Deliverable 3.1

State-of-the-Art on PHY Mechanisms Energy Consumption and Specification of Efficiency Enhancement Solutions

January 2024

Co-funded by
the European Union

6GSNS

Contractual Date of Delivery:	**September 30, 2023**
Actual Date of Delivery:	**January 28, 2023**
Editor(s):	**Vladica Sark (IHP)**
Author(s)/Contributor(s):	**Vladica Sark, Jesús Gutiérrez (IHP)** **Jordi Pérez-Romero, Oriol Sallent, Juan Sánchez-González,** **Anna Umbert (UPC)** **German Castellanos, Revaz Berozashvili, Simon Pryor (ACC)** **Govindarajan Mohandoss (ARM)** **Josep Xavier Salvat, Jose A. Ayala (NEC)** **Ory Eger (PW)** **Israel Koffman (REL)** **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	15.03.2023	Vladica Sark (IHP)	Initial version and initial ToC
	27.07.2023	Anna Umbert (UPC)	Chapter 4.3
	27.07.2023	Esteban Municio (I2CAT)	Initial draft for 4.1.5
0.2	01.08.2023	Eli Shasha, Israel Koffman, Baruch Globen (RunEL)	Chapters 3.1 and 4.2
0.3	02.08.2023	German Castellanos (ACC)	Contributions to Chapter 2 SotA of HW acceleration.
	11.09.2023	Jose Ayala Romero (NEC)	Contribution to 4.1.4, 4.1.5 and review of 4.3
0.4	30.10.2023	Govindarajan Mohandoss (ARM)	Chapter 2.2.2
0.5	02.11.2023	Ory Eger (PW)	Contributions to Chapter 2.2
0.6	15.11.2023	Vladica Sark, Jesús Gutiérrez (IHP)	Contributions to Chapter 4.1.1 on ISAC SotA
0.7	20.11.2023	Vladica Sark, Jesús Gutiérrez (IHP)	Contributions to Chapter 4.1.2 on ISAC
0.8	22.11.2023	Josep Xavier Salvat, Jose Ayala Romero (NEC)	Revisions to Chapter 4.1.4 on RIS
0.9	08.12.2023	Vladica Sark, Jesús Gutiérrez (IHP)	Revisions to Chapter 4
0.91	16.12.2023	Ory Eger (PW)	Revised Chapter 2.2.2
0.92	16.12.2023	Josep Xavier Salvat (NEC)	Revised Chapter 4.1.3
0.93	19.12.2023	Anna Umbert (UPC)	Final revision of Chapter 4
0.95	17.01.2024	Mir Ghoraishi (GIGASYS), Jesús Gutiérrez (IHP)	Final Review and Proof Reading
1.00	28.01.2024	Simon Pryor (ACC)	Submission to the EC

Table of Contents

List of Acronyms 7
Executive Summary 10
1 Introduction 11
2 Hardware Acceleration for Flexibility and Energy Usage Optimisation 12
2.1 Legacy and state-of-the-art L1/L3 energy efficient processing 12
2.2 Proposed improvement strategies 16
2.2.1 Using ARM architecture for L1/L3 processing 16
2.2.2 Power optimisation metrics and HW capabilities of the ARM core 19
2.2.3 Acceleration using GPU for L1 computationally extensive functions 20
2.3 Key Performance Indicators (KPIs) for energy efficiency 25
2.4 Takeaways for BeGREEN Reference Architecture 26
3 O-RAN Based Strategies for RU Power Consumption Optimisation 27
3.1 Legacy and state-of-the-art RU Power amplifiers optimisation 27
3.2 Near-RT and non-RT Applications for RU Management 30
4 PHY Layer Enhancements for Radio Network Energy Efficiency 32
4.1 ISAC, RIS and their contribution to network energy efficiency 32
4.1.1 ISAC state-of-the-art and developments for energy efficiency 32
4.1.2 BeGREEN approach on ISAC 36
4.1.3 ISAC aided strategies for improvement of energy efficiency 37
4.1.4 Self-configuring RIS 41
4.1.5 PHY layer and RIS/ISAC aided network level energy consumption optimisation 42
4.2 RU power optimisation 43
4.2.1 State-of-the-art status in RU power optimisation techniques 43
4.2.2 Problem statement 44
4.2.3 Proposed improvement strategies in RU power optimisation 46
4.3 Interference management in relay-enhanced scenarios 51
4.3.1 State-of-the-art and problem statement 51
4.3.2 Interference management strategies 54
4.4 Telemetry subsystem 58
5 Summary and Conclusions 60
6 Bibliography 61

List of Figures

Figure 2-1 Look-Aside and In-Line architectures according to Nokia [10] 13
Figure 2-2 Power consumption comparison of In-line and Look-Aside approaches 14
Figure 2-3 Open-RAN acceleration techniques needed by organisations 14
Figure 2-4 Demanded function needed for hardware acceleration offloading 15
Figure 2-5 O-RAN RU/DU Split 7.2 17
Figure 2-6 Legacy linear PUSCH Receiver using x86 architecture 18
Figure 2-7 Linear PUSCH Receiver using ARM architecture 18
Figure 2-8 CU framework implementation, a: traditional; b: with XDP. 19
Figure 2-9 Linear PUSCH receiver using ARM plus GPU architecture 21
Figure 2-10 Sphere decoder PUSCH receiver using ARM plus GPU architecture 21
Figure 2-11 Demapping example for QPSK 22
Figure 2-12 Message passing algorithm 25
Figure 3-1 Transition diagram for power-state node 28
Figure 3-2 O-RAN cell on/off flow diagram using non-RT RIC 31
Figure 3-3 O-RAN cell on/off flow diagram using Near-RT RIC 31
Figure 4-1 a. Deviceless sensing; and b. Device based sensing 38
Figure 4-2 Comparison of single beam and multi beam coverage 40
Figure 4-3 Reference diagram of a hybrid reconfigurable intelligent surface 42
Figure 4-4 Performance of Wiener PA 45
Figure 4-5 Performance of Wiener PA with memory 45
Figure 4-6 Performance of saturated PA 46
Figure 4-7 Performance of perfect PA 46
Figure 4-8 DPD training setup 47
Figure 4-9 State-target training samples versus perfect and trained model 48
Figure 4-10 Signal spectrum for Asat = 5 of perfect reference signal compared with plain PA output and DPD output 49
Figure 4-11 Constellation points for Asat=5 of reference, plain and DPD 49
Figure 4-12 Signal spectrum for $A_{sat} = 3$ of perfect reference signal compared with plain PA output and DPD output 49
Figure 4-13 Constellation points for $A_{sat}=3$ of reference, plain and DPD 50
Figure 4-14 RF signal and power envelope 50
Figure 4-15 Modified training setup to support power envelope 51
Figure 4-16 Relay operation modes 52
Figure 4-17 Interferences in in-band relay mode 52
Figure 4-18 Relay transmission and reception with TDM 52
Figure 4-19 Interferences in out-band relay mode with FDD or fixed TDD 53
Figure 4-20 Interferences in out-band relay mode with dynamic TDD configuration 54

List of Tables

Table 2-1 5G NR 3GPP Standard MCS 23
Table 2-2 Scenarios for DU Acceleration Performance Evaluated 25
Table 3-1 Measurements for Energy, Power and Environmental Parameters 28
Table 3-2 O-RU NES Mechanisms 29
Table 3-3 Energy Savings Mechanism Mapping to M-Plane and C-Plane Commands [36] 29

List of Acronyms

3GPP	3rd Generation Partnership Project
5GC	5G Core
5GNR	5G New Radio
ABS	Almost Blank Subframes
AI	Artificial Intelligence
AoA	Angle-of-Arrival
AP	Access Point
API	Application Programming Interface
ARM	Advanced RISC Machine
B5G	Beyond 5G
BPF	Berkeley Packet Filter
CAPEX	CApital EXpenditures
CoMP	Coordinated Multi-Point
COTS	Commercial Off-The-Shelf
CP	Control Plane
CPU	Central Processing Unit
CSI	Channel State Information
CU	Centralised Unit
DMRS	DeModulation Reference Signal
DoW	Description of Work
DPC	Dynamic Power Control
DU	Distributed Unit
DVFS	Dynamic Voltage and Frequency Scaling
E2E	End-to-End
E2SM	E2 Service Models
eICIC	enhanced ICIC
EIRP	Effective Isotropic Radiated Power
ERLLC	Extremely Reliable and Low Latency Communications
EVM	Error Vector Magnitude
FAPI	Functional Application Platform Interface
FEC	Forward Error Correction
FeICIC	further enhanced ICIC
FFR	Fractional Frequency Reuse
FFT	Fast Fourier Transform
FMCW	Frequency Modulated Continuous Wave
FPGA	Field-Programmable Gate Array
GPP	General-Purpose Processor
GPU	Graphics Processing Unit
GTP	GPRS Tunnelling Protocol
HRIS	Hybrid Reconfigurable Intelligent Surfaces
IA	Interference Alignment
IP	Internet Protocol
ISAC	Integrated Sensing and Communication
ISM	Industrial Scientific and Medical
KPI	Key Performance Indicator
L2	Layer 2

LLR	Log Likelihood Ratio
LNS	Linux Network Stack
LoS	Line-of-Sight
LTE	Long Term Evolution
MAC	Medium Access Control
MARISA	Metasurface Absorption and Reflection for Intelligent Surface Applications
MP	Management Plane
mMIMO	massive Multiple-Input Multiple-Output
MMSE	Minimum Mean Square Error
ML	Machine Learning
mmWave	Millimetre Wave
MNO	Mobile Network Operator
NDT	Network Digital Twin
near-RT	near Real-Time
NFV	Network Function Virtualisation
NIC	Network Interface Card
NLoS	Non-Line-of-Sight
NOMA	Non-Orthogonal Multiple Access
non-RT	non Real-Time
OFH	Open Fronthaul
O-RAN	Open RAN
PAPR	Peak-to-Average Power Ratio
PCIe	Peripheral Component Interconnect Express
PDCP	Packet Data Convergence Protocol
PHY	Physical layer
PMN	Perceptive Mobile Network
PoC	Proof-of-Concept
PUSCH	Physical Downlink Shared Channel
QAM	Quadrature Amplitude Modulation
RAN	Radio Access Network
RB	Resource Block
RIC	Radio Interface Controller
RIS	Reconfigurable Intelligent Surface
RISC	Reduced Instruction Set Computing
RLC	Radio Link Control
RLF	Radio Link Failure
RRM	Radio Resource Management
RSS	Received Signal Strength
RSSI	Received Signal Strength Indicator
RU	Radio Unit
SDR	Software Defined Radio
SDK	Software Development Kit
SFR	Soft Frequency Reuse
SIC	Successive Interference Cancellation
SIMD	Single Instruction Multiple Data
SLA	Service Level Assurance
SMO	Service Management Orchestator
SNR	Signal-to-Noise Ratio

TSG	Technical Specification Group
UPF	User Plane Function
uRLLC	ultra-Reliable and Low Latency Communications
vRAN	virtualised Radio Access Network
WFE	Wait for Event
WFI	Wait for Interrupt
WP	Work Package
XDP	eXpress Data Path

Executive Summary

BeGREEN proposes an evolved radio access network (RAN) that aims to improve the energy efficiency of this segment of beyond 5G (B5G) and 6G networks, targeting the European vision for a green digital society and economy by 2030. The strategies proposed in the project seek the maximisation of the performance demanded by current and future networks, considering power consumption as a factor.

This document, BeGREEN D3.1, titled "State-of-the-Art on PHY Mechanisms Energy Consumption and Specification of Efficiency Enhancement Solutions", presents the strategies and functionalities that may be deployed at lower layers (L1 to L3) targeting energy efficiency enhancements to B5G systems.

Firstly, the suitability of the widely used software-based implementations for RAN operations in general purpose processors (GPP), given their virtualisation benefits, is analysed and options such as opting for different architectures, e.g., ARM instead of x86 or using Graphics Processing Units (GPUs) are studied.

