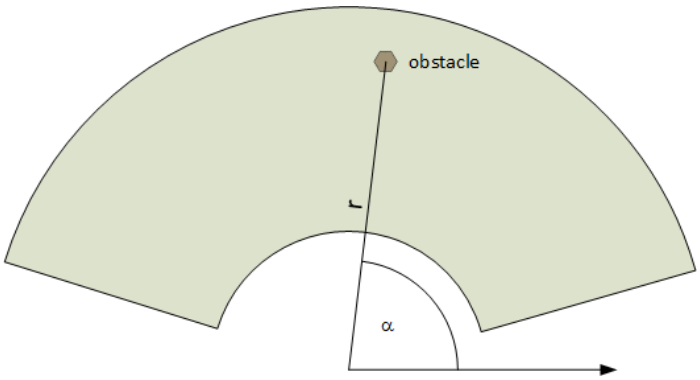


Figure 4-2 2D polar coordinate system

4.1.2 Waveforms for ISAC

Wireless communication and radar sensing systems have been developed separately for decades. However, with the increasing demands in next-generation communication systems, it becomes infeasible to simultaneously get high data rates and the desired sensing accuracy with separate design of communication and sensing systems. Also, these two systems share some aspects in terms of the system architecture, and they can use the same spectrum. This motivates researchers to study the joint design of these systems, which is an emerging topic known as ISAC [15].

ISAC is a technology that should enable sensing of the environment using the radio data communication network. Different approaches can be found in the literature, but the preferred approach is to use the data communication waveforms for sensing purposes. Using this approach, no additional overhead (spectrum usage) will be introduced, which means that the capacity of the data transmission channel will not be affected. The main disadvantage of this approach is that the data transmission waveforms are not necessarily optimal for sensing. This is the main motivation why some authors are adding additional waveforms to the data transmission waveforms to perform sensing.

Sensing is initially perceived as a radar like functionality of the radio network. The radar functionality can be easily implemented in a data communication system since most of the hardware components needed are already present. A typical radar sensing is performed by sending a known radio waveform, which is reflected by the surrounding objects and received back by the receiver. By estimating the Time of Flight (ToF) of the radio waveform, the distance between the radio transceiver and the object reflecting the transmitted waveform can be estimated. The angle between the radio transceiver and the object can be estimated by using an antenna array. By having a linear antenna array, only an angle in a single plane can be estimated, and having a planar antenna the angles in both azimuth and elevation plane can be estimated. This means that the 2D or 3D position of the obstacles/persons can be estimated, respectively.

An ideal waveform for performing sensing is a Dirac pulse. This pulse is sent from a transmitter and multiple pulses, which are reflected from the surrounding obstacles, will be received at the receiver. This is an ideal model, since it involves assumptions that are not practically feasible in a real system, such as the fact that these pulses have an infinite bandwidth. Even if relatively short pulses can be transmitted in some systems, e.g. Ultra-Wideband (UWB), the amplitude of these pulses will be limited, meaning they will carry limited amount of energy, strongly limiting the available SNR at the receiver. This is why waveforms with longer duration are the usual choice. Nevertheless, the ToF is usually much shorter than the waveform itself, which means that the received waveforms will be overlapped. To be able to resolve all these reflections, one must be able to discriminate between the different merged reflections. This is possible with a technique known as pulse compression [16].

Pulse compression is a signal processing method used in radar systems to increase the range resolution or SNR when the pulse width or bandwidth of the transmitted signal is constrained, respectively, i.e. it can be used to compress a longer waveform into a single pulse. In pulse compression, first, the transmitted signal is modulated and then the received signal is correlated with the modulated version of the transmitted signal. This is shown in Figure 4-3 and it is performed by correlating the received waveform with its own locally generated copy. This approach is only possible with some waveforms that can support pulse compression.

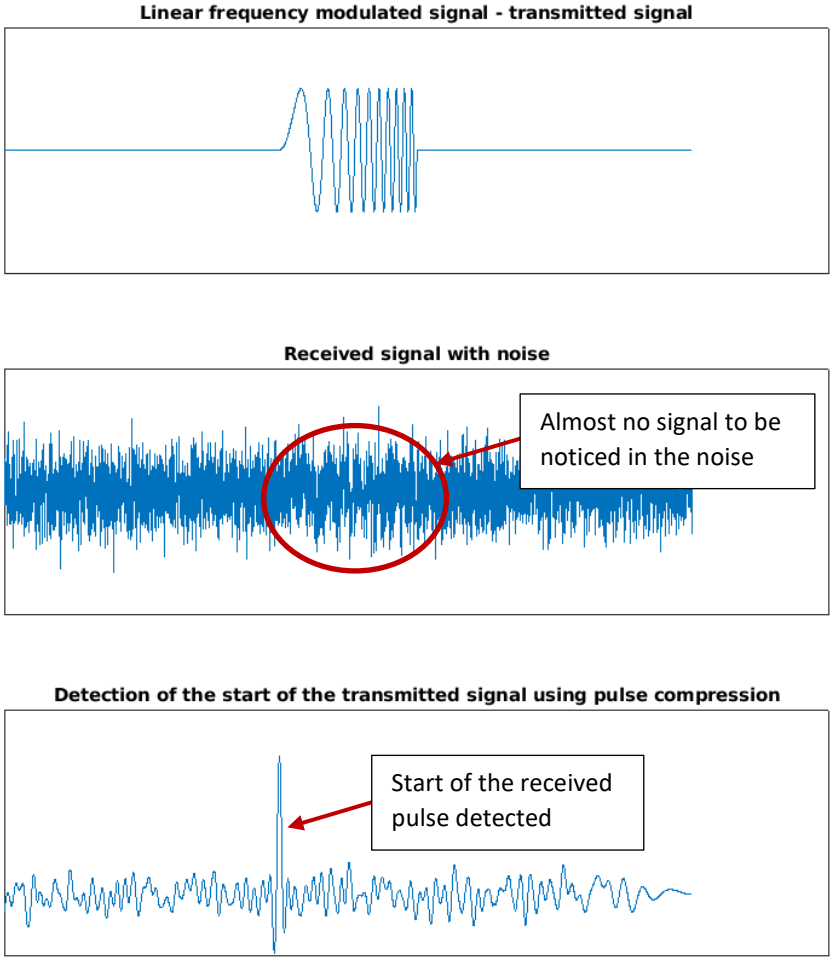


Figure 4-3 Pulse compression waveform in presence of noise and its detection

The main feature of the waveforms supporting pulse compression is their specific autocorrelation function. Namely, in an ideal case, the autocorrelation function should have a peak for $\tau = 0$, and to be 0 for $\tau \neq 0$. Finding a waveform with this autocorrelation function is usually not a simple task. Therefore, waveforms having an autocorrelation function that approximately complies to the above-mentioned rules are preferred. In real implementations, the waveforms used for sensing do have a strong and narrow peak for $\tau = 0$, and sidelobes that are not 0 but significantly lower than the main peak. In an ideal case, the main peak width is equal to $2/B$, where B is the bandwidth of the used waveform. It is also important to have a sensing waveform that has as narrow as possible peak for $\tau=0$ since, in that case, two reflections arriving from two nearby objects can be easily resolved. This also defines the range resolution, as:

$$R_r = \frac{c}{2 * B}$$

where c is the speed of light, B is the transmitted signal bandwidth and R_r is the range resolution.

In the case of **radar**, usually analogue waveforms are used and the main intention is to use waveforms that are easy to process in the analogue domain and to be able to simplify the digital part of the radar. Therefore, one of the most common radar waveforms used are continuous wave (CW) and frequency modulated continuous wave (FMCW) radars. In the **CW** case, just a carrier, i.e. a sine wave is transmitted. This type of waveform is used mainly for a Doppler radar, to estimate the velocity of a given target (object). In the case of **FMCW radar**, a linear modulated frequency waveform (LMFW) is used. This waveform is initially pre-processed in analogue domain, which enables later use of low end, low sample rate data converters for signal acquisition. The used analogue pre-processing further enables relatively simple processing in digital domain and enables detection of targets in 2D or 3D space, depending on the antenna array configuration.

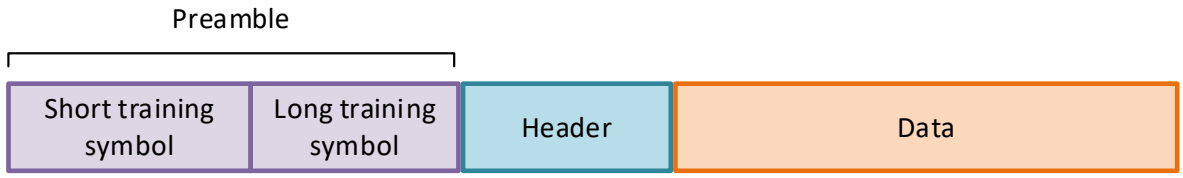


Figure 4-4 Typical wireless physical layer frame

In a wireless data transmission system, compared to relatively simple radar systems, the available digital HW is already quite powerful. This means that no analogue pre-processing is needed, since all the needed signal processing can be performed in digital domain. This is also preferred approach since no significant changes need to be introduced in the HW of the wireless data transmission system. Additionally, more complex waveforms, like OFDM, or single carrier data digital waveforms can be used.

As already mentioned, the waveforms used for data transmission in the wireless system should be used also for sensing. These waveforms should also have the above-mentioned autocorrelation properties. A typical frame structure for the most popular wireless transmission systems is shown in Figure 4-4. Two distinct fields that can be used for sensing are the preamble and the data field. Both of them usually have good autocorrelation function. The header field is not preferred for the purpose of sensing.

4.1.3 ISAC system model and methodology

It is assumed that the signal is sent by the data transmission system and reflected back from one or multiple targets. The data transmission system is equipped with a single antenna. The received signal is a superposition of delayed and attenuated copies of the transmitted signal. Assuming there are K stationary targets, i.e. no Doppler shift, the received signal at the radar can be written as

$$r\,(t) = \sum_{k=1}^{K} \beta_k s(t-\tau_k) + n(t), \tag{4-1}$$

where β_k is the reflection factor of the k-th user, τ_k is the is the time-of-arrival (ToA) of the signal of the k-th user, and $n(t)$ is the additive white Gaussian noise (AWGN).

To detect the ToA values, the pulse compression technique is employed. That is, the cross-correlation between the received signal and the modulated transmitted pulse is calculated to detect the peaks of the resulting signal. Mathematically, this operation is performed as

$$\langle s,r\rangle(t) = \int_{\{t'=0\}}^{\infty} s^*\left(t'\right) r\left(t+t'\right)\, dt'. \tag{4-2}$$

Here, $\langle s,r\rangle$ denotes the cross-correlation of s and r, being s^* is the conjugate of s.

In pulse compression, the range resolution capability of the receiver is increased by the pulse compression ratio compared to the conventional system. It is also equal to the SNR gain of pulse compression, and it can be evaluated as [17]

$$PCR = (c_o\tau/2\,)/\,(co/2B) = B\tau, \tag{4-3}$$

where B is the bandwidth and τ is the pulse duration. The term $B\tau$ is called the time-bandwidth product of the modulated pulse, which also equals to pulse compression ratio. It is used to compare different waveforms as it is inversely proportional to the Cramer-Rao lower bound, which is an indicator of the precision of the ToA estimation, i.e. how precise the distance to the target is estimated. It can also be translated to the precision of angle estimation in MIMO radar given the number of receive antennas.

[17] https://www.radartutorial.eu/08.transmitters/Intrapulse%20Modulation.en.html

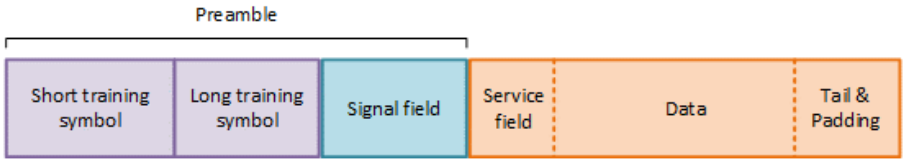


Figure 4-5 IEEE 802.11 Wi-Fi Preamble and Frame Structure

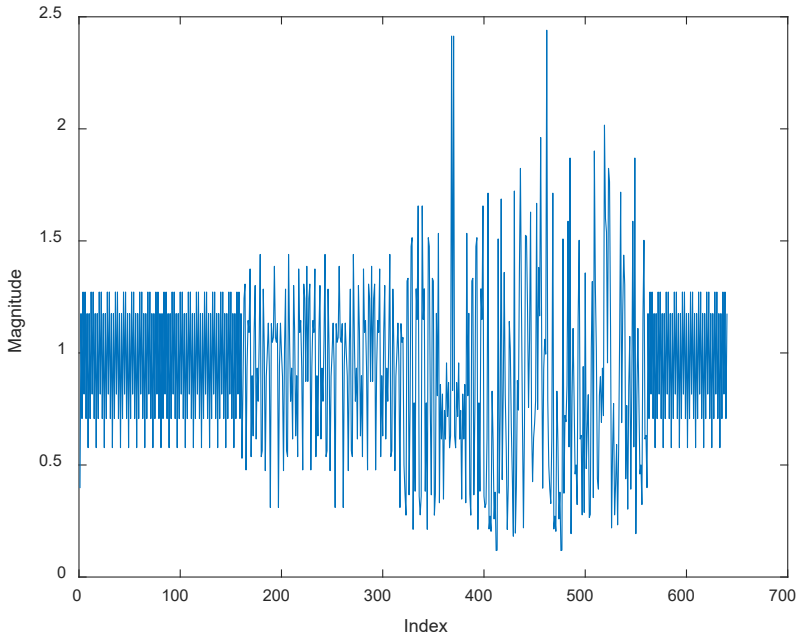


Figure 4-6 Preamble part of IEEE 802.11n Wi-Fi waveform

As transmitted waveform candidates, IEEE 802.11 WLAN (Wi-Fi) and 5G-NR waveforms are used which are OFDM-based waveforms. OFDM is known as an excellent waveform choice for LTE due to its spectral efficiency, easy implementation in hardware through FFT/IFFT, compatibility with high data rates, and robustness against selective fading. It is also used in the 5G-NR standard by a cyclic prefix insertion, called CP-OFDM. To mitigate the effects of multipath, the cyclic prefix length is chosen to be larger than the channel delay spread. In the following, the waveforms and frame structures in Wi-Fi and 5G-NR are briefly presented.

The PHY frame structure in Wi-Fi, as determined in the IEEE 802.11 standards, consists of a preamble, including a signal field, and the data field. The preamble part at the beginning of the frame is used for synchronization and channel equalization at the receiver, while the signal part (also known as header) carries information such as the data rate and the payload length. The data field carries the transmitted data through multiple OFDM symbols. The frame structure in Wi-Fi is depicted in Figure 4-5 [19] and the preamble part of the frame is plotted in Figure 4-6.

5G NR frames have a duration of 10 ms and consist of multiple subframes of 1 ms, which are further divided into slots. The number of slots in a subframe depends on the subcarrier spacing (SCS), where different SCSs correspond to different numerologies. Each slot consists of 14 OFDM symbols. The 5G-NR frame corresponding to Numerology 0, i.e., a SCS of 15 kHz, is illustrated in Figure 4-7. Note that in LTE, only a SCS of 15 kHz is supported.

4.1.3.1 Signal model

As mentioned previously, the so-called pulse compression waveforms are a perfect candidate for sensing. Nevertheless, to minimize the overhead and to efficiently use the spectrum, usually, the transmitted data frames are used also used for sensing. The properties of the autocorrelation function of the sensing signal are described above in section 4.1.2.