Secondly, O-RAN-based strategies that allow an optimal operation of the radio units (RUs) are presented. These include power optimisation in the power amplifiers (PAs) and the applications and developments that can be deployed at the RAN Intelligent Controllers (RICs) to manage the RUs.

Finally, this document includes various BeGREEN elements that are being proposed as extensions to the envisioned 3GPP/O-RAN network architectural framework at the physical (PHY) later. Examples are the implementation of integrated sensing and communication (ISAC) strategies, the use of reconfigurable intelligent surfaces (RISs), PHY-layer strategies for the optimisation of the RUs and, finally, the interference management in relay-enhanced scenarios. These developments share BeGREEN goals on energy efficiency improvements and reducing energy consumption.

For every development, enhancement, or optimisation strategy that BeGREEN proposes to reduce power consumption of the RAN, a State-of-the-Art (SotA) section is presented. Then, a brief description of the proposed enhancements and strategies are included.

BeGREEN D3.1 serves as the reference to BeGREEN D3.2 that is expected to provide the initial results and evaluations stemming from the different developments.

1 Introduction

Next generation networks, beyond 5G (B5G) and 6G, are expected to introduce architectural transformations that have ranged from an inflexible and monolithic system to a flexible, agile, and disaggregated architecture to support service heterogeneity, coordination among multiple technologies, and rapid on-demand deployments. B5G/6G networks will play an ambitious role towards sustainability, to reduce its footprint on energy, resources, and emissions and to improve sustainability in other parts of society and industry.

The radio access network (RAN) consumes 73% of the total energy of a 5G system [1], making it an ideal target for optimisation. By using short term optimisation, leveraging Machine Learning (ML) or Artificial Intelligence (AI), together with closed-loop control capabilities offered by Open RAN (O-RAN), energy efficiency can be increased [2]. Strategies like switching off underutilised RUs and/or de-activating antenna elements in MIMO setups can lead to improving energy efficiency by up to 22% [3] and 18% [4], respectively. The common conclusions stemming from these works are that it is key to balance Quality of Service (QoS) and energy consumption.

The dominant contributors to power consumption are Power Amplifiers (PAs), baseband process modules, digital intermediate frequency (DIF) and transceivers. By using new generations of chipsets, and further by smart management of the network functions using AI methods, it is estimated to achieve between 30-70% in energy saving. Of course, power consumption stemming from the use of ML/AI methods needs to be accounted for [5].

BeGREEN aims at the design and integration of innovative solutions for improving the energy efficiency in the RAN and reducing power consumption beyond 3GPP Release 18. The Technical Specification Group (TSG) Radio Access Network (RAN) groups, in the ongoing Release 18, have performed a study on network energy consumption model [6], to identify and study network energy savings techniques and reduce energy consumption in mobile networks.

Formal activities to develop the new 3GPP work item on AI/ML support in 5G RAN began at the start of 3GPP Rel-18. 3GPP has recently issued a ground-breaking technical report that lays out which information would need to be exchanged among nodes and functions of a 5G RAN to support AI/ML based optimisation. The report focuses on three reference use cases: load balancing, mobility optimisation and network energy saving.

Energy savings can derive from the flexibility and scalability that is an inherent feature of disaggregated networks. This disaggregation entails bringing network function or functions that were previously highly integrated in dedicated hardware, thus being efficient for that particular implementation. The deployment of such functions in more general-purpose computing environments can lead to less energy efficient implementations.

This deliverable presents the physical layer (PHY) mechanisms proposed by the BeGREEN project to reduce the energy consumption, together with the specification of efficiency enhancement solutions for the several parts of a disaggregated RAN. Additional technologies like integrated sensing and communication (ISAC), reconfigurable intelligent surfaces (RISs) and relays complement the existing architectural framework in search of reducing the energy consumption and improving energy efficiency.

2 Hardware Acceleration for Flexibility and Energy Usage Optimisation

One of the main challenges of O-RAN adoption is to achieve the desired performance to support the ever-increasing demands of RAN processing for beyond 5G (B5G) and 6G networks. This is coming at the cost of boosting the energy consumption, which O-RAN aims to reduce by assessing the amount of energy savings that can be obtained using potential energy efficiency techniques. In this context, the exclusive and abusive use of general-purpose processors (GPPs) may bring along excessive power consumption that could be relieved with the use of specialised hardware for certain PHY layer tasks.

Hardware acceleration involves using specialised hardware devices to speed up specific computing tasks, especially those that are repetitive and computing/resource intensive. This helps improve overall system performance and efficiency by allowing dedicated hardware to handle certain tasks, freeing up GPPs like Central Processing Units (CPUs) for other operations. In the context of O-RAN implementation, a function called layer 1 (L1) forward error correction (FEC) demands a lot of processing power. While it is possible to implement L1 FEC with a standard CPU and optimised software, in high-demand systems this would use up a significant amount of computing resources. Hence, there is a strong preference to offload this computation to a hardware accelerator, similar to historical advancements like floating-point accelerators that were designed to ease CPU workloads. This technology is crucial for achieving optimal O-RAN performance, as it allows the CPU to focus on more intricate RAN operations, ultimately enhancing overall performance. In the following sections, a more detailed explanation from the viewpoint of the BeGREEN Project is given across components such as Distributed Unit (DU), Centralised Unit (CU) and RIC, main focus being on using hardware accelerators for offloading computationally complex tasks and improving energy efficiency of the overall system.

2.1 Legacy and state-of-the-art L1/L3 energy efficient processing

Some of the most computational extensive gNB processing reside is L1 within the DU. It covers a set of estimation, signal processing and FEC algorithms. Example of such algorithms are defined in detail in Section 2.2.1. Due to the cloudification of RAN processing, it is preferable to implement L1 over GPPs and Commercial Off-The-Shelf (COTS) servers as it is done for the rest of the DU. However, in many cases, GPPs and other computing resources available in today's state-of-the-art servers still struggle with the required computational loads. Another issue that arises when using GPPs is that, even if their computation resources can handle the required processing load, they can introduce significant latency. Latency is becoming a very important factor in future cellular communication usages, for example with ultra-Reliable and Low Latency Communications (uRLLC) in 5G and Extremely Reliable and Low Latency Communications (ERLLC) in 6G. Additionally, GPPs are not optimised for signal/data processing tasks, which makes them nonoptimal candidate in terms of energy efficiency.

Hardware acceleration can be perceived in a particular way depending on the split used in O-RAN, its components or by the layer they compromise. Mobile Network Operators (MNOs) could take advantage of split 7.2 and leave the DU closer the site due its need to perform radio processing functions, including the baseband processing utilising virtualised DUs. In the context of 5G, the responsibility for real-time computations of these functions lies within the DU. These include tasks like Fast Fourier Transform (FFT), digital filtering, channel coding, and L1 and L2 processing. To achieve optimal performance at scale, these intensive computations cannot be performed on the CPU alone, necessitating the need for hardware acceleration, which enables the efficient processing of radio signals, ensuring they meet the real-time requirements of Radiation Power Factor. Additionally, it extends to service agility and network programmability, which requires that hardware acceleration to be software-defined. This approach allows MNOs to deploy readily available commercial hardware and install DU software, with subsequent system enhancements and alignment with 5G technology evolution achieved through software upgrades.

Several ways to improve HW acceleration for the DU are mainly focused on parallelised processing, as described below [7]:

- *Vectorisation with Single Instruction Multiple Data (SIMD)*: Utilising CPUs with vectorisation capabilities, such as 256-bit registers, to work on multiple sets of data simultaneously, resulting in potential speedup.
- *Multi-threading for parallel processing*: Employing multiple processor cores in parallel to process data using SIMD, potentially achieving significant speedup, especially with CPUs supporting a high number of cores.
- *Distributed processing through message-passing:* Distributing calculations across multiple hosts, enabling parallel processing and overall speed improvement.
- *Hardware acceleration with specialised co-processors:* Leveraging GP-GPUs, FPGAs, ASICs, or custom DSPs for co-processing, with GPUs being particularly advantageous for network programmability due to their many-core architecture and throughput optimisation, compared to CPUs designed for serial operations and latency optimisation.

Recently, considerable attention has been devoted to hardware accelerators, such Field-Programmable Gate Arrays (FPGAs) and Graphics Processing Units (GPUs), with the purpose of enhancing real-time processing capabilities for the fundamental layers of the radio baseband in the context of 5G technology. Ericsson and Nokia have been investigating the feasibility of leveraging GPU-based acceleration techniques to address specific virtualised Radio Access Network (vRAN) workloads, particularly those related to 5G massive Multiple-Input Multiple-Output (mMIMO) systems and Artificial Intelligence (AI) applications [8].

In general, implementation of the CU on server platforms is possible even without hardware acceleration. However, utilising hardware acceleration can potentially reduce the cost of the solution by offloading CPU cores. In scenarios where numerous DU and RU deployments exist, executing L1 processing on virtualised server platforms without hardware acceleration would demand a substantial number of CPU cores. Integrating hardware acceleration allows for the offloading of CPU cores, resulting in enhanced performance, reduced processor requirements, and significant power consumption reduction [9]. Open RAN hardware acceleration can take the form of either In-Line or Look-Aside implementation using FPGAs, ASICs, GPUs, or a combination of them. The hardware acceleration options within the ORAN architecture can be categorised into two main approaches: Look-Aside and In-Line architectures as shown in Figure 2-1 [10].

In the Look-Aside architecture, the CPU assumes the role of the master for processing L1 functions. However, specific key functions such as FEC, are offloaded to a hardware accelerator that can either be a separate Peripheral Component Interconnect Express (PCIe) card or integrated on the same die as the CPU. While this approach allows for offloading certain functions to an accelerator, the CPU still handles many real-time computations of L1, which may not be energy and computationally efficient. Additionally, the Look-Aside architecture allows the CPU to handle L2 and L3 processing tasks more effectively.

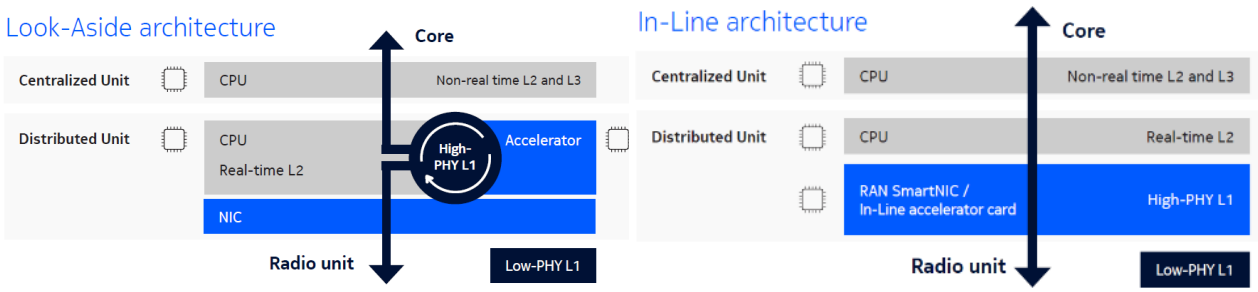


Figure 2-1 Look-Aside and In-Line architectures according to Nokia [10]

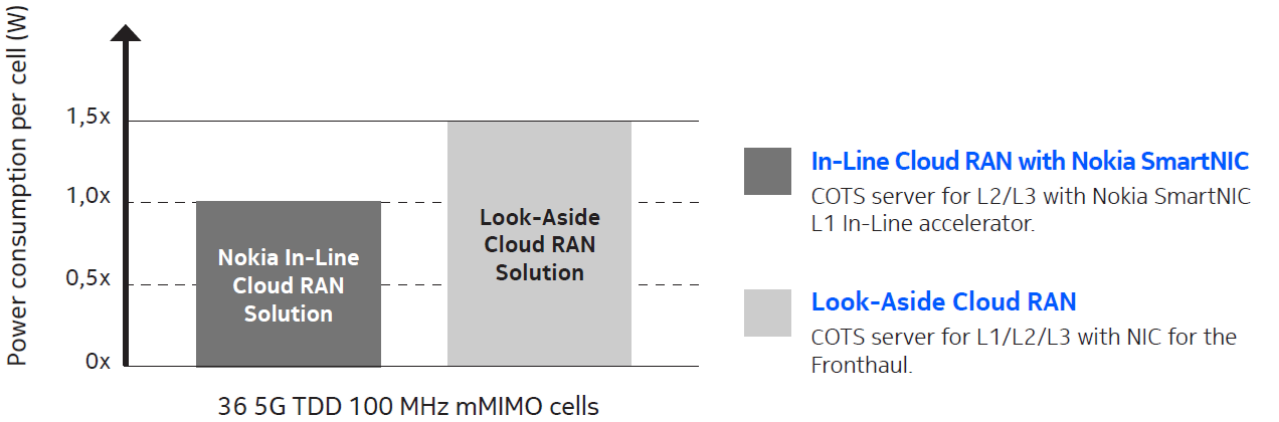


Figure 2-2 Power consumption comparison of In-line and Look-Aside approaches

Purpose-built silicon technology is widely acknowledged for its superior performance and energy efficiency, particularly for demanding workloads such as L1 processing. In the context of In-Line solutions, ARM[1]-based silicon technology is commonly utilised, not only in classic RAN networks but also in various network types, including cloud data centres operated by webscale companies. A recent benchmarking exercise was conducted to evaluate the performance of a high-capacity Cloud RAN configuration, utilising the available technology in 2023 and incorporating future roadmap advancements. The findings shown in Figure 2-2 revealed that the In-Line solution exhibited significantly lower power consumption per cell compared to the Look-Aside solution. Consequently, the In-Line approach offers a more cost-effective per-cell solution, in addition to its capacity and efficiency advantages.

In the "Accelerating Open RAN Platforms Operator Survey" [11] it is shown that the majority of mobile network businesses are interested in FPGA (52%), ACIS (45%) and GPU (45%) acceleration technologies for their hardware solutions (See Figure 2-3). Additionally, Figure 2-4 depicts what other functions they would want hardware accelerators to offload: 63% said they would want accelerators to offload Packet Data Convergence Protocol (PDCP) and security processing. Complete L1 processing and 5G core User Plane Function (UPF) offload is wanted by approximately 40%.

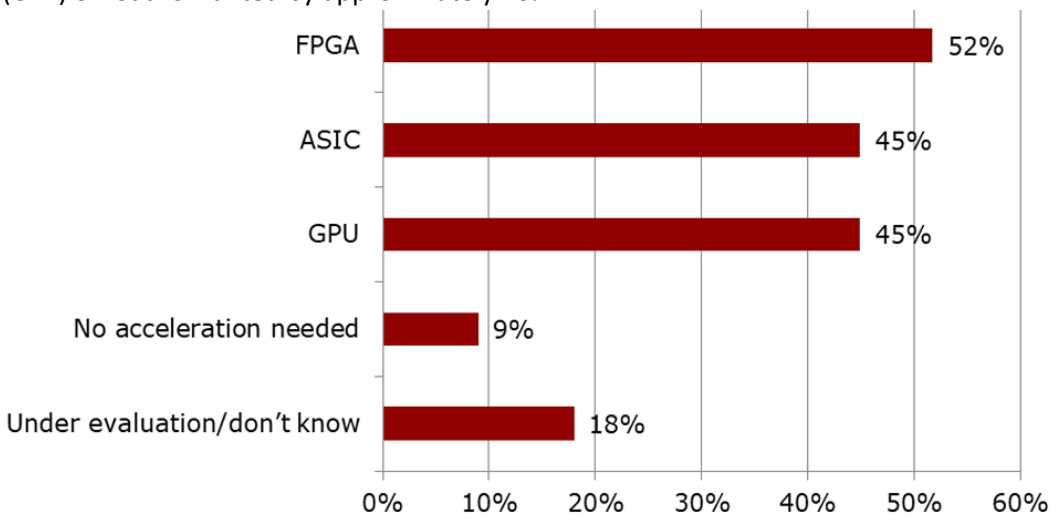


Figure 2-3 Open-RAN acceleration techniques needed by organisations

[1] www.arm.com

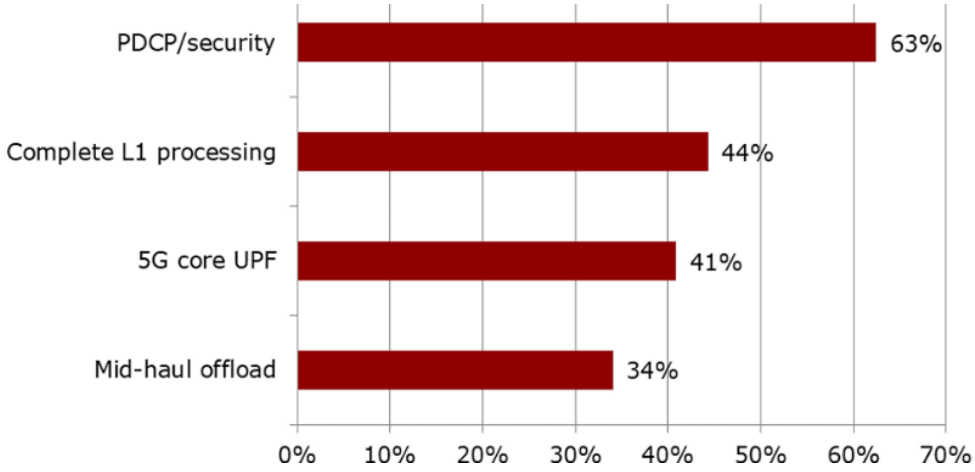


Figure 2-4 Demanded function needed for hardware acceleration offloading

The significance of GPUs extends beyond vRAN signal processing, particularly in the context of 5G and 6G systems where the integration of big data and wireless communication calls for the utilisation of AI/ML techniques to enhance network performance. GPUs have become the standard choice for model training and inference due to their capabilities in handling these tasks effectively. While a GPU-based hardware platform can support training, inference, and signal processing, the software aspect is equally critical. Programming NVIDIA[2] GPUs is achieved through CUDA, a widely adopted parallel programming framework that has achieved huge commercial success. Additionally, a comprehensive collection of GPU libraries, such as the NVIDIA RAPIDS software suite, enables the development of data analytics pipelines. These pipelines can be leveraged as services by the Service Management Orchestrator (SMO)/Non-RT RIC to update and fine-tune inference models executed under the Near-RT RIC [12].

Specifically, for L1, legacy gNBs solutions as well as in some of the current gNB solutions in the industry, the processing is run on dedicated HW (either silicon or FPGA). Usually, in these types of solutions, some of the algorithms are implemented in SW, and some in HW. An example of such a chip is Qualcomm® X100 5G RAN Accelerator Card [13].

In recent years, alongside the progress in O-RAN and in RAN cloudification, more and more L1 solutions tend towards SW implemented on COTS servers. Usually, these servers are based on Intel x86 chips. Intel has generated the FlexRAN™ reference Architecture [14] for wireless access which offers 4G and 5G L1 SW libraries and can be used as a starting point for L1 development on x86 architecture [15]. For example, one of the versions of the Mavenir L1 solution is FlexRAN™ based[3]. ARM also has a RAN Acceleration Library [16] which includes FEC processing and other L1 functions.

The performances and the resources of COTS servers are limited. They become easily exhausted with the ever-growing demand for high throughput, an increase in processing dimensions such as number of antennas, allocation sizes, number of layers, etc. This is the main reason for offloading the more computationally extensive algorithms such as FEC to accelerators. Two examples of such accelerators are:

- Intel® vRAN Dedicated Accelerator ACC100 [17], which includes Turbo Encoder/Decoder for 4G and Low-Density Parity Check (LDPC) Encoder/Decoder for 5G. It is used in addition to the FlexRAN SW for deployments where SW FEC processing is not feasible.
- NXP Layerscape® Access LA12xx programmable baseband processor [18], which includes LDPC Encoder/Decoder for 5G data channels and Polar Decoder/Encoder for 5G control channels.

[2] https://www.nvidia.com/

[3] https://www.mavenir.com/press-releases/industry-leaders-intend-to-collaborate-on-new-programmable-multi-generational-framework-that-extends-beyond-5g-open-ran/

Another type of L1 acceleration, that has been recently suggested, is the use of GPUs. NVIDIA has developed a full implementation of the L1 on a GPU in their NVIDIA Aerial Software Development Kit (SDK) [19]. In addition, investigation has been performed on implementing some of the PHY function on a GPU. One such example is implementation of an LDPC decoder on NVIDIA hardware [20].

Additionally, Dell and Marvel have developed an Open RAN Accelerator Card for L1/L2 [21] that, compared to FEC accelerators, does not focus on the Look-Aside approach, which increases the back-and-forth communication between the CPU and the FEC accelerator. Instead, it processes all L1 computations, freeing up valuable server CPU cores and eliminating the need for a fronthaul network interface card (NIC). As a result, all L2 computations are managed in a single CPU, reducing the number of them, helping to cut overall power consumption and overall costs.

Regarding power consumption, usually dedicated ASIC or FPGA will be the most energy-efficient, as they have much less overhead and control than that of CPUs or GPUs. However, they have many drawbacks, e.g. they are not flexible, not scalable and often under-utilised. As for comparing GPU and CPU, there can be many cases where GPU, when performing arithmetic operations that suits its architecture, can be more energy efficient than CPUs. An example that illustrates this can be found in [22], where energy efficiency of CPU, GPU and FPGA implementations for computer vision algorithms is compared.

2.2 Proposed improvement strategies

In this section we will go over the strategies that will be developed and proposed in BeGREEN to be used for accelerating L1 and L3 processing to enable energy consumption and compute resources reduction. The development of these strategies takes into consideration the architectural framework developed in BeGREEN [1]. First, improving energy efficient processing using ARM-based architecture will be investigated. Further, it will be explored if the use of a GPU for some more computationally complex processing can enable further improvement of the energy efficiency.

One of the platforms that will be used for this purpose is NVIDIA Jetson AGX Orin 64GB [23], which includes both ARM CPU cores and a GPU.

2.2.1 Using ARM architecture for L1/L3 processing

2.2.1.1 Architecture Definition

The most common O-RAN split option between RU and DU is 7.2, as shown in Figure 2-5. In this mode, the lower part of L1 is implemented at the RU and the higher part of L1, i.e. Medium Access Control (MAC) and Radio Link Control (RLC), are implemented at the DU. Usually the lower L1 modules have significant computational complexity, but there are many modules on the higher L1 that are also computationally complex.

In the UE transmitter (similarly to the gNB transmitter shown in the upper part of Figure 2-5), the payload bits are transferred from MAC to PHY, where they are being channel coded and mapped into Quadrature Amplitude Modulation (QAM) symbols over the time/frequency grid. A Cyclic Prefix is added to the symbols, and they are converted to the time domain. The time domain signal is up sampled, converted to analogue signal (using D/A converter), modulated to the carrier frequency and sent over the air via the UE antennas. Once transmitted, the signal undergoes a multipath fading channel and interference is added to it.

The RU at the gNB side then receives this signal from the gNB antennas, performs filtering, down sampling, analogue-to-digital conversion, additional down sampling in the digital domain, CP removal and conversion to the frequency domain using FFT. Note that different HW components in the RU (for example resistors) introduce thermal noise, which is added to the incoming signal on top of the interference.