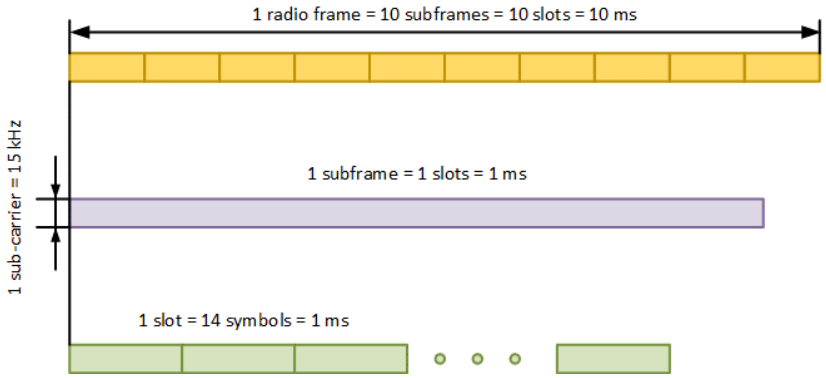


Figure 4-7 5G-NR frame structure for SCS 15 kHz[18]

The transmitted data frames usually do have these properties due to the low correlation between the symbols. Namely, due to the source coder in the data communication system, the correlation between the symbols is reduced, i.e. the source entropy is increased. The channel coder, on the other hand, adds additional redundancy to the data to be transmitted, increasing the correlation between the samples, hence worsening the autocorrelation function. Nevertheless, the autocorrelation function of the transmitted data frames is still good and useful for ToA, i.e. distance estimation.

4.1.3.2 Sensing signal processing strategies

The main objective of the sensing signal processing is to estimate the position of the obstacles from which the transmitted signal was reflected back and received by the receiver. This can be performed in 3-dimensional (3D) space or in 2-dimensional (2D) space. For both cases the required signal processing is similar. For the 3D case, the system and the required processing are slightly more complex compared to the 2D case. Namely, for a 3D case, a planar antenna array is needed where for the 2D case a linear antenna array is sufficient. Additionally, for a 3D case, an antenna array with larger number of antennas is required.

For many applications 2D sensing is sufficient. That is the choice opted for in BeGREEN. In this case it is necessary to find the positions of the obstacles reflecting the signal back to the RU in a polar coordinate system. This means finding the distance from the RU to the obstacle, and the direction at which this reflected signal is returning to the RU.

The distance estimation is performed by estimating the round-trip time (RTT) of the received signal. Namely, the transmitted signal travels to the obstacle that reflects it back and is received by the RU. The time needed for this is called the RTT. The distance to the obstacle can be calculated as:

$$d = c \cdot \frac{t_{oa} - t_{tx}}{2}, \tag{4-4}$$

where c is the speed of light, t_{oa} is the ToA of the signal at the receiver antenna and t_{tx} is the time of the transmission. The transmit time is known and the ToA needs to be estimated. The ToA is usually estimated by performing autocorrelation between the received signal and a copy of the transmitted signal. Having in mind that the transmitted signal has a strong autocorrelation peak for $\tau = 0$, by correlating the received signal with a copy of the transmitted signal, peaks will appear when a reflected signal is received. This peak will coincide with the ToA of the reflection at the receiver.

Estimation of the angle from which the reflected signal arrives at the receiver can be performed using different approaches. It is important to mention that, for estimating the angle, an antenna array of at least 2 receive antennas is needed. Nevertheless, having a larger number of receive antennas and a larger antenna will increase the sensing angular resolution.

[18] https://www.sharetechnote.com/html/5G/5G_FrameStructure.html

The receive antenna array, for 2D sensing is usually a uniform linear array (ULA) and for 3D sensing is usually a uniform planar array (UPA). The approaches to interconnect the antenna(s) and the RF chain can be: 1) have each antenna on a separate receiving chain, i.e. having low noise amplifier, downconverter and analogue-to-digital (A/D) converter for each antenna. This is the solution that can offer the best flexibility to the sensing system. This system is not always possible to be implement in a cost-efficient way. Therefore, other approaches are considered for implementation, such as 2) only analogue and 3) hybrid beamforming.

The **first approach** is the preferred one, i.e. when each antenna has a separate radio chain and a separate A/D converter, which means that each signal coming from each antenna will be fed to a separate A/D converter and will be available for further processing. This is a standard configuration used in digital beamforming systems and the most flexible one. It allows deployment of different approaches for direction-of-arrival (DoA) estimation, i.e. algorithms like MUltiple SIgnal Classification (MUSIC), Matrix Pencil, Estimation of signal parameters via rotational invariant techniques (ESPRIT) or digital beamforming with multiple parallel beams. Despite its associated costs, it is becoming a preferred solution in commercial for sub-6 GHz devices. For the case of mmWave systems, where larger channel bandwidths are used, this kind of solution is becoming too costly and is usually not implemented in commercial products.

The **second approach** to perform DoA estimation is using an analogue beamforming approach. With this approach, an analogue beam is formed using a phased antenna array, and this beam is scanned over the area of interest. This approach is much cheaper compared to the approach described above, since the signals from each antenna are led to an analogue phase shifter and further the phase shifted signals are summed up. The sum of these signals is brought to an A/D converter and digitized. With this approach a single A/D converter is needed, but the limitation is that algorithms for DoA estimation, like MUSIC, Matrix Pencil or ESPRIT, cannot be deployed. Additionally, if sensing in different directions is to be performed, the receiver antenna beam must scan the environment. This cannot be done for the duration of a single data transmission frame, which makes this approach spectrum and power inefficient, i.e. adds additional overhead of the system. Additionally, due to this limitation, a true ISAC cannot be performed since sensing of different directions while communication with a single UE cannot be performed.

Finally, the **third approach** is so called hybrid beamforming, i.e. is a combination of the previous two approaches. This approach has partly the advantages and disadvantages of the above-mentioned approaches. Nevertheless, it is rarely deployed in commercial devices and, therefore, it will not be considered within this project.

Within the BeGREEN project, the first approach was selected for the sub-6 GHz system. The main motivation is that a huge number of sub-6 GHz devices are already designed as MIMO systems. This means that the necessary hardware for the first approach described above is already available, and the sensing implantation will mainly require software changes. For the mmWave systems, due to the high sample rate, having a MIMO system is usually not an option.

4.1.3.3 Angle of arrival estimation

As already mentioned, estimation of the angle of the received reflection can be performed by using different algorithms like MUSIC, Matrix Pencil, ESPRIT or digital beamforming. In this project, initially, digital beamforming is usedto estimate the angle at which the signal is received at the receiver.

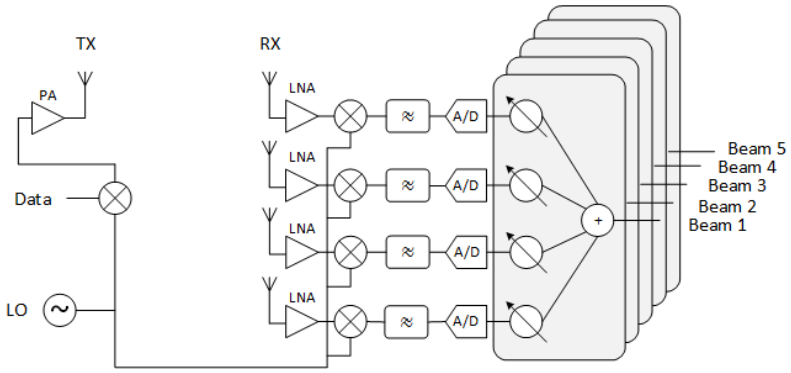


Figure 4-8 Digital beam forming capable transceiver

A simplified architecture of a transceiver capable of digital beamforming is given in Figure 4-8. This transceiver is a single-output-multiple-input system (SIMO), i.e. has one transmit and multiple receiver chains. This architecture enables digital beam steering for the receiver only, which is necessary to obtain the DoA of the signal reflected from different obstacles. On the transmitter side, it is not necessary to have multiple transmit antennas.

To perform beamforming at the receiver, the signals from the different antennas are multiplied by complex coefficients. A set of these complex coefficients for a single beam is called beamforming vector. The signal received from different beams can be calculated as:

$$s_B = B \cdot s, \tag{4-5}$$

Where s_B is a $m \times n$ matrix where n is the signal length and each row represents a signal coming from different direction, i.e. beam. The matrix B is a beam steering matrix where each row is a beamforming vector and has dimension of $m \times p$. Finally, the matrix s is a signal matrix where each row represents the signal received from each antenna and has a dimension of $p \times n$. Expanding (4-5) to obtain:

$$\begin{bmatrix} s_{B11} & \cdots & s_{B1n} \\ \vdots & \ddots & \vdots \\ s_{Bm1} & \cdots & s_{Bmn} \end{bmatrix}_{m \times n} = \begin{bmatrix} b_{11} & \cdots & b_{1p} \\ \vdots & \ddots & \vdots \\ b_{mp} & \cdots & b_{mp} \end{bmatrix}_{m \times p} \times \begin{bmatrix} s_{11} & \cdots & s_{1n} \\ \vdots & \ddots & \vdots \\ s_{p1} & \cdots & b_{pn} \end{bmatrix}_{p \times n}. \tag{4-6}$$

With this approach, a total of m different beams will be created. The number of beamforming vectors is also m, i.e. one beamforming vector for each beam. The number of antennas at the receiver is p and the signal length is n.

In Figure 4-9, an example of parallel beams obtained using digital beamforming is shown. The main advantage is that all of these beams can be formed at the same time in digital domain and larger environment can be sensed without additional spectrum usage.

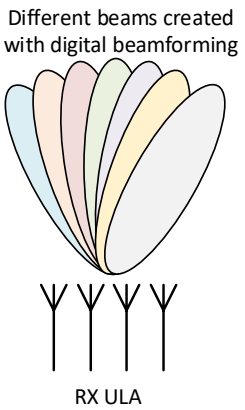


Figure 4-9 Multiple parallel receive beams created with digital beamforming

4.1.3.4 ISAC assisted beam training

Beamforming and electronic beam steering are technologies that exist for more than a few decades. These technologies allow to have a high gain antenna that can be electronically steered to a preferred direction. The main advantage of having a high gain antenna is that lower output power can be used, leading to increased power efficiency of the system. The main disadvantage of high gain antennas is that the antenna radiation pattern is relatively narrow, which significantly reduces the coverage area of such a system.

The main issue with the electronic beam steering approach is that beam training should be performed whenever an initial connection with the UEs needs to be established. This is needed to find the precise direction where a UE is located, i.e. direction to which pointing the beam. Additionally, if, from some reason, the connection between the two devices is lost, retraining must be performed.

Two main approaches for performing beam training are mainly used. The first one is called **exhaustive** search and the second one is called hierarchical search. In the first approach, all the possible directions are searched. Namely, if there are *n* directions in total, all of them will be searched. This means that the complexity of this approach will be *O(n)* if only one of the transceivers in a communication scenario has electronic beam steering possibility, and $O(n^2)$ if both of the transceivers in a communication scenario have electronic beam steering possibility. In case of **hierarchical** beam search, the complexities are *O(log(n))* and $O(log^2(n))$ respectively. Nevertheless, despite the lower complexity of the hierarchical beam search, is has a disadvantage because it must electronically generate wider radiation pattern of the phased array antenna, which means turning off some of the elements and transmitting less power, meaning less coverage. Both approaches are spectrum inefficient and relatively complex, therefore rarely used.

BeGREEN tries to make this approach more convenient and more spectrum efficient, therefore more attractive to be implemented in many devices. It is clear that this approach has good energy saving potential, if the main disadvantages mentioned above can be addressed. The approach proposed here is to leverage the sensing data in order to make the approach more spectrum efficient. Namely, having a heat map as in Figure 4-15, the beam training procedure can be optimized to be more spectrum efficient by searching more precisely the areas where more users are expected and less precise where no users are detected. This should improve the training procedure beyond the one of the hierarchical search and, at the same time, improve some of the disadvantages of this approach. This will make the approach more favourable in the future wireless system, leading to improvement of their energy efficiency.

4.1.4 Simulation results

4.1.4.1 Performance assessment of Wi-Fi and 5G-NR waveforms

To assess the performance of pulse compression for ToA estimation, Monte Carlo simulations are conducted with IEEE 802.11n (OFDM) high throughput Wi-Fi waveform and 5G-NR uplink waveforms. For simplicity, only one target is considered and is assumed that the reflection coefficient is 0.3. In this simulation, whether the ToA is correctly estimated or not (detection probability) is checked, and the root mean square error (RMSE) of the ToA estimation is calculated. The time-bandwidth product and the PAPR of the waveforms after pulse compression are analysed as well.

To generate the waveforms, the Wireless Waveform Generator App of the 5G Toolbox in MATLAB[19] is used. As some default parameters it assumed as channel bandwidth of 20 MHz, sample rate of 30.72 Msps, and SCS of 15 kHz, and assume that the modulation coding scheme is QPSK with a code rate of 1/2 and the message source is PN9. A summary of the default system configuration is given in Table 4-1.

[19] https://www.mathworks.com/

Table 4-1 Default System Parameters

Parameter	Value
Bandwidth	20 MHz
Sample Rate	30.72 Msps
SCS	15 kHz
Modulation	QPSK
Code rate	0.5

Firstly, the detection probability and RMSE of ToA estimation for Wi-Fi and 5G waveforms are compared using default system parameters. The results in Figure 4-10 demonstrate that the 5G-NR waveform provides an advantage of about 2 dB at a detection error rate of 10^{-3}. In terms of RMSE, the 5G-NR waveform requires approximately 2 dB lower SNR to reach an RMSE of 10^{-4} compared to the Wi-Fi waveform, as shown in Figure 4-13. The PAPR and time-bandwidth product of these two waveforms are also compared. To compare the time-bandwidth product, the ratio of peak powers of the pulse compressed waveforms are compared since the time-bandwidth product equals the pulse compression ratio, and the peak power of the signal is amplified with the same ratio as well [18]. The results in Table 4-2 show that the 5G-NR waveform has a higher time-bandwidth product and PAPR, which possibly improves its sensing capability for ToA estimation. Another possible reason is that 5G-NR waveforms are slightly longer than Wi-Fi waveforms under the same system parameters.