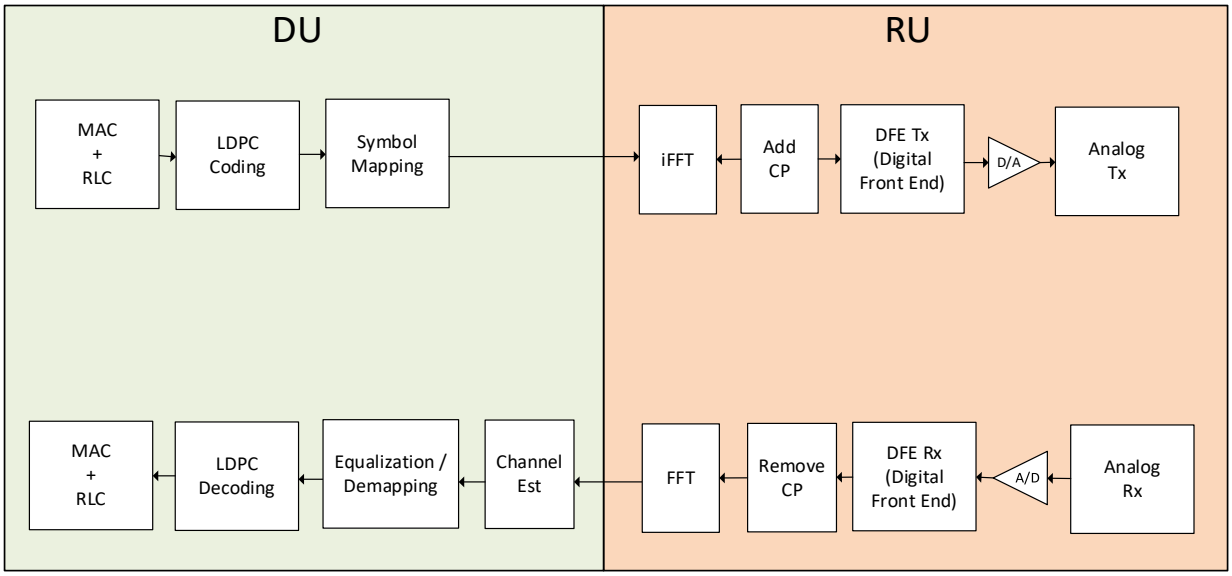


Figure 2-5 O-RAN RU/DU Split 7.2

The frequency domain samples produced by the RU are sent via the fronthaul to the DU. The purpose of the upper PHY in the DU is to receive these samples, detect the transmitted bits and to send them to the MAC layer.

In this mode, the lower part of L1 is implemented in the RU and the higher part of L1, MAC and RLC are implemented in the DU. Usually the lower L1 modules have significant computational complexity, but there are many modules on the higher L1 that are very computational complexity.

In the UE transmitter (similarly to the gNB transmitter shown in the upper part of Figure 2-5) the payload bits are transferred from MAC to PHY, where they are being channel coded and mapped into Quadrature Amplitude Modulation (QAM) symbols over the time/frequency grid. A Cyclic Prefix is added to the symbols, and they are converted to the time domain. The time domain signal is upsampled, converted to analogue, modulated to the carrier frequency and sent over the air via the UE antennas. Once transmitted, the signal undergoes a multipath fading channel and interference is added to it.

In 5G-NR the main UL data channel is called Physical Uplink Shared Channel (PUSCH). The main modules of a PUSCH receiver are:

- Channel Estimation – estimates the multipath channel response based on the Demodulation Reference Signal (DMRS) embedded in the PUSCH allocation.
- Equalisation – uses the channel estimates to reverse the effect of the channel and to mitigate noise and interference.
- Demapping – performs demapping from the equalised symbols domain to the bit domain and generates log likelihood ratios (LLRs). Their sign represents the bit values, and their amplitude represent the reliability of each of these bits.
- LDPC decoding – The channel coding scheme used in PUSCH is called LDPC. 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 has been reached. Note that a decoder that receives LLRs at its inputs rather than simply bits is called a soft decision decoder, and the LLRs are sometime referred to as soft values. Soft decision decoders performance is much better than hard decision ones, as the decoder uses the LLRs' amplitudes to learn how much it can "trust" their corresponding bits. Hence, in most cases gNB receivers (as well as UE receivers) use soft decision decoding.

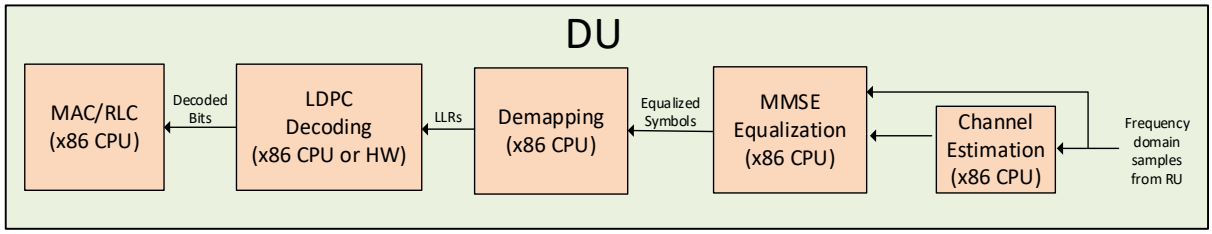


Figure 2-6 Legacy linear PUSCH Receiver using x86 architecture

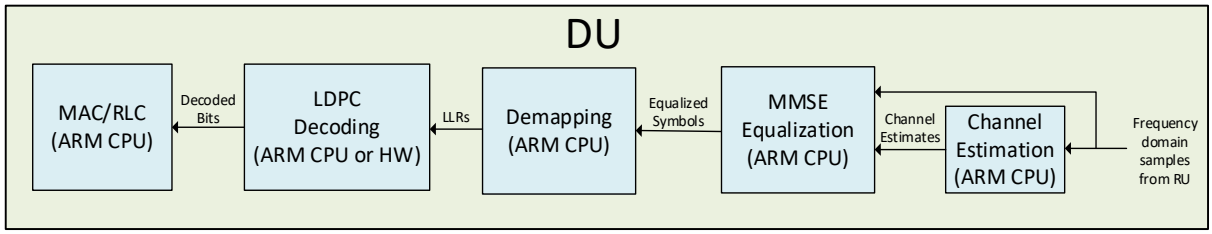


Figure 2-7 Linear PUSCH Receiver using ARM architecture

An example legacy DU implementation of a PUSCH receiver is shown in Figure 2-6. We can see there that most modules are implemented in SW on a CPU. The exception is the LDPC decoder, being the computational complexity for such a decoder very high. Therefore, a SW implementation is usually used when the number of UEs and number of total Resource Blocks (RBs) is relatively low, where for larger loads HW accelerators are used.

Legacy x86 architecture is inherently power hungry. In BeGREEN, we are planning to preserve the basic system as in the legacy receiver but use ARM-based architecture for the L1 processing rather than x86. We will quantify the power consumption in both architectures and expect that it would be lower using the ARM architecture compared to the x86 architecture. This is because ARM architecture is inherently less power hungry than Intel, as it evolved from low power devices. Unlike ARM, the x86 architecture focuses on performance. In addition, the feasibility of implementing the LDPC decoder over the ARM architecture will be examined. Note however, that as with the x86 case, a HW LDPC might still be needed for high loads.

2.2.1.2 CU/RIC implementation using ARM

In a traditional CU implementation, shown in Figure 2-8.a, all the traffic going from the NIC to the CU framework, has to pass through the Linux Network Stack (LNS), resulting in a duplicated processing of GPRS Tunnelling Protocol (GTP) messages, once by the LNS and other by the CU itself. In the Accelerated CU version (see Figure 2-8.b), an eXpress Data Path (XDP) is hooked into the NIC. The XDP uses the Berkeley Packet Filter (BPF) to intersect GTP messages and delivers it directly to the CU Framework without passing through the LNS. Here, the LNS will only manage no-GTP traffic, Control traffic and first GTP message, resulting in an offloading of nearly 60% of the traffic directly to the CU Framework, reducing the power consumption of the CU.

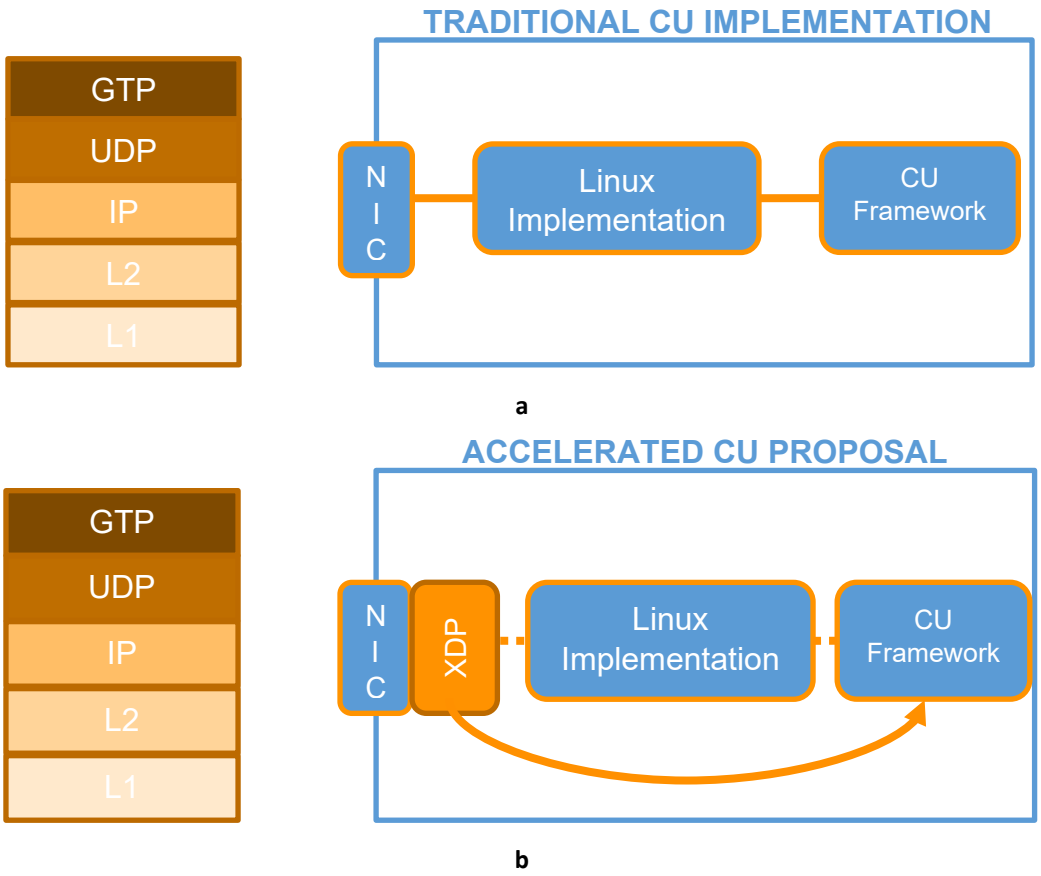


Figure 2-8 CU framework implementation, a: traditional; b: with XDP.

ARM servers are known for their superior energy efficiency compared to traditional x86 servers due to a combination of factors. Firstly, ARM processors employ a Reduced Instruction Set Computing (RISC) architecture, which simplifies instructions and enables tasks to be completed with fewer clock cycles, ultimately reducing power consumption. This is one of the mayor advantages for implementing XDP applications for the CU, since clock cycles will be reduced in higher traffic scenarios managed by the CU. Additionally, ARM chips are designed with power efficiency as a primary consideration, utilising smaller process nodes and boasting lower thermal design power ratings. This leads to diminished power consumption and less heat generation during operation. Moreover, ARM's customisation capabilities and scalability allows for tailored solutions, optimising power usage for specific workloads such as sequential traffic packet management. Furthermore, dynamic power management enables ARM chips to adapt clock speeds and voltages based on workload and traffic demands, further reducing energy consumption.

2.2.2 Power optimisation metrics and HW capabilities of the ARM core

The Neoverse N1[4] core provides mechanisms to control both dynamic and static power dissipation. Dynamic power management includes the following features:

- Architectural clock gating.
- Per-core Dynamic Voltage and Frequency Scaling (DVFS).

Static power management includes the following features:

- Dynamic retention.
- Powerdown.