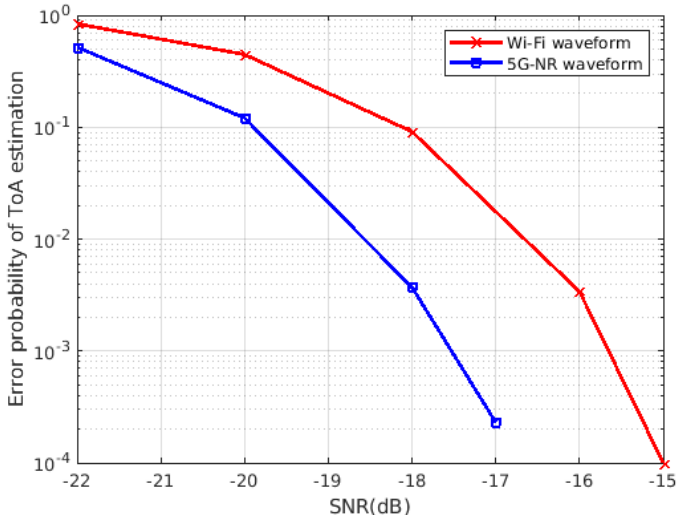


Figure 4-10 Success probability comparison of Wi-Fi and 5G-NR waveforms

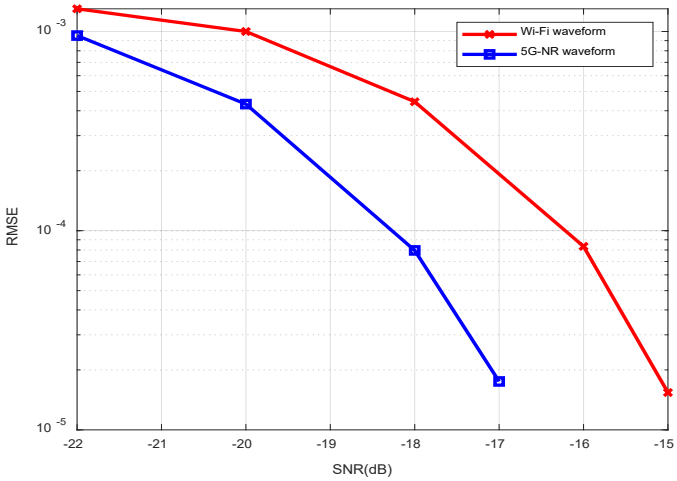


Figure 4-11 RMSE comparison of Wi-Fi and 5G-NR waveforms

Table 4-2 PAPR and Time-Bandwidth Product of the Waveforms

Waveform	PAPR [dB]	Time-Bandwidth Product
5G-NR Uplink	46.85	168
IEEE 802.11n Wi-Fi	46.22	105

The use of the preamble for pulse compression is also investigated as, for that purpose, the whole waveform is used to obtain the results in Figure 4-12 and Figure 4-13. The scenarios in which only the preamble is used are compared from Figure 4-5 of the Wi-Fi waveform, and the whole Wi-Fi waveform for pulse compression is used. The results shown in Figure 4-12 and Figure 4-13 demonstrate that using only preamble leads to a considerable performance loss (of about 13 dB) as it occupies a lower bandwidth. However, it is much simpler to use only the preamble, hence there is performance-complexity tradeoff.

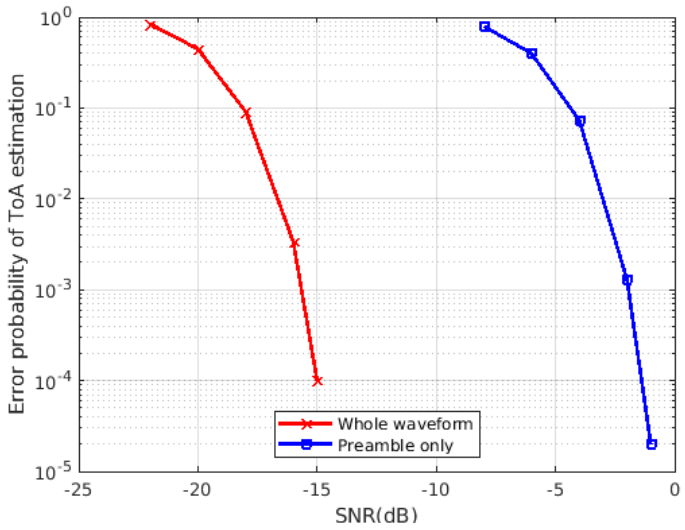


Figure 4-12 Success probability comparison of preamble and whole waveform for Wi-Fi

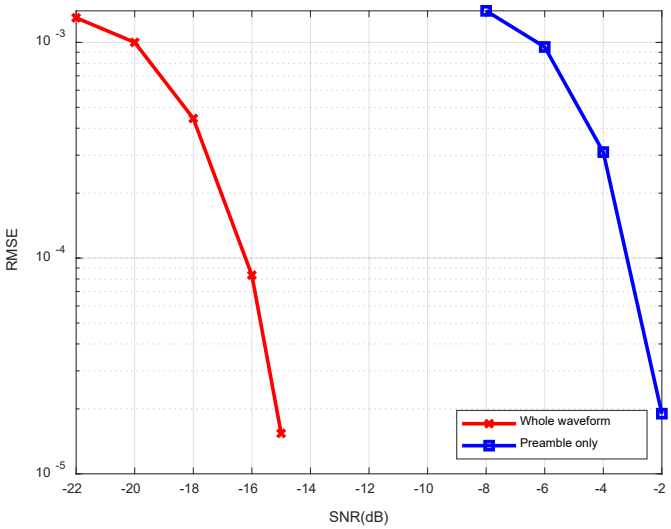


Figure 4-13 RMSE comparison of preamble and whole waveform for Wi-Fi

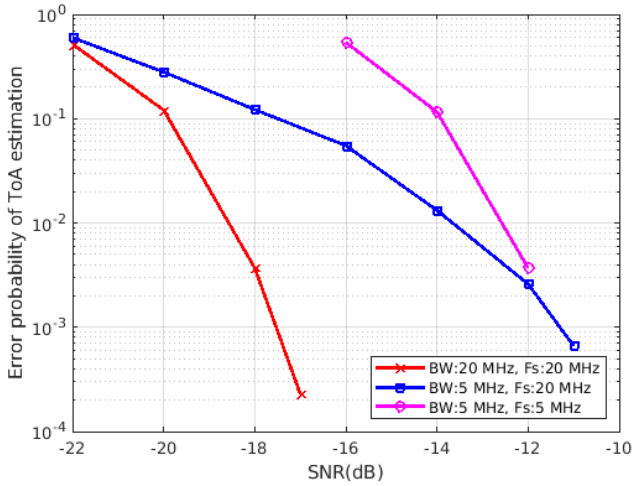


Figure 4-14 Success probability comparison of ToA with different system parameters

Finally, the effect of bandwidth and sample rate to the performance of ToA estimation for the 5G-NR waveforms is analysed. For this purpose, first the bandwidth is changed to 5 MHz in the default system parameters while keeping the sample rate the same, i.e. 20 Msps, and later reduce the sample rate to 5 Msps as well. The results in Figure 4-14 illustrate that decreasing the bandwidth while keeping the sample rate the same degrades the ToA detection performance considerably, especially in the high SNR regime. On the other hand, decreasing the sample rate while keeping the bandwidth greatly degrades the performance in low SNR regime, but in the high SNR regime the performance becomes similar.

4.1.4.2 User density estimation and optimal network resource assignment

With the sensing information that stems from a sensing system, e.g. that shown in Figure 4-16, users can be easily detected and their position can be estimated. This can be used on a larger scale to detect potential users and their positions in a street or in a town. Hence, a map representing spatial user density distribution can be easily created. The spatial distribution of the potential users is extremely important to enable resource optimization, hen. Namely, the network resources can be spatially assigned precisely where they are needed. This will guarantee that an optimal coverage is achieved in the radio network.

An example of a large-scale heat map representing user density obtained using the sensing functionality of the network is shown in Figure 4-15. This map is artificially generated in order to test different optimization algorithms. The red/yellow/light blue lines in the figure represent the streets in which more users are detected and the isolated objects with stronger reflections, i.e. red/yellow colour, represent buildings. The colours on the heat map represent the potential user density per unit of area, giving a valuable information of how many resources should be deployed in a given area.

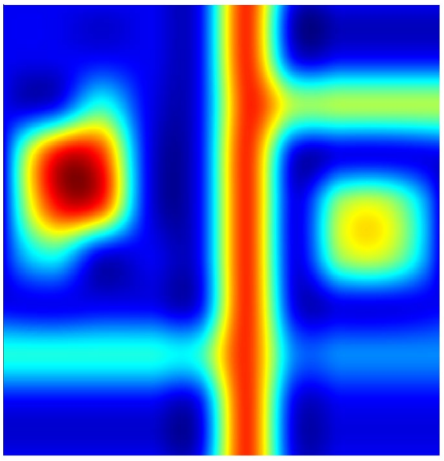


Figure 4-15 Large-scale heat map representing potential user density

4.1.5 BeGREEN sub-6 GHz sensing system

4.1.5.1 Sensing system description

For testing and evaluation of the proposed approaches, a sub-6 GHz ISAC sensing system was developed. The implementation was performed using USRP N321 software defined radios (SDR). The developed system is shown in Figure 4-16. Four SDRs are used to obtain a total of 8 receive channels, given that each radio has only two receiver channels. All the SDRs share the same local oscillator (LO). This is achieved by generating the LO at one of the SDRs and using cables to distribute it to the other SDRs. It is necessary to calibrate the phases between the different SDRs, since they will have random phases due to tolerances of the electronics components built in them. This will allow fully coherent operation of the SDRs, needed to perform beam forming.

The timing synchronization between the SDRs is achieved using the White Rabbit (WR) protocol, i.e. using a WR Ethernet switch model WRS-3/18. With this approach the SDRs time can be synchronized with precision of ±100 ps, which is sufficient for the current application.

The carrier frequency was selected to be in the 5 GHz Industrial-Scientific-Medical (ISM) band, since no explicit licence is needed for transmitting in this band. The lower 5 GHz ISM band is preferred since the SDR's output power depends on it. Namely, the higher the carrier frequency, the lower the output power. Having the carrier lowest possible in the 5 GHz ISM band, will result in output power of not more than 10 dBm, which is quite low, leading to a significantly reduced sensing range.

The ranging precision of the sensing system is proportional to the available bandwidth. In the 5 GHz ISM band, channel bandwidths of up to 160 MHz are available and bandwidths of 320 MHz are envisioned. The SDRs are supporting bandwidths of up to 200 MHz and this value is used in the experiments.

For sensing, different waveforms can be used, as previously discussed in section 4.1.2. In this case a maximum length sequence, with length of 16383 bits, i.e. symbol, was transmitted. This sequence is modulated using binary phase-shift keying (BPSK) modulation. An ULA consisting of 8 patch antennas was designed and manufactured, being the distance between the antennas of $\lambda/2$.

The software for sensing was written as two separate entities. The first part was written in C/C++ and is used to configure the SDRs and to acquire samples upon request. The second part was written in MATLAB and is used to perform the processing needed for the sensing part.

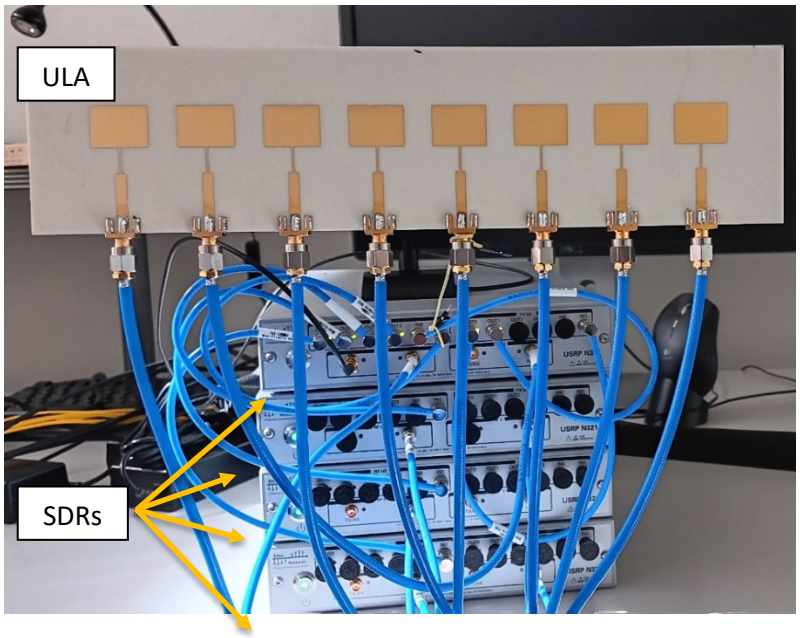


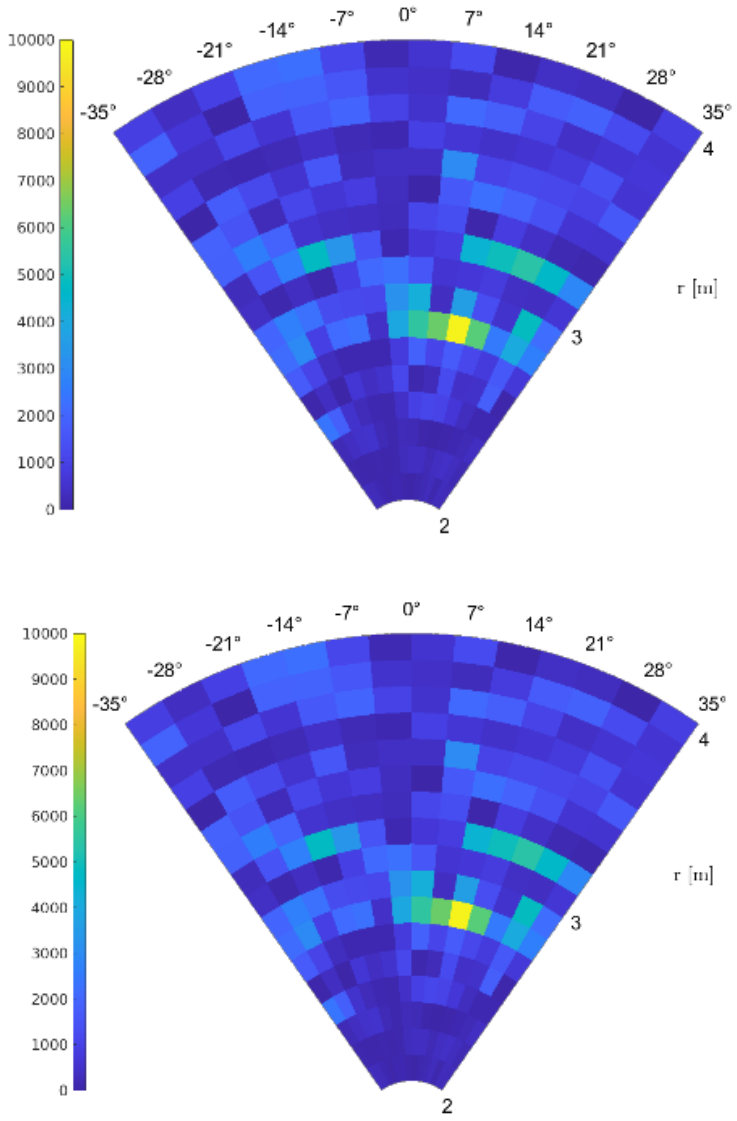
Figure 4-16 Sub-6 GHz sensing system

The system practically transmits a waveform that represents a data frame, but it only performs sensing and not data transmission, since the transmitted waveform is not received by another device. Even if there is no communication with another device, data transmission and sensing are performed simultaneously. There is no loss of generality, since data frames are being transmitted by the system and the sensing is performed.

4.1.5.2 Indoor sensing experiments

The initial tests of the sensing system were performed in an office environment, due to the low available transmit power and equivalent isotopically radiated power (EIRP), of around 10 dBm. The system shown in Figure 4-16 was used to perform the sensing. A background modelling of the surrounding was performed, to remove the static reflections, i.e. clutter. Furthermore, a person was moving in front of the sensing system and the reflections of the person were clearly to be seen. It can be noted that, even with relatively low bandwidth, i.e. not more than 200 MHz, it is clearly possible to detect persons or other moving objects.

The obtained results are shown in Figure 4-17. The figures represent heat map polar plots, where the radial direction represents the distance to the obstacle, and the azimuth represents the angle at which the obstacle is seen. The colour represents how strong the reflected signal is.