[4] Neoverse N1, https://www.arm.com/products/silicon-ip-cpu/neoverse/neoverse-n1

Architectural clock gating modes

When the Neoverse N1 core is in standby mode, it is architecturally clock gated at the top of the clock tree. Wait for Interrupt (WFI) and Wait for Event (WFE) are features of Armv8-A architecture that put the core in a low-power standby mode by architecturally disabling the clock at the top of the clock tree. The core is fully powered and retains all the state in standby mode.

Core dynamic retention

In this mode, all core logic and RAMs are in retention and the core domain is inoperable. The core can be entered into this power mode when it is in WFI or WFE mode. The core dynamic retention can be enabled and disabled separately for WFI and WFE by software running on the core.

For more information about WFI and WFE, see the Arm® Architecture Reference Manual Armv8[5] , for Armv8-A architecture profile. For more information about HW capabilities, see Power Management in Arm® Neoverse N1 Core Technical Reference Manual[6].

2.2.3 Acceleration using GPU for L1 computationally extensive functions

2.2.3.1 Proposed Architecture

As discussed in section 2.2.1, the computation complexity of LDPC decoding is very high. CPUs are general purpose processors and are not necessarily optimal for the type of the arithmetic and logical operations required for LDPC decoding. This inefficiency leads to increased power consumption when LDPC is implemented on a CPU and, what is even worse, in many cases (as discussed) the CPU resources are not sufficient for completing the required LDPC decoding capacity which requires the usage of expensive non-flexible HW accelerators. In BeGREEN we are planning to overcome these disadvantages by investigating the option of using the same ARM architecture, but with offloading the LDPC decoder to a GPU. This is shown in Figure 2-10.

Another disadvantage of the suggested receiver is its MIMO performance. Basically, MIMO relies on the spatial diversity of the channel to increase capacity, and by that to increase the number of users the channel can support and the achievable throughput for each user. The level of spatial diversity is determined by the correlation between the Tx antennas, the correlation between the Rx antennas and the fading channel. As an extreme example, when the channel is line-of-sight (LoS), the spatial diversity is very low, and the MIMO rank is not much larger than one. The equalizer used in the architecture shown in Figure 2-9 is a Minimum Mean Square Error (MMSE). The computational complexity of such an equalizer is relatively low due to its linearity, and it offers good performance for SISO receivers. As for MIMO receivers, MMSE performs well in cases where spatial diversity is high. However, with the intent of increasing the capacity of the channel, the scheduler strives for increasing the MIMO rank as much as it can. When the conditions (Tx antennas, Rx antennas & channel) are such that the spatial diversity is low, then the achievable rank is also low. This will result in the link adaptation mechanism to a low rank or even worse, insist on a high rank and suffer significant performance degradation that can result in Radio Link Failure (RLF).

For achieving good performance with significantly higher MIMO, BeGREEN suggests the sphere decoder-based architecture shown in Figure 2-10. This decoder is implementing an algorithm similar to Maximum Likelihood but with reduced complexity. The differences between these algorithms are described in section 2.2.3.2. In essence, the Maximum Likelihood demodulator goes over all hypotheses for transmitted symbols and chooses the most likely one.

[5] Armv8-M Architecture Reference Manual, https://documentation-service.arm.com/static/615c3754e4f35d2484678be7?token=
[6] Arm® Neoverse™ N1 Core Technical Reference Manual, https://documentation-service.arm.com/static/5e7e3d85b471823cb9de6210?token=

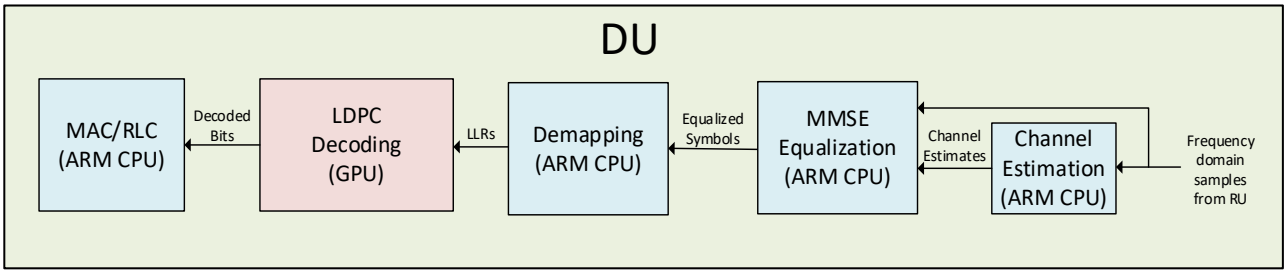


Figure 2-9 Linear PUSCH receiver using ARM plus GPU architecture

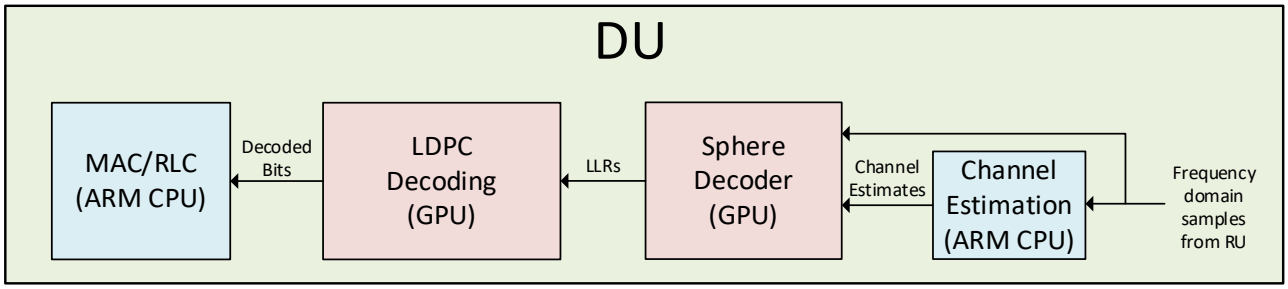


Figure 2-10 Sphere decoder PUSCH receiver using ARM plus GPU architecture

This receiver will truly find the best option and will have much better performance that the MMSE receiver. The fundamental challenge with this type of receiver is that it requires very high computational complexity. The higher the QAM modulation and the MIMO rank, the higher the number of hypotheses the receiver needs to evaluate. As an example, for 256QAM and rank 4, the number of hypotheses is 256^4, which is more than 4 billion. In addition, this needs to be performed over all PUSCH Resource Elements (REs) that can be up to ~40000 for 100 MHz BW resulting in 160 trillion hypotheses to go over. To reduce this number, instead of straight forward Maximum Likelihood decoder, the usage of a Sphere Decoder significantly reduces the number of hypotheses. The reduction strategy needs to be "intelligent". An example of a strategy is containing the hypotheses within a multi-dimensional sphere [24].

BeGREEN will investigate the sphere decoder in two different aspects. The first is to examine innovative strategies to reach a state where the number of hypotheses that are left is as small as possible and that these are the ones which are most likely to be the correct. The end goal is to evaluate as less hypotheses as possible while maintaining good receiver performance for a wide range of spectral diversity levels. As a result, the computational complexity will be significantly reduced and, therefore, the energy efficiency will be improved. However, running the sphere decoder on a CPU can still be very challenging and will result in extensive power consumption and increased number of required CPU cores. Therefore, this will result in higher Capital Expenditures (CAPEX) for the MNOs. Furthermore, this project will also focus on the implementation aspect of the Sphere Decoder on a GPU. The GPU potentially can be very effective in the sphere decoder processing, mainly due to it very high number of cores and threads. The calculations needed to be performed are similar to each RE and each hypothesis and can be kernelised and parallelised in the GPU. Therefore, it will reduce the load off the CPU, freeing it up to perform other DU tasks such as channel estimation and MAC/RLC processing. Most important, when utilising the available GPU resources, it potentially can reduce the overall power consumption of the HW. Note that in many cases the sphere decode is implemented in silicon or FPGA [25], which can also be very power efficient, as they are specifically dedicated for this task. However, when the number of users and the load of the network reduces (for example at night) the same GPU resources that are used for the sphere decoder can be used for other tasks, for example performing the learning needed for AI technologies used by the BeGREEN "Intelligent Plane" (described in Work Package 4 – WP4). This introduces an advantage for GPU usage over silicon or FPGA, which lack this versatility. In addition, GPU is inherently flexible, which makes it much easier to account for different scales and loads and also enables constant upgrades and improvements.

As shown in Figure 2-10, the LLRs are the outputs of the sphere decoder but also the inputs to the LDPC decoder. This offers an additional implementation and power consumption advantage to this architecture, which implements both on the same GPU. This decreases the overhead of transferring large amounts of data between the CPU and the GPU or between the CPU and silicon/FPGA implementations of the sphere decoder.

2.2.3.2 Algorithmic description of the modules being offloaded to GPU

2.2.3.2.1 Sphere decoder

As shown in Figure 2-5, the DU receives the frequency domain samples from the RU, and then uses the DMRSs to perform channel estimation. To generate the LLRs needed for the channel decoder, the receiver can either perform MMSE equalisation and then demapping, or to use a Sphere Decoder that generates LLRs internally.

The basic operation for an MMSE equaliser per PUSCH RE is given with:

$$\hat{\mathrm{d}} = \mathrm{H}^{\mathrm{H}}\left(\mathrm{H}\mathrm{H}^{\mathrm{H}} + \mathrm{R}_{\eta}\right)^{-1}\mathrm{r} = \mathrm{W}\mathrm{r}$$

where H is the channel matrix, R_{η} is the noise and interference covariance matrix, W is the equalisation weight matrix and $\hat{\mathrm{d}}$ is the equalised symbol. The equalised samples are then demapped into LLRs depending on the QAM modulation used. An example for QPSK is shown where each symbol represents two bits. For this case the LLRs are obtained as follows:

$$LLR_0 = \frac{real(\widehat{\mathrm{d}})}{norm\_factor} \qquad LLR_1 = \frac{imag(\widehat{\mathrm{d}})}{norm\_factor}$$

This is illustrated in Figure 2-11.

This type of receiver is based on finding the MMSE between the originally transmitted symbols and a linear operation over the received signal and works very well for SISO cases and for MIMO cases when spatial separation between the layers is high. However, when this is not the case, the performance degrades. The reason being that the relation between the received signal and the originally transmitted symbols cannot be easily represented as a linear operation. More specifically, the matrix being inverted $(\mathrm{H}\mathrm{H}^{\mathrm{H}} + \mathrm{R}_{\eta})$ has linear dependency between its rows, columns, or both, making it not invertible.

However, the cases where spatial separation is low are very common. For these cases, BeGREEN suggests to use a sphere decoder.

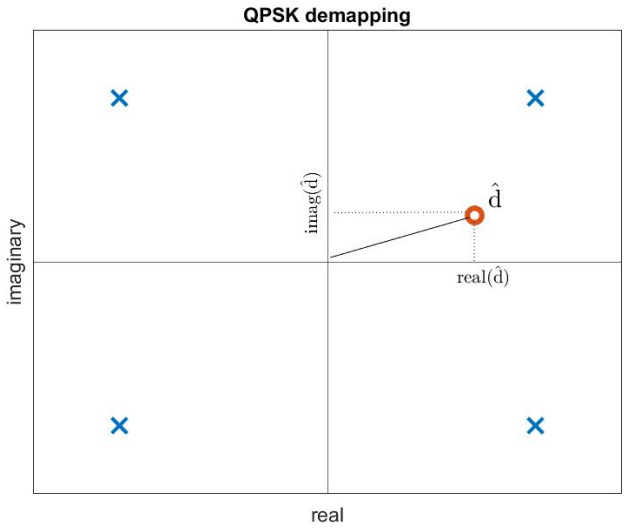


Figure 2-11 Demapping example for QPSK

This receiver is based on the Maximum Likelihood decoder that, in general, checks the most likely transmitted symbols based on the received signal as follows:

$$\hat{d} = \underset{d \in \mathfrak{D}}{\mathbf{argmin}} \{(r - Hd)^H R_\eta (r - Hd)\}$$

To elaborate, we calculate the distance metric $(r - Hd)^H R_\eta (r - Hd)$ for each hypothesis of d. This metric is the distance between the received signal and a single hypothesis of the originally transmitted symbols after undergoing the channel. We then calculate the minimal values between the distances for all hypotheses and to find the best estimation of the transmitted symbol. However, the number off hypotheses is very large. For example, for 4 layers and 256 QAM we have a total of 256^4≅4e9 distances to calculate.

Sphere decoder, like the ML decoder calculates distance metrics and chooses the one with minimal distance, but with reduced complexity. This is done by intelligently omitting a lot of the hypothesis which are less likely to achieve the minimal distance. BeGREEN will explore novel methods for hypotheses omission, and the reduction in the number of distance calculations will be translated into reduced power consumption.

LLR generation in the sphere decoder, unlike in the linear receiver (previously described), where in the first and second stage the MMSE equalisation and demapping between the equalised samples and the LLRs are performed respectively, the LLR calculation is not a two-stage process and is embedded within the general algorithm. It calculates the Euclidean distances between the received signal and the expected received signal corresponding to each possible transmitted QAM symbol (e.g. 256 options for 256 QAM) after undergoing the estimated fading channel. These Euclidian distances are used for calculating the LLRs for each of the received bits (e.g., 8 bits for each 256 QAM symbol).

2.2.3.2.2 LDPC Decoder

The soft decision LDPC decoder for PUSCH receives the LLRs either from the linear MMSE equaliser and demapper as shown in Figure 2-7 and Figure 2-9, or from the sphere decoder as shown in Figure 2-10. These include both the originally transmitted systematic bits and parity bits. The fading channel, noise and interference may cause some of these transmitted bits to be flipped. The added parity bits are meant to be used in the LDPC decoder for correcting the flipped bits using an iterative process. The LLRs are signed values. Their sign represents the bit value (1 stands for 0 and -1 stands for 1) while their amplitudes represent their reliability and help the decoder decide how much it can “trust” their corresponding sign.

The ratio between the number of originally payload bits and the total number of bits that are transmitted including the parity bits is called “code rate”. Lowering the code rate achieves better resilience to tough channel conditions. However, this comes at a cost of reducing the throughput. The 5G NR 3GPP standard [26] defines an MCS table shown in Table 2-1. This table uses 28 different Modulation and Coding Schemes (MCSs) that can be used. Starting from QPSK with low code rates, which are meant for very tough channel conditions and convey less than a quarter of a bit per symbol going all the way up to up to 256QAM with high code rates, which can be used only when channel conditions are very good and convey 7.4 bits per symbol. The LDPC decoder is required to be generic and support all code rates and is expected to achieve the best performance for each of these MCSs.

Table 2-1 5G NR 3GPP Standard MCS

MCS Index	Modulation Order Qm	Target Code Rate R x [1024]	Spectral Efficiency
0	2	120	0.2344
1	2	193	0.377
2	2	308	0.6016
3	2	449	0.877
4	2	602	1.1758

MCS Index	Modulation Order Qm	Target Code Rate R x [1024]	Spectral Efficiency
5	4	378	1.4766
6	4	434	1.6953
7	4	490	1.9141
8	4	553	2.1602
9	4	616	2.4063
10	4	658	2.5703
11	6	466	2.7305
12	6	517	3.0293
13	6	567	3.3223
14	6	616	3.6094
15	6	666	3.9023
16	6	719	4.2129
17	6	772	4.5234
18	6	822	4.8164
19	6	873	5.1152
20	8	682.5	5.332
21	8	711	5.5547
22	8	754	5.8906
23	8	797	6.2266
24	8	841	6.5703
25	8	885	6.9141
26	8	916.5	7.1602
27	8	948	7.4063
28	2	reserved	
29	4	reserved	
30	6	reserved	
31	8	reserved	

After adding the parity bits to the payload bits, the corresponding message becomes a codeword. The validity of the codeword can be checked by applying the parity check matrix. A resulting zero means that the codeword is valid, which means that the message has been conveyed properly. An example of a parity check matrix is:

$$\boldsymbol{H} = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 & 1 & 0 \end{pmatrix}$$

In this example, if Hc=0 → c is valid codeword.

The BeGREEN LDPC decoder uses the "message passing" algorithm and uses the parity check matrix (which is known at the receiver) iteratively to correct the erroneous bits. The message passing algorithms passes the LLRs back and forth from bit nodes to the check nodes. This is shown in Figure 2-12.

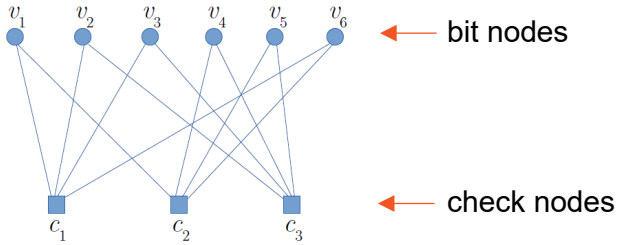


Figure 2-12 Message passing algorithm

Each LDPC decoding iteration is divided into two parts:

The first half of an iteration passes the messages from the bits nodes to the check nodes and it is called: q-messages. These are initialized to input LLRs at the first iteration, and for all other iterations require calculating:

$$q_n^{[i]} = LLR_n + \sum_{m:H_{mn}=1} r_{mn}^{[i]}$$

The second half of an iteration passes the messages from the check nodes back to the bit nodes and it is called r-messages. These are initialized to zeros at the first iteration, and for all other iterations require calculating:

$$r_{mn}^{[i]} = \left[\prod_{k:H_{mn}=1,k\neq n} \text{sign}\left(q_k^{[i-1]} - r_{mk}^{[i-1]}\right)\right] \cdot \min_{k:H_{mn}=1,k\neq n} \left|q_k^{[i-1]} - r_{mk}^{[i-1]}\right|$$

Successful decoding is achieved when all erroneous bits have been corrected before the maximal number of iterations has been reached.

2.2.3.3 L1/L3 candidate use cases for potential power improvement

To be able to verify the robustness of the suggested L1 receiver algorithms in terms of receiver performance and energy consumption, several scenarios should be tested. They should cover different MIMO configurations and different environments, such as those listed in Table 2-2.

All scenarios will be run with low and medium Tx and Rx antennas correlation configurations. A deployment of 100 MHz bandwidth, 273 PUSCH RBs and at least 6 sectors will be tested. They will run in C-band with a carrier frequency of ~3.5 GHz.

The scenarios presented above are related to the mMIMO and the Relay enhanced communication use cases defined in BeGREEN's deliverable D2.1 [1].

Table 2-2 Scenarios for DU Acceleration Performance Evaluated

Scenario	MIMO Mode	# of layers	Environment
1	SU-MIMO	4	Rural
2	SU-MIMO	4	Urban
3	MU-MIMO	4	Rural
4	MU-MIMO	4	Urban

2.3 Key Performance Indicators (KPIs) for energy efficiency

The legacy implementations described in section 2.1 will be compared to the proposed BeGREEN implementations/improvements described in section 2.2. The most suitable comparison would be in terms

of power consumption. However, it may be challenging to accurately measure its metrics and isolate the specific algorithm. Hence, the execution times and execution cycles can give a good estimation of the overall power consumption and can be measured much more easily. A tool that can be used to get insights on the CPU and GPU power consumption is the "Jetson Stats Package" [27].

We expect that the power consumption of the BeGREEN implementations will be reduced by at least 15% compared to legacy implementations.

2.4 Takeaways for BeGREEN Reference Architecture

In BeGREEN the modules being developed (LDPC decoder and sphere decoder) are part of the gNB PUSCH receiver. The PUSCH performance is measured in two ways:

- Block error rate (BLER) vs. Signal-to-Noise Ratio (SNR). The 10% point is of special interest.
- Throughput vs. SNR. The 10% point is of special interest.

As basic performance, the tests defined in section 8.2.1 of the 3GPP conformance tests [62] will be performed. The expectation is that the receiver needs to pass the requirements with a margin of at least 2dB.

At next stage, BLER and throughput tests that represent the BeGREEN use cases defined in section 2.2.3.3 will be performed.

These results will give insight on the Data Rate, Bandwidth and connection density KPIs defined in BeGREEN deliverable D2.1 Reference Architecture [1].

3 O-RAN Based Strategies for RU Power Consumption Optimisation

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 SMO to control RUs, but the RU is not directly manageable by the Near-Real-Time RIC (Near-RT RIC). To achieve advanced power control and beamforming management of the RU, it is necessary to extend the capabilities of the RU, DU, RIC and SMO entities for energy saving management. The hybrid OFH M-Plane model in the O-RAN architecture enables the DU to control some aspects of the RU. However, there are no E2 Service Models (E2SM) to enable Near-RT RIC control of the RU. The Small Cell Forum's (SCF) 5G functional application platform interface (FAPI) standards [29][30][31] specify interfaces to control and monitor the RU from the DU, but their alignment with future O-RAN and 3GPP specifications requires further research.

This chapter proposes extensions to the O1 and OFH M-Plane and the development of energy-aware rApps to control power usage of the RUs. It also proposes studying and developing E2 extensions to expose the hybrid OFH M-Plane or leverage the SCF P19/P4 [30][31] interfaces for energy-aware xApps, demonstrating their capabilities. The overall best strategies for combinations of energy-aware rApps and xApps will be researched, prototyped, and demonstrated for future market-focused applications.

The general principles of the RU architecture and the use cases related with Network Energy Saving (NES) are described in [32] and extended in section 3.1. The role of the RIC based xApps/rApps in terms of power consumption management is described in section 3.2. Detailed information about the test demos related will be described in deliverables D4.1 [38] and D5.1 [39].

3.1 Legacy and state-of-the-art RU Power amplifiers optimisation

RAN energy saving depends on planning and configuration. Due to the varying nature of the network traffic, load and to user mobility, the optimisation of the RAN energy consumption is complex and there is a risk that RAN equipment may consume much energy while serving low traffic.

Different energy saving features are investigated by the industry. Examples are deep sleep mode, e.g., shut down of a base station (BS) of a given technology; carrier shut down; and RF channels' switch off/on. More recently, short time scales energy saving advanced mechanisms have been proposed, at the symbol-, subframe and frame levels.

The RAN O-RUs are responsible for a major part of the mobile network energy consumption. Proposed solution for the O-RAN use cases include:

1. Energy saving by carrier and cell switching.
2. Energy saving by massive MIMO (mMIMO) RF channel switching.
3. Advanced Sleep Mode energy saving. This feature is expected to reduce power consumption by partially switching off O-RU components.

The current version of the O-RAN specifications includes a non-mandatory standard that deals with energy efficiency within the RU module as follows:

- O-RAN RU is hosting the Low-PHY layer (FFT/iFFT, PRACH extraction) and RF processing based on a lower layer functional 7.2 split. O-RAN WG1.Use-Cases-Analysis-Report-R003-v11.00 [33] considers the RAN energy consumption as an important topic for network operators, especially for 5G network.
- Energy savings are further discussed and studied at O-RAN Work Group 1 (Use Cases and Overall Architecture) Network Energy Saving Use Cases Technical Report [34]. For the O-RU Specific KPIs, - Energy efficiency and power consumption KPIs shall be provided by real-time metering.

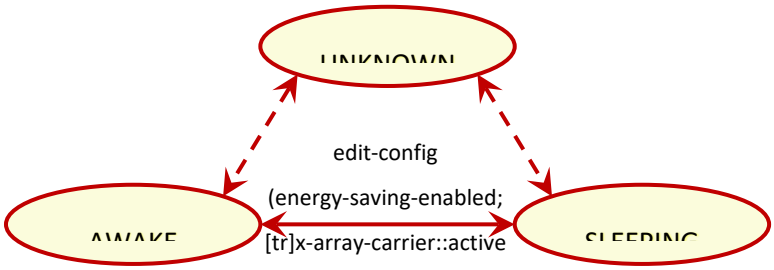


Figure 3-1 Transition diagram for power-state node

- O-RAN Control, User and Synchronisation Plane Specification [35] is the major input for O-RU requirements and implementation methods. Therein, the only related feature for energy savings is transmission blanking and it is an optional requirement.
- O-RAN Management Plane Specification [36] also defines the management requirements and techniques to accomplish O-RU energy savings.