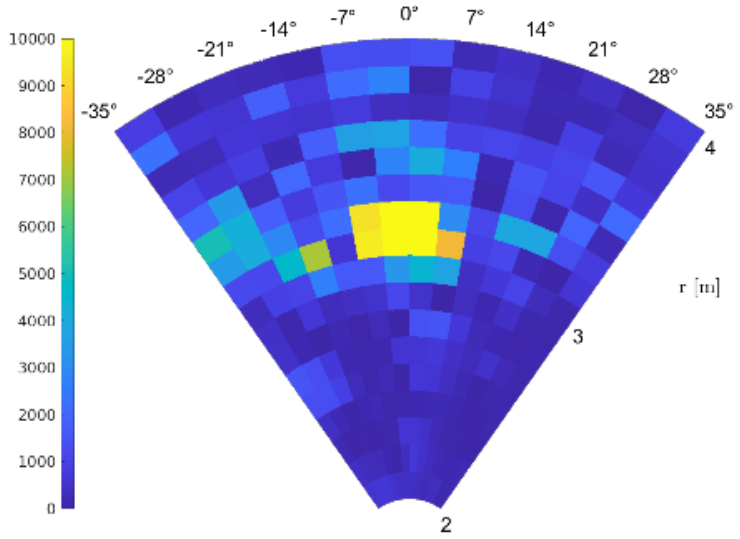


Figure 4-17 Initial tests of of the ISAC system. A person being detected in front of the sensing system

4.2 Self-configuring RIS

BeGREEN D3.1 [1] introduced the initial design of a novel solution, namely Metasurface Absorption and Reflection for Intelligent Surface Applications (MARISA), which leverages Hybrid Reconfigurable Intelligent Surfaces (HRISs), as plug-and-play devices with reflection and power-sensing capabilities. The reference analytical model for the considered wireless HRIS-aided network scenario is introduced here. To start with, stating the HRIS configuration problem accounting for the practical limitations of such HW platform is described.

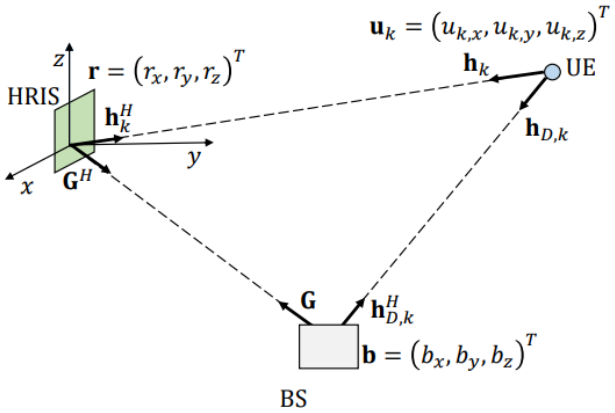


Figure 4-18 Geometrical representation of the considered scenario including the BS, the RIS and the UE.

4.2.1 System model

The scenario is depicted in Figure 4-18, in which a BS equipped with M antennas serve K single-antenna UEs with the aid of an HRIS. The BS is modelled as an ULA, and the HRIS as a planar linear array (PLA) equipped with $N = N_x \times N_z$ meta-atoms, where N_x and N_z denote the number of elements along the x and z axis, respectively. It is assumed that the inter-distance of the BS and HRIS array elements is $\lambda/2$, where $\lambda = \frac{c}{f_c}$ denotes the carrier wavelength, f_c is the corresponding carrier frequency and c is the speed of light. The joint reflection and absorption capabilities of the HRIS are realized through directional couplers whose operation is determined by the parameter $\eta \in [0, 1]$, which is the fraction of the received power that is reflected for communication, while $1 - \eta$ is the amount of absorbed power.

The locations of the BS center, the HRIS center and the k-th UE are denoted by $\boldsymbol{b} \in \mathbb{R}^3$, $\boldsymbol{r} \in \mathbb{R}^3$ and $\boldsymbol{u}_k \in \mathbb{R}^3$, respectively. Focusing on the downlink, the BS transmits data to the k -th UE over a direct line-of-sight (LoS) link $\boldsymbol{h}_{D,k} \in \mathbb{C}^{M\times 1}$ and a reflected link through the HRIS. Such path can be decomposed into the LoS channel $\boldsymbol{h}_k \in \mathbb{C}^{M\times 1}$ through which the HRIS reflects the impinging signal towards the UE, and the LoS channel $\boldsymbol{G} \in \mathbb{C}^{N\times M}$ between the BS and the HRIS. The array response vector at the BS towards the location $\boldsymbol{p} \in \mathbb{R}^3$ is denoted by $\boldsymbol{a}_{BS}(\boldsymbol{p}) \in \mathbb{C}^{M\times 1}$ whose elements are defined as:

$$\{\boldsymbol{a}_{BS}(\boldsymbol{p})\}_{m=1}^{M} \triangleq e^{j\langle \boldsymbol{k}_{PB},(\boldsymbol{b}_m-\boldsymbol{b})\rangle}. \tag{4-7}$$

where $\boldsymbol{k}_{PB}$ is the wave vector, defined as

$$\boldsymbol{k}_{PB} \triangleq \frac{2\pi}{\lambda}\frac{\boldsymbol{p}-\boldsymbol{b}}{\|\boldsymbol{b}-\boldsymbol{p}\|}. \tag{4-8}$$

with $\boldsymbol{b}_m$ denoting the coordinates of the m-th BS antenna element. Likewise, the HRIS array response vector towards the location p is denoted by $\boldsymbol{a}_R(\boldsymbol{p}) \in \mathbb{C}^{M\times 1}$ whose elements are:

$$\{\boldsymbol{a}_R(\boldsymbol{p})\}_{n=1}^{N} \triangleq e^{j\langle \boldsymbol{k}_{PR},(\boldsymbol{r}_n-\boldsymbol{r})\rangle}. \tag{4-9}$$

where $\boldsymbol{k}_{PR}$ is the corresponding wave vector

$$\boldsymbol{k}_{PR} \triangleq \frac{2\pi}{\lambda}\frac{\boldsymbol{p}-\boldsymbol{r}}{\|\boldsymbol{r}-\boldsymbol{p}\|}. \tag{4-10}$$

and $\boldsymbol{r}_n$ is the coordinate of the n-th meta-atom of the HRIS. The overall gain of a generic communication path between two given locations $\boldsymbol{p}, \boldsymbol{q} \in \mathbb{R}^3$ is defined as:

$$\gamma(\boldsymbol{p},\boldsymbol{q}) \triangleq \gamma_0 \left(\frac{d_0}{\|\boldsymbol{p}-\boldsymbol{q}\|}\right)^{\beta}. \tag{4-11}$$

where γ_0 is the channel power gain at a reference distance d_0 and β is the path-loss exponent. Hence, the BS-HRIS and the HRIS-UE$_k$ channels can be expressed as

$$\boldsymbol{G} \triangleq \sqrt{\gamma(\boldsymbol{b},\boldsymbol{r})}\boldsymbol{a}_R(\boldsymbol{b})\boldsymbol{a}_{BS}^H(\boldsymbol{r}) \in \mathbb{C}^{N\times M}. \tag{4-12}$$

$$\boldsymbol{h}_k \triangleq \sqrt{\gamma(\boldsymbol{u}_k,\boldsymbol{r})}\boldsymbol{a}_R(\boldsymbol{u}_k) \in \mathbb{C}^{N\times 1}. \tag{4-13}$$

while the direct BS-UE$_k$ channel is

$$\boldsymbol{h}_{D,k} \triangleq \sqrt{\gamma(\boldsymbol{b},\boldsymbol{u}_k)}\boldsymbol{a}_{BS}(\boldsymbol{u}_k) \in \mathbb{C}^{N\times 1}. \tag{4-14}$$

Thus, the received signal at the k-th UE is

$$y_k = \left(\sqrt{\eta}\boldsymbol{h}_k^H\boldsymbol{\Theta}\boldsymbol{G} + \boldsymbol{h}_{D,k}^H\right)\boldsymbol{W}_{\boldsymbol{s}} + n_k \in \mathbb{C}. \tag{4-15}$$

where $\boldsymbol{\Theta} = \mathrm{diag}\left[\alpha_1 e^{j\theta_1}, \dots, \alpha_N e^{j\theta_N}\right]$ with $\theta_i \in [0, 2\pi]$ and $|\alpha_i| \leq 1, \forall i$ being the phase shifts and the gains introduced by the HRIS, $\boldsymbol{W}_{\boldsymbol{s}} \in \mathbb{C}^{M\times K}$ is the transmit precoding matrix whose k-th column $\boldsymbol{w}_k$ is the transmit precoder of UE$_k$, $\boldsymbol{s} = [s_1, \dots, s_K]^T$ is the transmit symbol vector with $\mathrm{E}[|s_k|^2] = 1, \forall k$, and n_k is the noise term whose distribution is $CN(0, \sigma_n^2)$ (H denotes the Hermitian operator). For tractability, it is assumed that the same HRIS configuration is applied to the incident signals to compute the absorbed and reflected power. Although the two branches of the directional couplers may have dedicated phase shifters, the assumption is that they are the same. This assumption allows us to find a simple and useful relationship between the optimal phase shift configuration for both the reflection and absorption functions, which is beneficial for optimizing the HRIS. Moreover, it is assumed that $\{\alpha_i\}_{i=1}^N$ and $\{\theta_i\}_{i=1}^N$ can be independently optimized.

4.2.1.1 HRIS optimization

Here is described how an HRIS can be endowed with self-configuring capabilities, and how the absence of a

dedicated control channel results in the need for the HRIS of locally estimating the channels towards the BS and the UE, in order to establish and maintain a high-quality reflected path. Firstly, the optimization problem without imposing the absence of the control channel is formulated, and then the difficulty of solving the obtained problem by relying only on local Channel State Information (CSI) at the RIS is elaborated.

The signal-to-interference-plus-noise ratio (SINR) at the k-th UE can be written as

$$\mathrm{SINR}_k = \frac{\left\|\left(\sqrt{\eta}\boldsymbol{h}_k^H\boldsymbol{\Theta}\boldsymbol{G}+\boldsymbol{h}_{D,k}^H\right)\boldsymbol{w}_k\right\|^2}{\sigma_n^2+\sum_{j\neq k}\left\|\left(\sqrt{\eta}\boldsymbol{h}_k^H\boldsymbol{\Theta}\boldsymbol{G}+\boldsymbol{h}_{D,k}^H\right)\boldsymbol{w}_j\right\|^2}. \tag{4-16}$$

where $\boldsymbol{w}_k$ is assumed to be given during the optimization of the configuration of the HRIS. More precisely, $\boldsymbol{w}_k$ is optimized by the BS after the channel estimation phase. The disjoint optimization of $\boldsymbol{w}_k$ and the RIS configuration facilitates the design and deployment of a control channel-free HRIS, which is the focus of this solution. The optimization of $\boldsymbol{w}_k$ is elaborated in further text. The aim is to find the optimal HRIS configuration that maximizes the network sumrate, which is directly related to the SINR at every UE, as exemplified in the toy scenario of Figure 4-19. More precisely, the network sum-rate is defined as

$$R \triangleq \sum_{k=1}^{K}\log_2\left(1+\frac{\left\|\left(\sqrt{\eta}\boldsymbol{h}_k^H\boldsymbol{\Theta}\boldsymbol{G}+\boldsymbol{h}_{D,k}^H\right)\boldsymbol{w}_k\right\|^2}{\sigma_n^2+\sum_{j\neq k}\left\|\left(\sqrt{\eta}\boldsymbol{h}_k^H\boldsymbol{\Theta}\boldsymbol{G}+\boldsymbol{h}_{D,k}^H\right)\boldsymbol{w}_j\right\|^2}\right). \tag{4-17}$$

which results in the following optimization problem:

Problem 1 (SINR-based HRIS configuration):

$$\begin{aligned}&\max_{\boldsymbol{\Theta}} R.\\ &\text{s.t.}\,|\Theta_{ii}|^2 \leq 1, \mathrm{i}=1,\ldots,\mathrm{N}\end{aligned} \tag{4-18}$$

Problem 1 is a fractional program (FP), since it falls into the family of optimization problems involving at least one ratio of two functions. To tackle it, the FP Quadratic Transform method is applied to the objective function, and obtain the following equivalent optimization problem:

Problem 2 (SINR-based HRIS configuration reformulated):

$$\begin{aligned}&\max_{\boldsymbol{\Theta},\mu_k} \log_2(2\mu_k\sqrt{A_k(\boldsymbol{\Theta})}-\mu_k^2 B_k(\boldsymbol{\Theta})).\\ &\text{s.t.}\,|\Theta_{ii}|^2 \leq 1, \mathrm{i}=1,\ldots,\mathrm{N}\\ &\mu_k \in \mathbb{R}, k=1,\ldots,K\end{aligned} \tag{4-19}$$

where:

$$\begin{aligned}A_k(\boldsymbol{\Theta}) &\triangleq \sum_{j=1}^{K}\left\|\left(\sqrt{\eta}\boldsymbol{h}_k^H\boldsymbol{\Theta}\boldsymbol{G}+\boldsymbol{h}_{D,k}^H\right)\boldsymbol{w}_j\right\|^2+\sigma_n^2.\\ B_k(\boldsymbol{\Theta}) &\triangleq \sum_{j\neq k}\left\|\left(\sqrt{\eta}\boldsymbol{h}_k^H\boldsymbol{\Theta}\boldsymbol{G}+\boldsymbol{h}_{D,k}^H\right)\boldsymbol{w}_j\right\|^2+\sigma_n^2\end{aligned} \tag{4-20}$$

Problem 2 has the notable property of being convex in the variables μ_k and $\boldsymbol{\Theta}$ separately. Therefore, it can be efficiently (but sub-optimally) tackled by alternating the optimization over $\boldsymbol{\Theta}$ and over the auxiliary variables $\{\mu_k\}_{k=1}^K$. However, it cannot be locally solved at the HRIS for three reasons: i) the lack of digital signal processing units at the surface, ii) the capability of the HRIS to execute only power measurements, and, more importantly, iii) because Problem 2 requires the CSI of the direct link between the BS and each UE, which can be obtained only through a control and feedback channel.

These considerations call for an alternative approach that considers the constraints on the design and deployment of HRISs, while taking advantage of the massive availability of HRISs in Internet-of-Surfaces (IoS)-based networks.

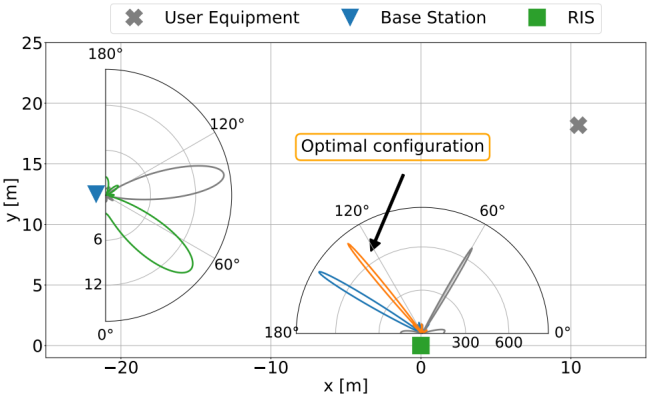


Figure 4-19 Scenario involving BS, UE and RIS

A solution that fulfils these requirements can be obtained by direct inspection of the SINR, and by considering the limitations of the HRIS in terms of RF and signal processing hardware, as well as the absence of a control channel.