Regarding the latter, it includes basic description of how an O-RU's carrier can be deactivated to achieve the energy savings within an O-RU and synchronisation aspects for carrier deactivation to accomplish energy saving. The power-state transition diagram for O-RU is presented in Figure 3-1 (taken from the above spec). This state can be indirectly controlled by editing the parameters energy-saving-enabled and [tr]x-array-carrier: active, as illustrated in the figure.

States description is as following:

- AWAKE: This value of power-state node indicates that the O-RU is operating normally, i.e., not in energy saving mode.
- SLEEPING: This value of power-state node indicates that the O-RU is in energy saving mode. M-plane connection and functions are alive.
- UNKNOWN: This value of power-state node can be exposed by the O-RU e.g., in case the O-RU does not know its power-state value is AWAKE or SLEEPING. This value of power-state node is optional.

The O-RU *epe-stats* [37] statistics shall include the performance measurements for energy, power and environmental parameters as shown in Table 3-1.

O-RAN Work Group 7 (White-box Hardware Workgroup) document on Network Energy Savings (NES) Procedures and Performance Metrics [32] describes NESs methods and requirements associated mainly with the O-RU. Energy savings KPIs and estimated power reductions are also presented. The mapping of energy savings to existing M-Plane and C-Plane commands is discussed. Follow-on versions will elaborate, e.g. on advanced sleep modes.

Table 3-2 summarizes the expected influence of the presented methods. The presented savings column includes an estimate for the O-RU energy saving mechanisms expected power savings.

Table 3-1 Measurements for Energy, Power and Environmental Parameters

Measurement Objective	Description
POWER	Value of measured power consumed by identified hardware component
TEMPERATURE	Value of measured temperature of identified hardware component
VOLTAGE	Value of measured voltage of identified hardware component
CURRENT	Value of measured current of identified hardware component

Table 3-2 O-RU NES Mechanisms

NES Mechanism	Traffic Capacity Impact	Wake-Up Latency	Traffic Disruption During Re-activation of Disabled Resources	Energy Savings Range Percentage vs Active Mode Reference Power Consumption	Comment
All RF carriers shutoff	No user data traffic	100s of ms or several secs	N/A (already no traffic)	80-99%	-
RF Band shutoff	Yes	-	No	40-50%	Entire front end for disabled RF band can be put in very low power state
CC shutoff within a band	Yes	-	FFS	Few %	Limited reduction in RF Front-end power; main power reduction will be via reducing JSED bandwidth and DUC/DDC filtering
RX/TX Array order reduction	Some impact	ms or 10s of ms	Likely	10s of %	Reducing number of RF front-end chains will directly reduce front-end power, which is dominant power consumer for medium and wide-area BSs
Reduced # of MIMO spatial layers	Likely	-	-	Unclear	Any change in # of antennas required could be offset by need for additional data symbols in time domain

Table 3-3 presents the energy savings mechanisms mapping to M-plane and C-plane commands. For existing commands, e.g., RF band disable, commands follow up is referenced. For the commands that do not exist, further study and discussions will be required, e. g., advanced sleep mode. To provide an effective low power solution, the above network energy savings features may imply changes to the O-RU hardware architecture and also its associated firmware.

Table 3-3 Energy Savings Mechanism Mapping to M-Plane and C-Plane Commands [36]

NES Feature	Document Section [36]	M-Plane or C-Plane Control	Commands Exist?	Comment/ Follow-Up
Disabling all RF bands	4.2.1	M-Plane	Yes	As per M-Plane Section 15.3.2
Disabling one or more bands in an O-RU	4.2.2	M-Plane	Yes	As per M-Plane Section 15.3.2
Disabling one or more CCs in an O-RU band	4.2.3	M-Plane	Yes	As per M-Plane Section 15.3.2
RF Channel Reconfiguration	4.3	M-Plane, possibly C-Plane	No	WG4 discussion topic
Active Low Power Mode	4.4	M-Plane	No	M-Plane command proposal pending

Advanced Sleep Mode	4.5	C-Plane, M-Plane	No	Command formats now in WG4 discussion

3.2 Near-RT and non-RT Applications for RU Management

Near-real time (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. For instance, Dynamic Spectrum Sharing (DSS) is a near real-time application that promptly allocates radio spectrum resources, either 4G-LTE and 5G-NR, which can be optimised to be allocated based on immediate demand within a Near-RT xApp. Another example is interference management applications, which swiftly identify and mitigate sources of interference, ensuring optimal network performance.

In the realm of RU management, these near real-time applications provide a series of benefits in terms of allocation of resources and optimisation of performance parameters such network traffic and power consumption. They could dynamically adjust RU configurations in response to rapidly changing network conditions or user demands, meaning that they can quickly allocate different frequency bands to RUs for load balancing, capacity and coverage optimisation and power control, if the proper interfaces and mechanisms are available at the RU/DU.

Non-real-time (non-RT) applications (rApps), in contrast, 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. For example, traffic forecasting and planning applications analyse historical data to predict future traffic patterns, aiding in capacity planning and energy optimisation. Network analytics and reporting tools generate performance reports and analytics for long-term planning and optimisation. In terms of RU management, non-real-time applications focus on long-term configuration planning, analysing trends and planning RU configurations for future network expansions or upgrades. Additionally, they can process historical performance data to identify patterns and make strategic adjustments to RU parameters. Their benefits lie in their ability to make informed, strategic decisions for long-term network optimisation, either for capacity, coverage, or power.

The technical report O-RAN WG1.NESUC-R003-v02 [34] provides a general description of use case for energy management and control of the O-RUs within the O-RAN specifications. Here, use cases such as Carrier and Cell Switch off/on, RF Channel reconfiguration, advanced sleep mode selection and O-Cloud resource energy saving mode; are described and evaluated from the Real-Time and Non-Real Time RIC as described in section 3.1. Figure 3-2 shows an example of the flow diagram interaction proposed by the O-RAN to control the RU under cell on/off schemes through the non-RT RIC. This one is compared with the Near-RT RIC proposed version shown in Figure 3-3, where the decision-making is envisioned in a real time scale to provide instant RU configurations to save energy.

In BeGREEN, the implementation of the RIC will include different deployments and flavours for the NearRT 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 Intelligent 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 Intelligent Plane for easy integration with the Intelligent SDK. Here, an Intelligent xApp will control the RU utilisation, power transmission and cell status to reduce its energy consumption. More details of this implementation can be found in Section 4 of the BeGREEN D5.1 [39].