A feasible strategy to optimize the SINR for every UE is to optimize the HRIS configuration so that the intensity of $\boldsymbol{h}_k^H \boldsymbol{\Theta} \boldsymbol{G} \boldsymbol{w}_k$ is maximized, $\forall k$ i.e., the end-to-end RIS-assisted channel gain of each user is enhanced. This approach is approximately equivalent to maximizing the SINR_k while ignoring the interfering term $B_k(\boldsymbol{\Theta})$. Upon completion of this optimization, the BS can optimize the precoding matrix $\boldsymbol{W}$, in order to co-phase the direct and the path reflected through the HRIS. Indeed, even though the BS cannot control the HRIS configuration due to the absence of a control channel, it can always estimate the direct channel towards each UE and the equivalent RIS-assisted link.

4.2.1.2 Channel estimation and HRIS configuration

Let us now focus our attention on the optimization of the HRIS configuration, by considering that it is equipped with a single RF power detector. To this end, firstly a closed-form expression for the HRIS configuration that maximizes the reflected power is derived. It is assumed that a training phase exists, during which the BS and each UE transmit a pilot symbol s in order to realize the initial beam alignment procedure. Without loss of generality, it is assumed that a certain degree of synchronization, i.e., the BS and the UEs transmit at different times, but all UEs transmit simultaneously. The latter assumption will be relaxed.

For ease of presentation, the following vector is defined

$$\boldsymbol{v} \triangleq \left[\alpha_1 e^{-j\theta_1}, \dots, \alpha_N e^{-j\theta_N}\right]^T \in \mathbb{C}^{N\times 1}. \tag{4-21}$$

such that $\boldsymbol{\Theta} = \text{diag}(\boldsymbol{v}^{\boldsymbol{H}})$. The signals at the output of the RF combiner, which are obtained from the pilot signals transmitted by the BS and the UEs, can be formulated as

$$\begin{aligned} y_B &= \sqrt{(1-\eta)}\boldsymbol{v}^{\boldsymbol{H}}\boldsymbol{G}\boldsymbol{w}_R s + n \in \mathbb{C}. \\ y_U &= \sqrt{(1-\eta)}\boldsymbol{v}^{\boldsymbol{H}}\boldsymbol{h}_{\Sigma} s + n \in \mathbb{C} \end{aligned} \tag{4-22}$$

where it is assumed that the BS and the UEs emit the same amount of power P and $n \sim N(0, \sigma_n^2)$ is the additive noise term. Let $\boldsymbol{w}_R$ be the optimal BS precoder for the BS-HRIS link. It will be shortly shown that the knowledge of $\boldsymbol{w}_R$ is not explicitly needed to optimize the HRIS configuration. Also, $\boldsymbol{h}_{\Sigma} \triangleq \sum_{k=1}^{K} \boldsymbol{h}_k$ is defined Since the UE-HRIS channel $\boldsymbol{h}_k$ corresponds to the uplink, to use it in the downlink, it is assumed that the channel reciprocity holds.

Therefore, the detected power P_B and P_U from the pilot signals emitted by the BS and the UEs, respectively, can be formulated as

$$P_B = \mathrm{E}[|y_B|^2] = (1-\eta)\left|\boldsymbol{v}^{\boldsymbol{H}}\boldsymbol{G}\boldsymbol{w}_R s\right|^2 + \sigma_n^2.$$
$$P_U = \mathrm{E}[|y_U|^2] = (1-\eta)\left|\boldsymbol{v}^{\boldsymbol{H}}\boldsymbol{h}_{\Sigma} s\right|^2 + \sigma_n^2 \quad (4\text{-}23)$$

In order to be self-configuring, an HRIS needs to infer the channels $\boldsymbol{G}$ and $\boldsymbol{h}_{\Sigma}$ only based on P_B and P_U respectively. This is equivalent to finding the configuration of the HRIS that maximizes P_B and P_U, which in turn corresponds to estimating the directions of incidence of the signals on the HRIS. As a result, the following optimization problem is formulated, whose solution is the HRIS configuration that maximizes P_B.

$$\max_{\boldsymbol{v}} \left|\boldsymbol{v}^{\boldsymbol{H}}\boldsymbol{G}\boldsymbol{w}_R s\right|^2.$$
$$\text{s.t.} \, |v_i|^2 \leq 1, \mathrm{i} = 1, \ldots, \mathrm{N} \quad (4\text{-}24)$$

where v_i is the ith element of $\boldsymbol{v}$. The objective function can be recast as

$$\left|\boldsymbol{v}^{\boldsymbol{H}}\boldsymbol{G}\boldsymbol{w}_R\right|^2 = \boldsymbol{v}^{\boldsymbol{H}}\boldsymbol{a}_R(\boldsymbol{b})\boldsymbol{a}_R^H(\boldsymbol{b})\boldsymbol{v}\left|z_{R,R}\right|^2. \quad (4\text{-}25)$$

where $z_{R,R} \triangleq \sqrt{\gamma(\boldsymbol{b},\boldsymbol{r})}\boldsymbol{a}_{BS}^H(\boldsymbol{r})\boldsymbol{w}_R \in \mathbb{C}$ is the projection of the BS precoding vector $\boldsymbol{w}_R$ onto the BS-HRIS direction. Hence, the optimal HRIS configuration for maximizing the absorbed power from the BS is $v_B \in \mathbb{C}^{N\times 1}$ with

$$v_{B,i} = e^{j<a_{R,i}(\boldsymbol{b})} \; i = 1, \ldots, N. \quad (4\text{-}26)$$

Analogously, the optimal HRIS configuration that maximizes P_U is $\boldsymbol{v}_U \in \mathbb{C}^{N\times 1}$ with

$$v_{U,i} = e^{j<h_{\Sigma,\mathrm{i}}} \; i = 1, \ldots, N. \quad (4\text{-}27)$$

From the previous two equations, it is observed that the optimal HRIS configuration that maximizes the sensed power depends only on the HRIS array response vectors towards the BS and UE directions, but it is independent of the (optimal) BS precoding vector.

Based on $\boldsymbol{v}_B$ and $\boldsymbol{v}_U$, a distributed approach for optimizing the HRIS could be proposed. In particular the following optimization problem is formulated.

Problem 3 (Multi-UE SINR-based HRIS configuration):

$$\max_{\boldsymbol{\Theta}} \frac{\left|\boldsymbol{h}_{\Sigma}^H\boldsymbol{\Theta}\boldsymbol{G}\boldsymbol{w}\right|^2}{\sigma_n^2}.$$
$$\text{s.t.} \, |\Theta_{ii}|^2 \leq 1, \mathrm{i} = 1, \ldots, \mathrm{N} \quad (4\text{-}28)$$

Problem 3 is independent of the direct channels between the BS and the UEs, as well as of the BS precoder $\boldsymbol{w}$: these are fundamental requirements due to the lack of control channel. With the aid of Cauchy–Schwarz's inequality:

$$\left|\boldsymbol{h}_{\Sigma}^H\boldsymbol{\Theta}\boldsymbol{G}\boldsymbol{w}\right|^2 \leq \sum_{k=1}^{K}\left|\boldsymbol{h}_{\mathrm{k}}^H\boldsymbol{\Theta}\boldsymbol{G}\boldsymbol{w}\right|^2. \quad (4\text{-}29)$$

implying that the objective function is a lower bound for the sum of the powers of the signals transmitted by the UEs independently, which are sensed by the HRIS. Notably, the previous inequality becomes an equality if and only if the channels $\boldsymbol{h}_k$ are orthogonal to each other. By using the inequality, the objective function can be reformulated as:

$$\frac{\left|z_R\boldsymbol{v}^H\hat{\boldsymbol{h}}\right|^2}{\sigma_n^2}. \quad (4\text{-}30)$$

where $\hat{\boldsymbol{h}} \triangleq \boldsymbol{h}_{\Sigma}^* \circ \boldsymbol{a}_R(\boldsymbol{b})$ is the equivalent channel that accounts for the overall effect of the aggregate UE-HRIS channels from the HRIS standpoint, and $z_R \triangleq \sqrt{\gamma(\boldsymbol{b},\boldsymbol{r})}\boldsymbol{a}_{BS}^H(\boldsymbol{r})\boldsymbol{w} \in \mathbb{C}$ is the reflected path between the BS and the HRIS for a given precoder $\boldsymbol{w}$ at the BS.

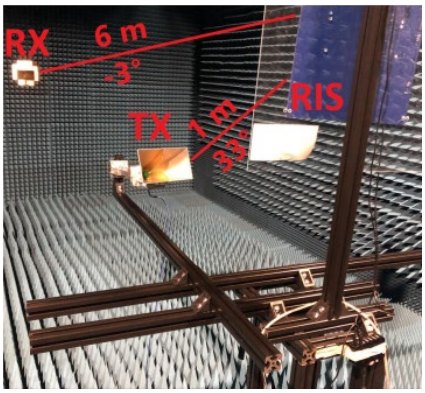


Figure 4-20 Testbed in an anechoic chamber

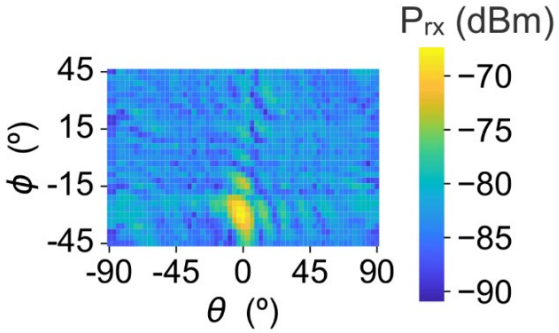


Figure 4-21 Received power over different codebook configurations

Therefore, the HRIS optimal configuration solution of Problem 3 is:

$$v_{BU} = e^{j<\hat{\boldsymbol{h}}} = e^{j<(\boldsymbol{h}_{\Sigma}^{*} \circ \boldsymbol{a}_R(\boldsymbol{b}))} = \boldsymbol{v}_U^{*} \circ \boldsymbol{v}_B. \quad (4\text{-}31)$$

which proves that the HRIS configuration in the absence of a control channel can be inferred solely from v_U and v_B.

4.2.2 RIS Empirical characterization

In what follows, one 10x10 RIS board is characterised in an 8m×5m anechoic chamber to give the reader an upper bound on the gain using and RIS beamforming to a user. Two SDR devices attached to horn antennas with gain $G = 13.5$dBi is used to generate ("TX") and receive ("RX") a continuous stream of OFDM QPSK-modulated symbols with 5 MHz of bandwidth and numerology that meets 3GPP LTE requirements. The transmission power of TX is -30 dBm per subcarrier, and the reference signal received power (RSRP) at RX is sampled. The distances RIS-TX and RIS-RX are $d_{RIS-TX} = 1.1$ m and $d_{RIS-RX} = 6.3$ m, respectively. The azimuth angle of the RX and TX are fixed as $\theta_{RX} = \theta_{TX} = 0°$, and the elevation angles of RIS-TX and RIS-RX are fixed to $\phi_{RX} = -3°$ and $\phi_{TX} = -33°$, respectively. Figure 4-20 shows our testbed.

The received power as a function of the angles θ and ϕ of the pointing beam of the RIS are measured. Figure 4-21 shows the measured power with a colour range for every combination of θ (x-axis) and ϕ (y-axis). Note that for the configuration with the phase difference between the TX and the RX, that is $(\theta_{RX} - \theta_{TX}, \phi_{RX} - \phi_{TX})$, is where the maximum received power is measured. The average difference with other configurations is about 10-15dB.

4.3 ISAC and RIS coexistence and convergence

ISAC is a functionality that will be implemented in future B5G and 6G networks. This functionality will be implemented in the RUs where the data transmission functionality is also implemented. Due to the higher carrier frequencies used in the B5G and 6G systems, the coverage of the radio network will deteriorate due

to the poorer RF penetration through walls and buildings. This will affect both data transmission and sensing.

RIS is a solution that will be used to alleviate this problem and it will improve the coverage at the back of obstacles, without worsening the energy footprint of the network. Nevertheless, for RIS to support the sensing capabilities of the joint ISAC system, its operation must be tightly coordinated with the operation of the RU performing sensing. BeGREEN is developing algorithms and considers an architecture that will enable sensing functionality over RIS (see PoC2 in BeGREEN D5.1 [5]). This approach will facilitate sensing over RIS in areas which have denied coverage without additional energy usage.

4.4 Interference analysis in relay-enhanced scenarios

The deployment of relays to improve a B5G network involves, among other tasks, the decision of which radio resources the relay will use. As described in BeGREEN D3.1 ([1] Section 4.3.1), based on spectrum utilisation, two operation modes are considered for relaying: in-band and out-band. In in-band mode, relays use the same spectrum resources as the cellular system, thus no additional spectral resources are required. Whereas in out-band mode, different spectrum resources are allocated to backhaul and access links. While the operation in out-band is usually considered to avoid mutual interferences between BSs and relays, 5G/B5G networks offer new elements like beamforming and PRBs management that allow to play with the spatial dimension, and to use the available spectrum more efficiently. In this respect, this study consider a B5G scenario enhanced with relays that includes mMIMO at the BS, and then it analyses three possible spectrum allocation configurations between BS and relays that differ in the way how the spectrum is shared by the BS and the relays and thus on the associated interferences.

The model assumes a 5G NR mMIMO system with M antennas at the BS and K single-antenna user equipment (UE) devices numbered as k=1, …, K, with M>>K. Frequency-domain beamforming is considered, so that it is possible to have multiple simultaneous beams using different resource blocks. In particular, the BS has N resource blocks and generates K beams, each with a given number of resource blocks. The downlink direction is considered in a single cell scenario.

A relay is also connected to the BS. The relay does not use massive MIMO and serves a total of K_R UE devices, numbered as $\mathbf{k}=K+1,\ldots, K+K_R$. From the perspective of the connectivity BS-relay, the relay is just modeled as one of the other UEs. Without loss of generality, it is assumed that the K-th UE is the relay, and the rest of UEs are uniformly distributed throughout the coverage area.

Each UE connected to directly the BS or through the relay requires a certain bit rate denoted as R_k for $k=1,\ldots,K+K_R$. The bit rate of the relay is given by the aggregate of all it served UEs, that is:

$$R_K = \sum_{k=K+1}^{K+K_R} R_k. \tag{4-32}$$

The following three strategies are considered to decide how the available spectrum is distributed between the BS and the relay:

- **Strategy 1 "Out-band":** the relay operates in out-band mode, so there are no interferences from the relay to UEs served directly by the BS, neither from the BS to the users served by the relay.
- **Strategy 2 "Partial in-band":** the relay operates with a subset of the resources assigned to the BS, in other words that is partial in-band relay operation.
- **Strategy 3 "Full in-band":** the relay operates with the same resources assigned to the BS, in other words that is full in-band relay operation.

The three strategies are detailed below, then the power consumption model is developed, and the performance of the different solutions is currently being analysed and compared. Power consumption results and energy efficiency benefits will be included in BeGREEN D3.3.

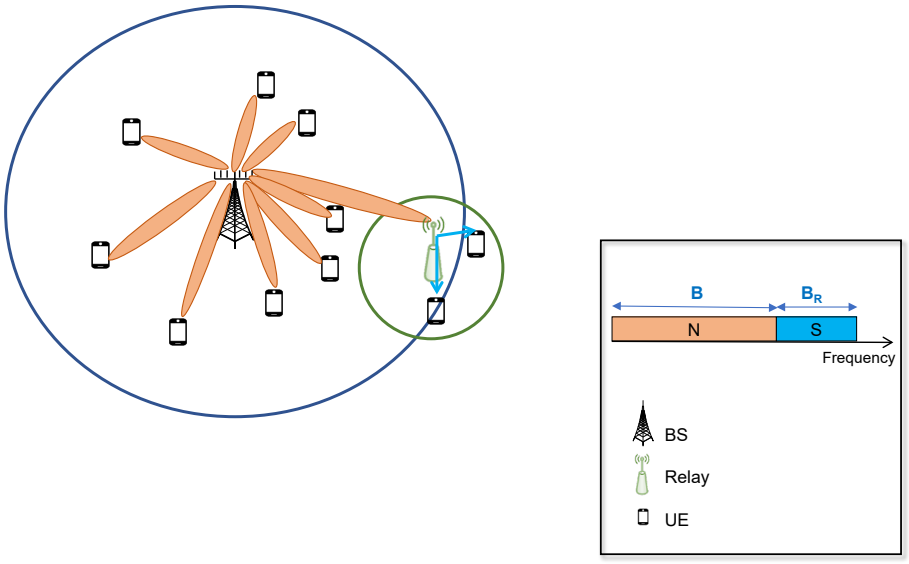


Figure 4-22 - Relay-enhanced strategy 1

4.4.1 Strategy 1

In this first configuration (strategy 1) a separate spectrum band is assigned to the deployed relays, so that they operate in out-band mode, as shown in Figure 4-22.

In this strategy, the BS uses a total bandwidth *B*, corresponding to N resource blocks, to serve the *K* UEs, the relay uses a total bandwidth B_R, corresponding to S resource blocks, to serve the K_R UEs connected to it, and each UE uses a different portion of this bandwidth B_R.

4.4.2 Strategy 2

In order not to increase the spectrum needed by the network when relays are deployed and reduce the interferences from the relay, the strategy 2 is proposed where the relay operates with a subset of the resources assigned to the BS, as shown in Figure 4-23, in other words it is a partial in-band relay operation. In strategy 2, the relay uses a subset (*S*) of the *N* resource blocks in the BS to serve the K_R UEs connected to it and each UE uses some of the *S* resource blocks.

The "Relay's Incidence Area" (RIA) is defined as the entire area delimited between the imaginary lines that join the BS and the edges of the relay coverage area, marked in grey in Figure 4-23.

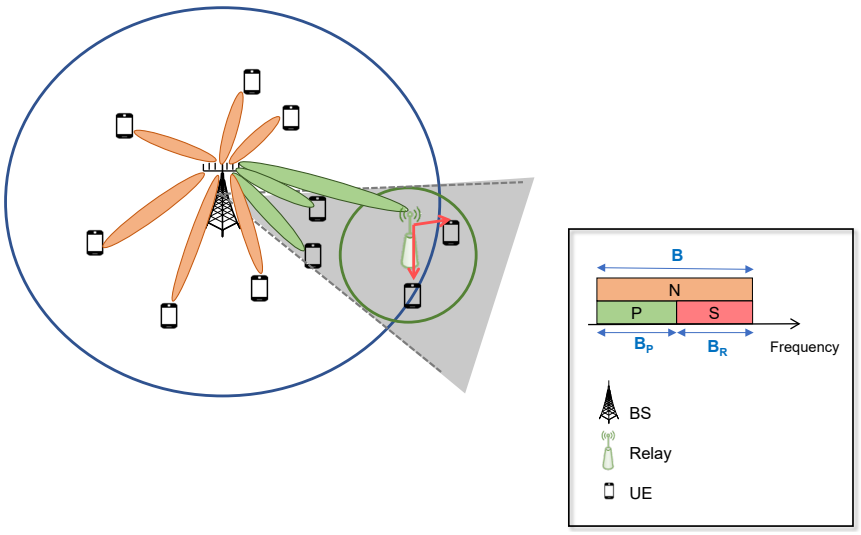


Figure 4-23 Relay-enhanced strategy 2

Since the Relay needs S resource blocks to serve its UEs, in order to avoid interference, it is assumed that it will use resources that are not being used in the beam that serves the Relay, nor in the beams that serve UEs in the RIA, therefore in the direction where the Relay is located, the BS cannot use all N resources but N-S. Strict synchronization between Relay and BS is required. A number of 'i' UEs are considered inside the RIA, numbered as k=K-i, ..., K-1. With this proposal for the reuse of radio resources, there is no interference, ideally, from the BS towards the UEs served by the Relay.

The BS transmitter general architecture to implement frequency-domain beamforming is shown in Figure 4-24. Note that frequency-domain beamforming requires digital beamforming [17][21]. The MAC scheduler, residing in the BS, takes scheduling decisions based on channel-quality reports obtained from the devices, and controls the set of resource blocks for each transmission. The use of resources by the beams and the Relay can be changed every T (scheduling decision, typically once per slot). The resource-block mapping takes the modulation symbols to be transmitted on each antenna port and maps them to the set of available resource elements in the set of resource blocks assigned by the MAC scheduler for the transmission [1].

4.4.3 Strategy 3

In strategy 3 (see Figure 4-25) the relay and the BS use the same spectrum resources without any coordination between them. Then, the relay uses all the total bandwidth *B* assigned to the BS to serve the K_R UEs connected to it and each UE uses a different portion of this total bandwidth *B*.

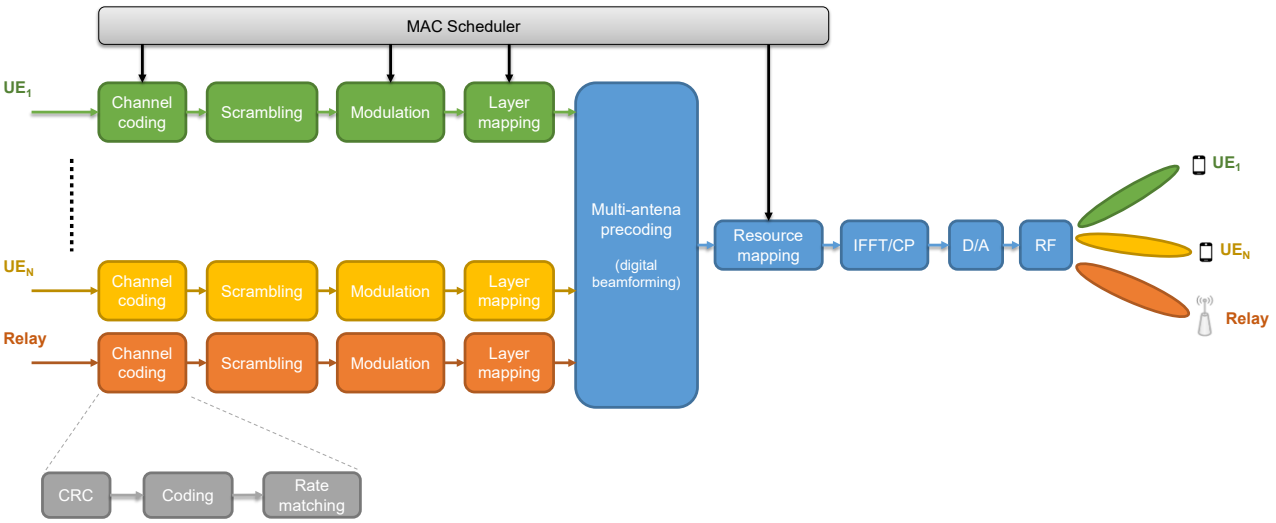


Figure 4-24 BS transmission architecture for frequency-domain beamforming

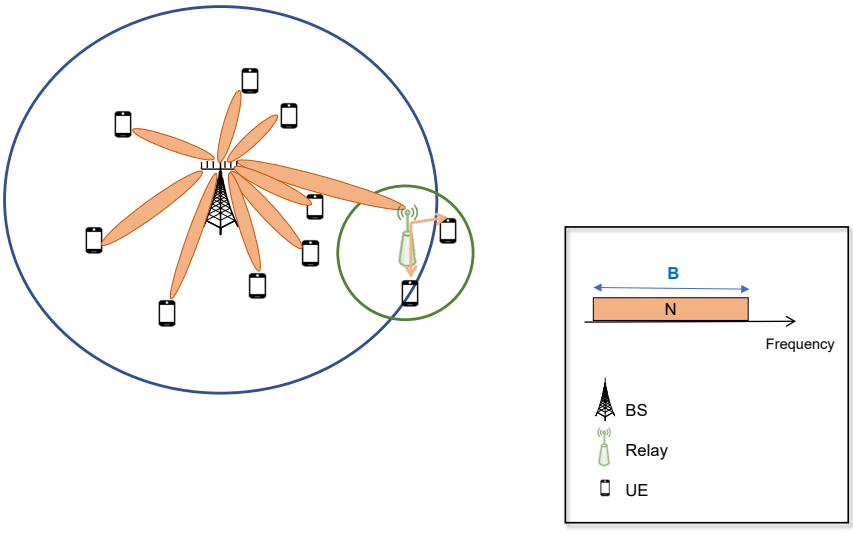


Figure 4-25 Relay-enhanced strategy 3

4.4.4 Power consumption model

This section presents the power consumption model in the considered strategies. For this purpose, the formulation of the required transmitted power at the BS and the relay to support the bit rate of the UEs is presented first. Then, the total power consumption in the scenario is presented.

4.4.4.1 Power consumption in strategy 1

The performance of the massive MIMO system depends on the precoding strategy that determines how to transmit the signals of the K UEs through the M antennas of the BS. Different precoding techniques exist for computing the precoding matrix. This document considers the Maximum Ratio (MR) technique, based on the expressions from chapter 3 of [22].

Following the expressions of chapter 3.3 of [22] adapted to the notation used in this document, the SINR obtained by users connected to the BS is given by:

In strategy 1:

$$SINR_k = \frac{M \cdot P_{T,k} \cdot \beta_k \cdot c_{CSI}}{P_N + \sum_{k'=1}^{K} P_{T,k'} \cdot \beta_k}. \quad (4\text{-}33)$$

where $P_{T,k}$ is the transmit power of the BS to the k-th UE, k=1,...,K. This power is split across all the antennas in accordance with the precoding matrix, and $P_{T,R}$ is the transmit power of the relay. $P_N = N_o \cdot B$ is the noise power measured over all the bandwidth B. β_k is the large scale fading in the link between the BS-UE k. It accounts for the path loss, shadowing and antenna gains, that is:

$$\beta_k = \frac{G_T \cdot G_R}{L_k}. \quad (4\text{-}34)$$

where G_T and G_R are the transmit and receive antenna gains and L_k is the total propagation loss in the link BS-UE k, that includes the path loss and the shadowing.

Moreover, the term c_{CSI} ranges between 0 and 1 and reflects the quality of the channel estimation that impacts on the computation of the precoding matrix. Following [22] and [23] it is defined as:

$$c_{CSI} = \frac{\gamma_k}{\beta_k}. \quad (4\text{-}35)$$

where γ_k is the mean square of the channel estimate that includes both fast fading and large-scale fading. MMSE channel estimation is assumed. The mean square error of the channel estimation is $(\beta_k - \gamma_k)$. For ideal channel estimation the error is 0 so $\beta_k = \gamma_k$ and $c_{CSI} = 1$.

The term in the denominator in (4-33) accounts for the noise and the total intra-cell interference.

In strategy 1, this interference in (4-33) results, on the one hand, from the non-orthogonality of the channels of the different users, so that each UE k' different from k will generate interference to the UE k. On the other hand, the interfering term in the denominator corresponding to the same user $k'=k$ is due to the beamforming gain uncertainty [24].

The spectral efficiency of the k-th UE can be estimated as:

$$S_k = log_2(1 + SINR_k). \quad (4\text{-}36)$$

And the bit rate achieved by the UEs, is given by:

$$R_k = B \cdot \varepsilon \, log_2(1 + SINR_k). \quad (4\text{-}37)$$

where the term ε is an efficiency factor that accounts for the overheads due to cyclic prefix, reference signals, control plane signalling, etc.

Using (4-37), the minimum required SINR to achieve the bit rate R_k of the k-th UE is given by:

$$SINR_{min,k} = 2^{\frac{R_k}{B \cdot \varepsilon}} - 1. \quad (4\text{-}38)$$

From (4-33), the required transmitted power by the BS to the *k*-th UE is given in terms of the *SINR* requirement $SINR_{min,k}$ as:

$$P_{T,k} = \frac{SINR_{min,k}(P_N + \beta_k P_{T,BS})}{M \cdot \beta_k c_{CSI}}. \quad (4\text{-}39)$$

where $P_{T,BS}$ is the total transmitted power of the BS for all the UEs, that is:

$$P_{T,BS} = \sum_{k=1}^{K} P_{T,k}. \quad (4\text{-}40)$$

And finally, by substituting (4-39) in (4-40):

$$P_{T,BS} = \sum_{k=1}^{K} \frac{SINR_{min,k}(P_N + \beta_k P_{T,BS})}{M \cdot \beta_k c_{CSI}}. \quad (4\text{-}41)$$

From this relationship, the total required transmit power at the BS to support all the UEs in strategy 1 is given by:

$$P_{T,BS} = \frac{\sum_{k=1}^{K} \frac{SINR_{min,k} P_N}{M \cdot \beta_k c_{CSI}}}{1 - \sum_{k=1}^{K} \frac{SINR_{min,k}}{M \cdot c_{CSI}}}. \quad (4\text{-}42)$$

Regarding the required transmitted power at the relay, since the relay does not support massive MIMO, it is assumed that the transmissions of the K_R UEs are multiplexed in frequency. Then, considering that the total bandwidth available at the relay is denoted as B_R, the bandwidth occupied by the transmission to the *k*-th UE is a fraction of the total bandwidth denoted as $B_k = \alpha_k B_R$, where $0 < \alpha_k < 1$. The general term α_k allows modelling different strategies. For example, an equal bandwidth allocation per UE can be modelled by setting $\alpha_k = 1/K_R$, or an allocation proportional to the bit rate requirement can be modelled by setting $\alpha_k = R_k/(R_{K+1} + ... + R_{K+KR})$, etc.

The *SINR* requirement of the *k*-th UE in the link relay-UE is given by:

$$SINR_{R,min,k} = 2^{\frac{R_k}{B_k \cdot \varepsilon_R}} - 1. \quad (4\text{-}43)$$

where the term ε_R is an efficiency factor $0 < \varepsilon_R \leq 1$ that accounts for the overheads associated to cyclic prefix, reference signals, control plane signaling, etc.

The required *SINR* relates with the transmitted power by the relay to the *k*-th UE as follows:

$$SINR_{R,min,k} = \frac{P_{TR,k} \beta_{R,k}}{N_o B_k}. \quad (4\text{-}44)$$

where $\beta_{R,k}$ is the large scale fading in the link relay-UE_k accounting for antenna gains, path loss and shadowing, and $N_o \cdot B_k$ is the noise power in the band used by the *k*-th UE. Similar to β_k defined in (4-34), $\beta_{R,k}$ is:

$$\beta_{R,k} = \frac{G_{TR} \cdot G_R}{L_{Rk}}. \quad (4\text{-}45)$$

where G_{TR} and G_R are the transmit (at Relay) and receive antenna gains L_{Rk} is the total propagation loss in the link Relay-UE *k* respectively, that includes the path loss and the shadowing.

Then, the required transmitted power by the relay to the *k*-th UE is given by:

$$P_{TR,k} = \frac{SINR_{R,min,k} N_o B_k}{\beta_{R,k}}. \quad (4\text{-}46)$$

And the total power transmitted by the relay, is the aggregate of the transmitted powers of all the UEs:

$$P_{T,R} = \sum_{k=K+1}^{K+K_R} \frac{SINR_{R,min,k} N_o B_k}{\beta_{R,k}}. \quad (4\text{-}47)$$

The total power consumption at the BS and the relay will be linearly dependent on the transmitted powers $P_{T,BS}$ and $P_{T,R}$ including a constant term per antenna (TRX) that accounts for the circuit power consumption. This yields:

$$P_{TOT} = a_{BS}P_{T,BS} + M \cdot P_{o,BS} + a_R P_{T,R} + P_{o,R}. \quad (4\text{-}48)$$

where a_{BS}, a_R are the scaling factors of the transmitted power at the BS and relay, respectively. Similarly, $P_{o,BS}$ and $P_{o,R}$ are the circuit power consumption per antenna at the BS and relay, respectively.

4.4.4.2 Power consumption in strategy 2

Following the same procedure as in the previous section, in strategy 2, the *SINR* obtained by users connected to the BS is given by:

- for the Relay and the UEs inside its incidence area, i.e. users *k*=*K*-*i* to *k*=*K*-1, and the relay *k*=*K*:

$$SINR_k = \frac{M \cdot P_{T,k} \cdot \beta_k \cdot c_{CSI}}{P_{N_p} + \sum_{k'=K-i}^{K} P_{T,k'} \beta_k + \sum_{k'=1}^{K-i-1} \rho \cdot P_{T,k'} \beta_k}. \quad (4\text{-}49)$$

- for the UEs outside the RIA, i.e. *k*=1 to *k*=*K*-*i*-1:

$$SINR_k = \frac{M \cdot P_{T,k} \cdot \beta_k \cdot c_{CSI}}{P_N + \sum_{k'=1}^{K} P_{T,k'} \beta_k + P_{T,R} \cdot \beta_{R,k}}. \quad (4\text{-}50)$$

where $P_{T,k}$ is the transmit power of the BS to the *k*-th UE, *k*=1,...,*K*. $P_N = N_o \cdot B$ is the noise power measured over all the bandwidth *B* with N_o the noise power spectral density, and $P_{N_p} = N_o \cdot B_P$ is the noise power measured over bandwidth B_P. β_k and $\beta_{R,k}$, defined in (4-34) and (4-45), are the large scale fading in the link between the BS-UE *k* and Relay-UE *k* respectively. They account for the path loss, shadowing and antenna gains.

The term c_{CSI} is defined in (4-35), and the denominator in (4-49) and (4-50) accounts for the noise and the total intra-cell interference in the corresponding area.

This interference results in (4-49), on the one hand, from the non-orthogonality of the channels of the *i* users located in the RIA that use beams with the same part of the spectrum. On the other hand, the interfering term in the denominator corresponding to the same user *k'*=*k* is due to the beamforming gain uncertainty [22]. And finally, the (*K*-*i*-*1*) UEs who are outside the RIA, generate an interference on user *k* proportional to ($\rho = \frac{B_P}{B}$) the part of the spectrum that coincides the spectrum used by the users in the RIA.

Similarly, the term in the denominator from (4-50) accounts for the noise and the total intra-cell interference in the case of UEs outside the RIA. It is worth noting that this interference results, on the one hand, from the non-orthogonality of the channels of the different users (all), so that each UE *k'* different from *k* will generate interference to the UE *k*, and the same user *k*. On the other hand, the interference from the relay to users outside the RIA.

The spectral efficiency of the *k*-th UE can be estimated as (4-36), and the bit rate achieved by the UEs outside the RIA, i.e. *k*=1,..*K*-*i*-1, is given by (4-37). Accordingly, the minimum required *SINR* for these UEs is (4-38).

In the case of the Relay and UEs inside its incidence area, i.e. *k*=*K*-*i*,..,*K*, since the beam that serves each one uses a bandwidth B_P,. the bit rate achieved is given by:

$$R_k = B_P \cdot \varepsilon \, log_2(1 + SINR_k). \quad (4\text{-}51)$$

And the minimum required *SINR* to achieve the bit rate R_k is given by:

$$SINR_{min,k} = 2^{\frac{R_k}{B_P \cdot \varepsilon}} - 1. \quad (4\text{-}52)$$

From (4-49) and (4-50), the required transmitted power by the BS to the *k*-th UE is given in terms of the *SINR*

requirement $SINR_{min,k}$, respectively as:

- for the Relay and the UEs inside its incidence area:

$$P_{T,k} = \frac{SINR_{min,k}\left(P_{N_p}+\beta_k\cdot\left(P_{T,BS}-\sum_{k'=1}^{K-i-1}(1-\rho)\cdot P_{T,k'}\right)\right)}{M\cdot\beta_k\, c_{CSI}}. \quad (4\text{-}53)$$

- for the UEs outside the RIA:

$$P_{T,k} = \frac{SINR_{min,k}\left(P_N+\beta_k\cdot P_{T,BS}+P_{T,R}\cdot\beta_{R,k}\right)}{M\cdot\beta_k\cdot c_{CSI}}. \quad (4\text{-}54)$$

where $P_{T,BS}$ is the total transmitted power of the BS for all the UEs, as defined in (4-40).

And finally, by substituting (4-53) and (4-54) in (4-40):

$$P_{T,BS} = \sum_{k=K-i}^{K}\frac{SINR_{min,k}\left(P_{N_p}+\beta_k\cdot\left(P_{T,BS}-\sum_{k'=1}^{K-i-1}(1-\rho)\cdot P_{T,k'}\right)\right)}{M\cdot\beta_k\cdot c_{CSI}} + \sum_{k=1}^{K-i-1}\frac{SINR_{min,k}\left(P_N+\beta_k\cdot P_{T,BS}+P_{T,R}\cdot\beta_{R,k}\right)}{M\cdot\beta_k\cdot c_{CSI}}. \quad (4\text{-}55)$$

And now by substituting (4-54) in (4-55):

$$P_{T,BS} = \sum_{k=K-i}^{K}\frac{SINR_{min,k}\left(P_{N_p}+\beta_k\cdot\left(P_{T,BS}-\sum_{k'=1}^{K-i-1}(1-\rho)\cdot\frac{SINR_{min,k'}\left(P_N+\beta_{k''}\cdot P_{T,BS}+P_{T,R}\cdot\beta_{R,k'}\right)}{M\cdot\beta_{k''}\cdot c_{CSI}}\right)\right)}{M\cdot\beta_k\cdot c_{CSI}} + \sum_{k=1}^{K-i-1}\frac{SINR_{min,k}\left(P_N+\beta_k\cdot P_{T,BS}+P_{T,R}\cdot\beta_{R,k}\right)}{M\cdot\beta_k\cdot c_{CSI}}. \quad (4\text{-}56)$$

From this relationship, the total required transmit power at the BS to support all the UEs in strategy 2 is given by:

$$P_{T,BS} = \frac{\sum_{k=K-i}^{K}\frac{SINR_{min,k'}\cdot P_{N_p}}{M\cdot\beta_k\cdot c_{CSI}}-\sum_{k=K-i}^{K}\frac{SINR_{min,k'}\cdot(1-\rho)}{(M\cdot c_{CSI})^2}\cdot\sum_{k'=1}^{K-i-1}\frac{SINR_{min,k'}\left(P_N+P_{T,R}\cdot\beta_{R,k'}\right)}{\beta_{k'}}+\sum_{k=1}^{K-i-1}\frac{SINR_{min,k'}\cdot\left(P_N+P_{T,R}\cdot\beta_{R,k}\right)}{M\cdot\beta_k\cdot c_{CSI}}}{1-\sum_{k=K-i}^{K}\frac{SINR_{min,k}}{M\cdot c_{CSI}}+\sum_{k=K-i}^{K}\frac{SINR_{min,k}\cdot(1-\rho)}{(M\cdot c_{CSI})^2}\cdot\sum_{k'=1}^{K-i-1}SINR_{min,k'}-\sum_{k=1}^{K-i-1}\frac{SINR_{min,k}}{M\cdot c_{CSI}}}. \quad (4\text{-}57)$$

Regarding the required transmitted power at the relay, it is assumed that the interference coming from the beams pointing outside the relay incidence area is negligible for the UEs connected to the relay, so it can be calculated as in strategy 1 with (4-47), and the total power consumption can be computed following (4-48).

4.4.4.3 Power consumption in strategy 3

Likewise, the two previous strategies, in strategy 3, the *SINR* obtained by users connected to the BS is given by:

$$SINR_k = \frac{M\cdot P_{T,k}\cdot\beta_k\cdot c_{CSI}}{P_N+\sum_{k'=1}^{K}P_{T,k''}\beta_k+P_{T,R}\cdot\beta_{R,k}}. \quad (4\text{-}58)$$

according to the terms definitions that have already been described.

In that case the interference in the denominator of (4-58) results from the non-orthogonality of the channels of the different users (all), and the interference from the relay. Note that in the case of the relay, i.e. *k*=*K*, the term L_{RK} represents the degree of cancellation of the self-interference that exists between transmission and reception.

From (4-58) the required transmitted power by the BS in strategy 3 to the *k*-th UE is given in terms of the *SINR* requirement $SINR_{min,k}$ as:

$$P_{T,k} = \frac{SINR_{min,k}\left(P_N+\beta_k\cdot P_{T,BS}+P_{T,R}\cdot\beta_{R,k}\right)}{M\cdot\beta_k\cdot c_{CSI}}. \quad (4\text{-}59)$$

By substituting (4-59) in (4-40):

$$P_{T,BS} = \sum_{k=1}^{K} \frac{SINR_{min,k}\left(P_N+\beta_k \cdot P_{T,BS}+P_{TR}\cdot\beta_{R,k}\right)}{M\cdot\beta_k\cdot c_{CSI}}. \quad (4\text{-}60)$$

From this relationship, the total required transmit power at the BS to support all the UEs in strategy 3 is given by:

$$P_{T,BS} = \frac{\sum_{k=1}^{K} \frac{SINR_{min,k}\left(P_N+P_{T,R}\cdot\beta_{R,k}\right)}{M\cdot\beta_k\cdot c_{CSI}}}{1-\sum_{k=1}^{K}\frac{SINR_{min,k}}{M\cdot c_{CSI}}}. \quad (4\text{-}61)$$

Regarding the required transmitted power at the relay, in the case of strategy 3, the users served by the relay will have interference from the beams of the BS pointing to the relay incidence area (i.e. users *k*=*K*-*i* to *k*=*K*-1, and the relay *k*=*K*), so the required *SINR* relates with the transmitted power by the relay to the *k*-th UE as follows:

$$SINR_{R,min,k} = \frac{P_{TR,k}\beta_{R,k}}{N_o B_k+\sum_{k'=K-i}^{K} P_{T,k'}\beta_k\cdot\frac{B_k}{B}}. \quad (4\text{-}62)$$

where β_k , defined in (4-34), is the large scale fading in the link BS-UE$_k$ accounting for antenna gains, path loss and shadowing, and B_k/B is for the portion of the BS transmitted power to UE *k'*, in the band used by the *k*-th UE. Then, the required transmitted power by the relay to the *k*-th UE is given by:

$$P_{TR,k} = \frac{SINR_{R,min,k}\left(N_o B_k+\sum_{k'=K-i}^{K} P_{T,k'}\beta_k\cdot\frac{B_k}{B}\right)}{\beta_{R,k}}. \quad (4\text{-}63)$$

And the total power transmitted by the relay, in strategy 3, is the aggregate of the transmitted powers of all the UEs:

$$P_{T,R} = \sum_{k=K+1}^{K_R+1} \frac{SINR_{R,min,k}\left(N_o B_k+\sum_{k'=K-i}^{K} P_{T,k'}\beta_k\cdot\frac{B_k}{B}\right)}{\beta_{R,k}}. \quad (4\text{-}64)$$

And by substituting (4-59) in (4-64):

$$P_{T,R} = \frac{\sum_{k=K+1}^{K_R+1}\frac{SINR_{R,min,k}\cdot N_o B_k}{\beta_{R,k}}+\sum_{k=K+1}^{K_R+1}\frac{SINR_{R,min,k}\cdot\beta_k\cdot B_k}{M\cdot\beta_{R,k}\cdot B\cdot c_{CSI}}\cdot\left(\sum_{k'=K-i}^{K} SINR_{min,k'}\left(\frac{P_N}{\beta_{k'}}+P_{T,BS}\right)\right)}{1-\sum_{k=K+1}^{K_R+1}\frac{SINR_{R,min,k}\cdot\beta_k\cdot B_k}{M\cdot\beta_{R,k}\cdot B\cdot c_{CSI}}\cdot\left(\sum_{k'=K-i}^{K} SINR_{min,k'}\cdot\left(\frac{\beta_{R,k'}}{\beta_{k'}}\right)\right)}. \quad (4\text{-}65)$$

Then by substituting (4-65) in (4-61) the total required transmit power at the BS to support all the UEs without dependence on the power of the relay are obtained as:

$$P_{T,BS} = \frac{\frac{1}{\left[1-\sum_{k=1}^{K}\frac{SINR_{min,k}}{M\cdot c_{CSI}}\right]}\cdot\left(\sum_{k=1}^{K}\frac{SINR_{min,k}\cdot P_N}{M\cdot\beta_k\cdot c_{CSI}}+\sum_{k=1}^{K}\frac{SINR_{min,k}\cdot\beta_{R,k}}{M\cdot\beta_k\cdot c_{CSI}}\cdot\frac{\sum_{k''=K+1}^{K_R+1}\frac{SINR_{R,min,k''}\cdot N_O B_{k''}}{\beta_{R,k''}}+\sum_{k''=K+1}^{K_R+1}\frac{SINR_{R,min,k''}\cdot\beta_{k''}\cdot B_{k''}}{M\cdot\beta_{R,k''}\cdot B\cdot c_{CSI}}\sum_{k'=K-i}^{K}\frac{SINR_{min,k''}\cdot P_N}{\beta_{k'}}}{\left[1-\sum_{k''=K+1}^{K_R+1}\frac{SINR_{R,min,k''}\cdot\beta_{k''}\cdot B_{k''}}{M\cdot\beta_{R,k''}\cdot B\cdot c_{CSI}}\left(\sum_{k'=K-i}^{K} SINR_{min,k'}\cdot\left(\frac{\beta_{R,k'}}{\beta_{k'}}\right)\right)\right]}\right)}{\left[1-\frac{1}{\left[1-\sum_{k=1}^{K}\frac{SINR_{min,k}}{M\cdot c_{CSI}}\right]}\cdot\sum_{k=1}^{K}\frac{SINR_{min,k}\cdot\beta_{R,k}}{M\cdot\beta_k\cdot c_{CSI}}\cdot\frac{\sum_{k''=K+1}^{K_R+1}\frac{SINR_{R,min,k''}\cdot\beta_{k''}\cdot B_{k''}}{M\cdot\beta_{R,k''}\cdot B\cdot c_{CSI}}\sum_{k'=K-i}^{K} SINR_{min,k'}}{\left[1-\sum_{k''=K+1}^{K_R+1}\frac{SINR_{R,min,k''}\cdot\beta_{k''}\cdot B_{k''}}{M\cdot\beta_{R,k''}\cdot B\cdot c_{CSI}}\left(\sum_{k'=K-i}^{K} SINR_{min,k'}\cdot\left(\frac{\beta_{R,k'}}{\beta_{k'}}\right)\right)\right]}\right]}. \quad (4\text{-}66)$$

Then, the transmit power at the relay $P_{T,R}$ can be obtained from (4-65) using the result from (4-66). Finally, the total power consumption can also be computed following (4-48).

The impact of the three strategies on the system's power consumption is currently being studied with simulations, and the results will be included in the upcoming BeGREEN D3.3.

5 Summary and Conclusion

BeGREEN aims at reducing the power consumption and enhancing the efficiency of current and future cellular networks. This document, BeGREEN D3.2 describes several methods for reducing mostly the PHY layer power consumption. The proposed enhancements and optimizations include using new CPU architectures and acceleration for flexibility and energy usage optimization in the DU and CU; RU on/off schemes and general RU operation optimisation; and enhanced transmit power and beamwidth allocation using ISAC and RIS. In addition, interference mitigation in relay-enhanced scenarios and its relevance to energy consumption are discussed.

The CU acceleration shows the transition from x86 to ARM architectures to improve energy efficiency and performance in O-RAN networks. The CU is crucial for managing control functions and user data transmission. Given the significantly higher traffic volume handled by the CU-UP, the porting process prioritizes the Arm architecture due to its superior power efficiency and processing capabilities. This transition involves adapting C++ code to ARM using a cross-compilation toolchain and deploying the CU components as Docker images for both x86 and Arm platforms. The integration with Kubernetes further facilitates seamless multi-architecture support and optimized resource utilization. Initial test results reveal that while x86 shows better energy efficiency in scenarios with a limited number of simultaneous users, the Arm architecture excels in large-scale, high-throughput environments, achieving greater overall power savings. Specifically, as the data throughput increases, the energy savings with Arm become more pronounced compared to x86, as this is attributed to Arm's efficient handling of parallelized tasks. A collaborative approach between project partners (ACC, ARM and PW) was pivotal in these advancements. Joint efforts in testing and validation have shown that the aggregated power reduction when using Arm covers a wide range of scenarios, making it a highly suitable choice for future scalable and energy-efficient network deployments. For some of the methods initial power measurement results are shared, which show significant power reduction compared to legacy network for some of the scenarios. Future work includes further developments implementations and calibration of the mentioned methods. In addition, different variant of these solution will be explored as well as they will be thoroughly tested. This will be described in BeGREEN D3.3, to include final evaluation and benchmarking of the implemented solutions. Specifically for DU acceleration, BeGREEN D3.3 will also include evaluation of the sphere decoder processing as well as introducing GPU implementation of both the LDPC and the sphere decoder.

In the RU domain, which is the major power consumer in cellular network, several power saving techniques were explored and simulated with promising results such as AI based Digital-Pre-Distortion and Envelope Tracking, which can lead to more than 60% RF PA energy savings. In addition, a novel technique called PA Blanking that powers-off the RU PA when there is no data in the downlink stream was developed and implemented on a commercial RU and demonstrated in the late June 2024 EuCNC in Antwerp showing a PA energy consumption reduction higher than 40%. The full results of the PA Blanking module performance will be described in the upcoming BeGREEN D3.3.

Further RU Energy saving research will be performed in the following areas: capacity and energy optimization and balancing; avoidance of RF power race between neighbouring cells while providing optimum performance, and power to bandwidth optimization to maximize performance.

It was proved and demonstrated that ISAC can be used to detect potential users, i.e., persons, and based on their distribution to optimally assign network resources. RIS has enormous potential for energy efficiency improvement. It can be used to extend the coverage of the RAN in urban areas, especially where the propagation is blocked due to buildings. It is extremely low power and can substitute an RU. BeGREEN tested an autoconfiguring RIS integration and functionality and its applicability towards energy efficiency improvement is being evaluated.

Finally, efficient interference management can strongly affect the energy efficiency of the network. A few strategies for interference management were proposed and evaluated to improve the network energy efficiency. The power consumption model for three different spectrum allocation strategies is developed. These differ in the way how the spectrum is shared and managed between the BS and the relay. The performance of each solution is being analysed and compared. The power consumption results and energy efficiency benefits will be included in BeGREEN D3.3.

This document presented a few different approaches for network energy efficiency improvement. The approaches presented here are the ones which are believed to bring significant improvement of the energy consumption of the network. The final results of these solutions will be presented in the upcoming BeGREEN D3.3.

6 Bibliography

[1] BeGREEN D3.1, "State-of-the-Art on PHY Mechanisms Energy Consumption and Specification of Efficiency Enhancement Solutions", January 2024. Available Online: https://www.sns-BeGREEN.com/deliverables

[2] Mavenir, "A Holistic Study of Power Comsumption and Energy Savings Strategies for Open vRAN Systems," White Paper, February 2023.

[3] 3GPP TS 38.141, "NR; Base Station (BS) conformance testing, Part 1: Conducted conformance testing (Relase 18)", September 2023.

[4] 3GPP TS 38.114, "NR; Base Station (BS) radio transmission and reception (Relase 18)", September 2023.

[5] BeGREEN Deliverable D5.1, "Use Case Identification and Demonstration Plan", December 2023, https://www.sns-begreen.com/deliverables

[6] 3GPP TS 38.401 "5G; NG-RAN; Architecture description (Release 15)"

[7] 3GPP 37.480 "5G; E1 general aspects and principles (Reelase 17)"

[8] B. Doorgakant, T. P. Fowdur, "Derivation and Comparative Analysis of End-To-End Power Models for 5G RAN Architectures," in https://www.researchsquare.com/article/rs-3235078/v1.

[9] O-RAN Alliance, O-RAN.WG1.NESUC-R003-v02.00 - O-RAN Work Group 1 (Use Cases and Overall Architecture) Network Energy Saving Use Cases, https://orandownloadsweb.azurewebsites.net/specifications

[10] BeGREEN, D4.1, "State-of-the-Art Review and Initial Definition of BeGREEN O-RAN Intelligent Plane", December 2023. [Online] Available: https://www.sns-begreen.com/deliverables

[11] R. Liu, M. Li, H. Luo, Q. Liu, and A. L. Swindlehurst, "Integrated sensing and communication with reconfigurable intelligent surfaces: Opportunities, applications, and future directions," IEEE Wireless Commun., vol. 30, no. 1, pp. 50-57, Feb. 2023.

[12] H. Luo, R. Liu, M. Li, and Q. Liu, "RIS-aided integrated sensing and communication: Joint beamforming and reflection design," IEEE Trans. Veh. Technol., vol. 72, no. 7, pp. 9626-9630, Jul. 2023.

[13] H. Luo, R. Liu, M. Li, Y. Liu, and Q. Liu, "Joint beamforming design for RIS-assisted integrated sensing and communication systems," IEEE Trans. Veh. Technol., vol. 71, no. 12, pp. 13393-13397, Dec. 2022.

[14] K. Kim, J. Kim and J. Joung, "A Survey on System Configurations of Integrated Sensing and Communication (ISAC) Systems," 2022 13th International Conference on Information and Communication Technology Convergence (ICTC), Jeju Island, Korea, Republic of, 2022.

[15] ETSI INDUSTRY SPECIFICATION GROUP (ISG) INTEGRATED SENSING AND COMMUNICATIONS (ISAC), https://www.etsi.org/committee/isac

[16] H. O. Ramp and E. R. Wingrove, "Principles of Pulse Compression," in IRE Transactions on Military Electronics, vol. MIL-5, no. 2, pp. 109-116, April 1961

[17] O-RAN.WG4.CUS.0-R003-v14.00: "O- O-RAN Working Group 4 (Open Fronthaul Interfaces WG) Title (Release 14)", November 2023.

[18] Pulse Compression, https://www.sciencedirect.com/topics/computer-science/pulse-compression

[19] V. Ninkovic, A. Valka, D. Dumic and D. Vukobratovic, "Deep Learning-Based Packet Detection and Carrier Frequency Offset Estimation in IEEE 802.11ah," in *IEEE Access*, vol. 9, pp. 99853-99865, 2021.

[20] K. Shen and W. Yu, "Fractional programming for communication systems−−Part I: Power control and beamforming," IEEE Transactions on Signal Processing, vol. 66, no. 10, pp. 2616–2630, 2018.

[21] Ericsson, "The Massive MIMO handbook-Technology Primer", in extended second edition, EN/LZT 4/28701-FGB1010987 Uen Rev B, 2023.

[22] T. L. Marzetta, E. G. Larsson, H. Yang, H. Q. Guo, Fundamentals of Massive MIMO, Cambridge University Press, 2016.

[23] E. Björnson, "A Basic Way to Quantify the Massive MIMO Gain", March, 2018. Online: https://ma-mimo.ellintech.se/2018/03/23/a-basic-way-to-quantify-the-massive-mimo-gain/.

[24] E. Dahlman, S. Parkvall, and J. Skold, 5G NR: The Next Generation Wireless Access Technology, Academic Press (Elsevier), 2018.

[25] 3GPP TS 38.306 v17.1.0, "NR; User Equipment (UE) radio access capabilities (Release 17)", June, 2022.

[26] "5G TDD Synchronisation. Guidelines and Recommendations for the Coexistence of TDD Networks in the 3.5 GHz Range", GSMA, White Paper, Apr. 2020, Accessed: Apr. 6, 2023. Available: https://www.gsma.com/spectrum/wp-

content/uploads/2020/04/3.5-GHz-5G-TDD-Synchronisation.pdf

Appendix: Building Energy Efficient 5G Infrastructure using arm Neoverse Systems

Arm Neoverse family of CPUs are designed for Datacentre, Edge and HPC markets. Neoverse CPUs deliver the speed, energy efficiency, performance per watt and performance per dollar needed to build 5G/6G Infrastructure. Neoverse family consists of V-Series, N-Series and E-Series CPUs to cater different infrastructure markets based on the workload.

- Neoverse V-series delivers maximum per core performance for systems running demanding compute and memory intensive applications.
- Neoverse N-series provides a balanced CPU design optimized to deliver both performance per watt and performance per dollar.
- Neoverse E-series supports high data throughput with minimal power consumption.

Several Neoverse class products are available in the market. To mention a few, one can name: Amazon Graviton 2/3/4 instances, Nvidia Grace CPU and Ampere Altra CPU based COTS Servers. These Neoverse CPU based COTS servers are built by several OEMs and ODMs. Some of them to mention are HPE, Super Micro, Gigabyte, Foxconn and Wiwynn. These Neoverse family based HW are best-in-class and are well suited in building energy efficient vRAN/Open RAN Infrastructure.

Guideline to get best performance on Arm Neoverse CPU based Servers:

To get the best performance, it is important to check and set some of the System configurations available through BIOS and GRUB Settings. Accurate BIOS settings must be obtained through OEM/ODM Vendors. Following are general high-level guidelines to get the best performance.

1. Update the latest BIOS and BMC Firmware from OEM/ODM Vendor.
2. Power/Performance Settings:
    a. For Max performance: Disable ACPI, CPPC and LPI in BIOS. On Ampere Altra based Servers, these settings will allow the cores to run at maximum CPU frequency of 3GHz. LPI and CPPC are like Intel C States and P States. LPI is Low Power Idle state and CPPC is Collaborative Processor Performance Control. CPPC is a mechanism to manage the performance of CPU cores using DVFS (Dynamic Voltage and Frequency Scaling) technique.
    b. For Max power savings: Enable CPPC/LPI in BIOS. Enabling CPPC will allow the Operating System to control the CPU clock speed. CPU frequency can be managed by Operating System via Governor Policy. Set the Scaling governor settings to “ondemand” in “/sys/devices/system/cpu/cpufreq/policy0/scaling_governor”. On Ampere Altra based Servers, Minimum CPU frequency is 1 GHz. The minimum and maximum frequency can be set at per core level in “/sys/devices/system/cpu/cpu<id>/cpufreq/cpuinfo_min_freq” and “/sys/devices/system/cpu/cpu<id>/cpufreq/cpuinfo_max_freq”. Current CPU frequency can be read from “/sys/devices/system/cpu/cpu<id>/cpufreq/cpuinfo_cur_freq”.

    LPI has 4 idle states. 0 is the highest idle power state and 3 is the lowest idle power state. For Max power savings, cores must idle in LPI-3 state. Arm architecture provides two instructions for reducing power consumption when the processor is idle - Wait for Interrupt (WFI) and Wait for Event (WFE). The key difference between WFI and WFE is that WFI simply puts the processor to sleep until the next interrupt wakes it up, while WFE allows the processor to sleep until a specific event occurs. Refer Neoverse Technical Reference Manual to learn about WFI and WFE instructions. Following links are also useful:

https://developer.arm.com/documentation/ka001283/latest/

https://s-o-c.org/what-is-the-difference-between-wfi-and-wfe-arm-cortex/

c. For balance of performance and power savings: Disable CPPC to run CPU at max frequency. On Ampere altra, max frequency is 3 GHz. Choose the desired LPI states from 0 to 4 based on the idle conditions of core.

For more details about Power Management, check the following links:

Refer the Neoverse N, V, E Series Technical reference manuals - https://developer.arm.com/documentation/#numberOfResults=48&&cf-navigationhierarchiesproducts=%20IP%20Products,Processors,Neoverse

https://neoverse-reference-design.docs.arm.com/en/latest/power-management/index.html

For Ampere altra based Servers, Check Power Management section in Ampere documentation – https://amperecomputing.com/assets/Altra_Max_UM_v1_15_20230617_b380fdbc19.pdf

3. To get deterministic and maximum data plane performance, it is recommended to isolate CPU cores from Linux Scheduler, disable all the maskable interrupts at core level and use huge pages to avoid TLB cache miss. These can be achieved through the following GRUB settings. The following is an example taken from 80 Core Ampere Altra based Server running Ubuntu OS.

   File: /etc/default/grub

   GRUB_CMDLINE_LINUX_DEFAULT="default_hugepagesz=1G hugepagesz=1G hugepages=64 tsc=reliable iommu.passthrough=1 audit=0 isolcpus=1-69 nohz_full=1-69 irqaffinity=0 rcu_nocbs=1-69 rcu_nocb_poll nosoftlockup cma=256m kpti=off"

   In the above config, 1-69 CPU cores are isolated from Linux Scheduler, 1G Huge pages are reserved and they are used to run CU/DU functionality. All the interrupts are redirected to Core 0. IOMMU is disabled.

4. Use the latest GCC/LLVM compiler. GCC options to choose: -march=native -mtune=native. More details can be obtained from the link below:

   https://community.arm.com/arm-community-blogs/b/tools-software-ides-blog/posts/compiler-flags-across-architectures-march-mtune-and-mcpu

5. Integrate arm acceleration libraries in CU/DU stack. The following libraries boost the performance of CU and DU stack.

   DU - Arm RAN Acceleration Library (RAL) for PHY layer.

   https://community.arm.com/arm-community-blogs/b/infrastructure-solutions-blog/posts/arm-ral-is-now-open-source

   CU - IPSec MB library for PDCP Crypto - https://gitlab.arm.com/arm-reference-solutions/ipsec-mb

Guidelines to measure CPU and System Power:

CPU utilization of CU/DU Network functions vary based on number of Cells, MIMO layers, number of active UEs, total DL and UL bandwidth. Performance per watt can be measured for various traffic load conditions as mentioned below:

1. Idle System Power when CU/DU stack is not running.

2. System Power while CU/DU stack is up and running. UEs are not attached.
3. System Power while CU/DU stack is up and running. Power can be measured with variable number of Cells/UEs with active DL and UL traffic on each UE.
4. Idle CPU Power when CU/DU stack is not running.
5. CPU Power while CU/DU stack is up and running. UEs are not attached.
6. CPU Power while CU/DU stack is up and running. Power can be measured with variable number of Cells/UEs with active DL and UL traffic on each UE.

Performance per watt at CPU level and System level can be determined by “Total UL+DL throughput in bits per second / Power in Watts”. System and CPU power can be measured using tools like Redfish or ipmitool. Check OEM/ODM Vendor documentation to measure System and CPU power.