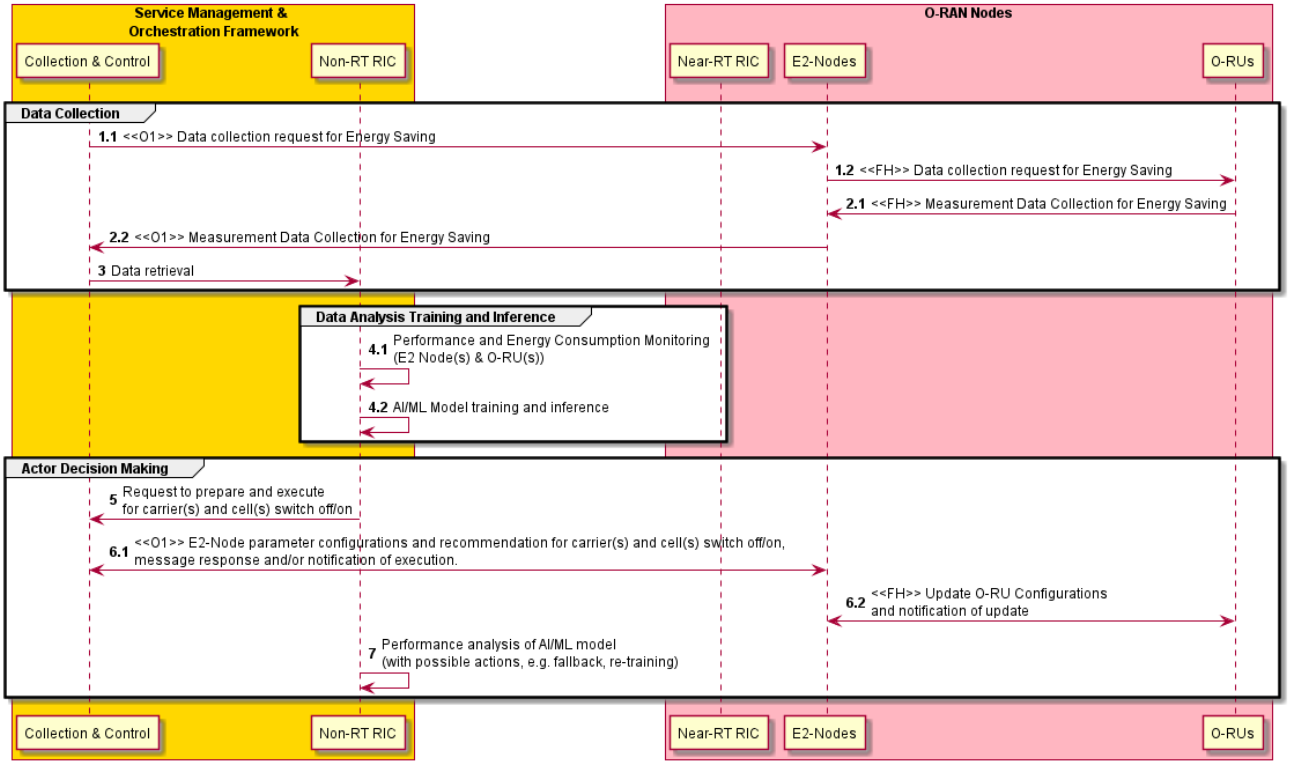


Figure 3-2 O-RAN cell on/off flow diagram using non-RT RIC

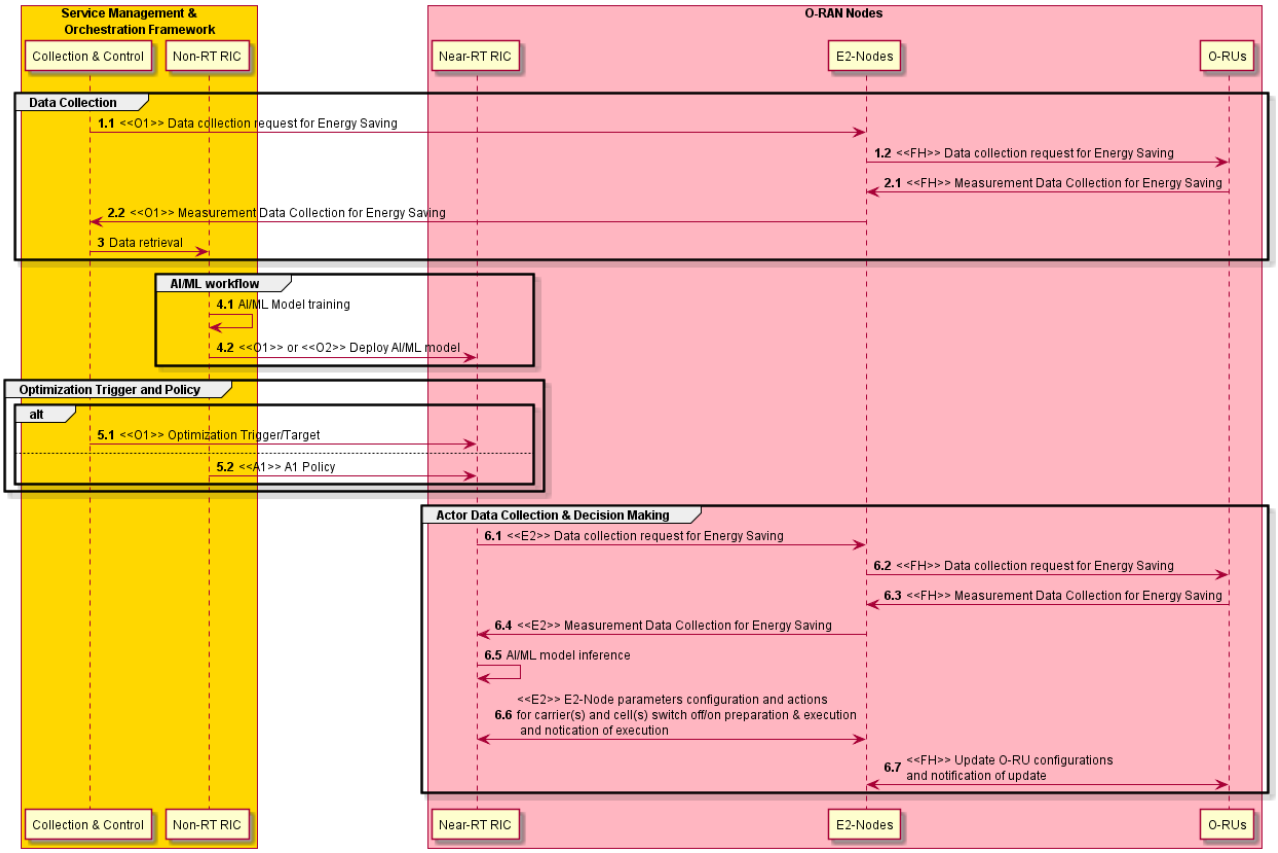


Figure 3-3 O-RAN cell on/off flow diagram using Near-RT RIC

4 PHY Layer Enhancements for Radio Network Energy Efficiency

In the recent years, a few different physical layer technologies arose, promising to offer new features implemented in the upcoming 6G mobile networks. Some of these physical layer technologies include ISAC as well as RIS, forecasting benefits such as improved energy efficiency. Additionally, techniques that could help in this direction such as envelope tracking for the PA as well as digital pre-distortion in the RU are actively being developed. All these new PHY-layer technologies are expected to offer new services or to improve the power consumption of the RAN.

4.1 ISAC, RIS and their contribution to network energy efficiency

ISAC has recently emerged as an additional functionality to be inherent of 6G networks. It is supposed to enable the networks to wirelessly sense the physical environment and to create, the so called, perceptive mobile network (PMN) [51]. PMNs are supposed to sense the environment to enable different types of services. The ISAC functionality of the network is similar to RADAR, with a difference that the same waveform is used for data communications as well as for the sensing/RADAR function. The sensing functionality of ISAC is intended to be used for detecting objects, activities, and event recognition [52] while, at the same time, being able to pinpoint the location of mentioned features of interest. Additionally, sensing assisted communications are becoming a popular solution for further network and radio interface optimisation.

Initially, the sensing functionality was intended for **deviceless sensing**, i.e. RADAR-like sensing, where the object to be detected do not possess any RF devices. Instead, the RUs must be able to function as a RADAR i.e. being able to receive an RF signal simultaneously during transmission, which is usually doable with minor changes in the RUs' hardware. This way the reflections from surrounding objects are received back at the RU, making possible the estimation of their positions.

The sensing concept can be extended to **device-based sensing**, enabling that multiple network nodes (RUs, UEs, etc.) can function as a complete sensing system that enables the so-called network sensing. With this type of sensing one can precisely pinpoint the position of the UEs, and an exact UE density will be obtained. In this approach the positions of the transmitting UEs are estimated based on the time of arrival of the transmitted signals towards different RUs. This type of sensing requires timestamping of the received frames and synchronisation of the RUs, capabilities that usually RUs offer.

If we consider the capability of a network to perform **network sensing** itself, it does mean that all hardware nodes, e.g., RUs, UEs, in the network feature already the sensing functionality.

4.1.1 ISAC state-of-the-art and developments for energy efficiency

The early investigation of ISAC systems concentrated on waveform design based on MIMO and beamforming techniques used in communications [40][41]. In [42], an orthogonal frequency division multiplexing (OFDM) symbol-based ISAC signal processing method was proposed, which overcomes the typical drawbacks of correlation-based radar signal processing and satisfies both the radar ranging and communication requirements. In [43] the authors proposed a reconfigurable and unified multifunctional receiver for data fusion services of radar sensing and radio communication based on time-division platform. In each time slot assigned for radar sensing or radio communication modes, the system can perform localisation function or data communication, respectively.

Emerging ISAC systems promise to affect different wireless communications aspects, as well as, to offer new services [63]. Further, RIS is also being studied for improving radar and communications [45][46]. ISAC systems can share the unified transceiver, the same spectrum and digital signal processing HW to exploit the reflected echo of communication beam to perform radar sensing function [47], and achieve immediate improvement in both the spectrum and energy efficiency, compared with the conventional communication

system where the echo signal is not of interest and not used for sensing. The ISAC system also benefits from mutual sharing of sensing information for improved reliability and performance, e.g., using the range and Doppler knowledge to assist beamforming and channel prediction.

BeGREEN will investigate how the sensing functionality can improve the energy efficiency of the network. A few different directions will be pursued.

- First, **deviceless** sensing approach will be used to sense the geometry of the environment as well as the presence of potential users and the user density in such an environment. This information will be used to dynamically optimise network coverage and to reduce the RU power consumption, or to optimise, for example, some PHY-layer functions such as beam training, which is time-consuming and, therefore, increases energy consumption.
- The **device-based** approach will be of a great advantage when dynamic network optimisation is to be performed. Additionally, having knowledge of the environment, together with the current user density, an optimal coverage can be made to optimize the network for energy consumption.

4.1.1.1 ISAC aided beam training and tracking

Phased array antenna technologies used for beam training and beam tracking are well studied and known. Nevertheless, their deployment in commercial products is quite limited. These technologies are important in the sense that, by transmitting narrow beams, interference to other users is minimised and, additionally, a higher Effective Isotropic Radiated Power (EIRP) is achieved using lower transmit power. However, despite technologies like beam forming and beam training/tracking are well investigated, their deployment is not straightforward. The main issue is the beam training and search procedures, which can be time-consuming and energy inefficient. Additionally, it uses the wireless medium extensively, making it unavailable for other users. Therefore, beam training/search procedures are usually used for stationary connections, where the beam training and search is performed only at the start-up for initial beam alignment.

Having an efficient beam search and beam tracking approach would greatly improve the network efficiency. It would enable beamforming and beam steering techniques to be deployed, without extensive use of the wireless medium and thus saving energy, while enabling higher EIRP using less transmit power.

To solve the issues in mobile networks with such complicated, time and energy consuming beam training and searching approach, different approaches are being investigated in the recent years. Some of them are trying to reconstruct the environment to simplify the beam training procedures. In [48], for example, light detection and ranging (LIDAR) is used for environment reconstruction and reducing the overhead associated with the beam training procedures. Additionally, in [49], computer vision-based approaches are used to find the best beams while at the same time reducing the overhead. Anyway, the computer vision-based approaches will always be less preferred due to the privacy infringement issues. Due to this, position-based methods for optimising beam search and overhead reduction are being investigated in [50]. They can use different positioning technologies like GPS or eventually indoor positioning technologies. The main disadvantage of this approach is that the position of the UE should be estimated using additional technology, which normally is completely independent of the data transmission technology used in the device. This brings additional hardware complexity to the system, as well as additional power consumption. Furthermore, not all positioning technologies are available everywhere, which limits the use of positioning technologies for aiding the beam steering process only in certain areas where coverage is available.

Radar technologies for assisting beam search i.e. beam prediction, are becoming relevant, especially in vehicular systems. The main motivation is that radars are widely available in today's cars and can be reused for aiding the beam selection process. Nevertheless, radar is usually not available in wireless communication systems, making this approach not easily applicable. Moreover, the advantage of the ISAC concept and existing implementations using a variety of wireless technologies, make it suitable for aiding the beam

selection process. This will introduce a minimum complexity of the system, but it will allow for faster and more efficient beam search/tracking.

4.1.1.2 Applications and AI/ML approaches

ISAC is not being developed only to serve the purpose of PHY-layer function optimisation. It is a functionality that will be used for many different applications like environmental sensing [63], road traffic monitoring [64], 3D scene reconstruction, i.e., digital twin [65], etc. This can be extremely beneficial for the mobile network in terms of its perception of the environment.

In the current 5G, beyond 5G and the future 6G networks, many machine learning/artificial intelligence (ML/AI) algorithms are being introduced to optimize the network performance in terms of throughput, latency, availability, energy consumption [66], etc. These ML/AI algorithms require different telemetry data which is information of the current performance of the network. The telemetry data basically gives metrics of different parameters of the network in time. This data can include throughput, latency, number of connected users, etc. and can be used for further optimisation of the overall network and especially for power efficiency optimisation.

In the past, mobile networks were planned to be static and, usually, the optimal position and coverage of the BSs were estimated based on the expected traffic and its spatial distribution. However, the traffic and its spatial distribution are not static, i.e., they change over time. This change can be on an hourly basis, day-night, daily, business day versus weekend, and season basis, summer vs winter. The networks were not able to adapt the coverage based on these short time changes, making its power efficiency not optimal. This is mainly because the network is static and fast changes are not easily implementable. This kind of network optimisations are a perfect candidate for the ML/AI approach described before. The ML/AI algorithms can predict, based on the received telemetry, the spatial load distribution and adapt the coverage accordingly. This can mean for example turning off some of the RUs during the low load hour in the night and turning on additional RUs during the peak hours when additional capacity is needed. With this approach the network can save a large amount of energy, simply by reducing the standby periods, in which a large portion of the energy is used for no useful task.

4.1.1.3 RIS-assisted communication

Turning off some of the RUs will for sure reduce power consumption, but this should not reduce the network coverage, since this still might affect some users. The main goal is reducing the power consumption by sacrificing the network capacity when it is not needed. The logical approach will be to turn off some of the RUs and to increase the coverage of the others by increasing their transmit power. In densely populated areas, at least in some cases, this might not be possible, if there are large buildings blocking the signal propagation. This would require having some of the RUs to be powered on, even there is no need for the capacity they offer. To solve this issue, RIS can play a huge role by allowing coverage in some areas without activating additional RUs. The RIS will be used to reflect the signal towards the areas not covered by the active RUs, while at the same time consuming significantly less power. This approach will enable further optimisation even in densely populated areas by further reducing the number of active RUs needed.

To perform a network wide power consumption optimisation of the RAN, a telemetry for the traffic load will be needed. This way the network will know how many resources are needed and where. The main issue with this approach is that the exact location of the UEs is not precisely known, which hinders optimal resource allocation. Namely, the UEs will choose the RU with the strongest signal and will monitor the other nearby RUs in case a handover is needed. This will work for a static scenario of the network, but if the network coverage is to be changed dynamically, this approach will not work. The main issue is if the network is to be reconfigured, it must know the position of the users in order to choose the optimal RUs to be turned off. The estimated received signal strength indicator (RSSI) at the UE and the RU can be used for initial estimation of

the UE position, but this will not give a good estimate, especially in urban scenarios where multipath rich environment is to be expected. Precise position of users and their density is essential for this approach and, therefore, ISAC is the solution which will be pursued in this project.

Strong changes in traffic demand are mainly expected in dense urban areas where huge number of users are present. It is not unusual to have high building in these areas which will obstruct the radio signal propagation, and demand installation of additional RUs in order to secure a good coverage. This will mean using additional RUs just to secure coverage, and not capacity. To reduce the power consumption of the RU, in these cases, a RIS can be used. The RIS will secure the coverage in the obstructed areas, but will not increase the capacity, as well as the power consumption. The ISAC functionality of the network can be used in these scenarios to detect the obstacles and the users, which will be used to better control the RISs, and to help estimating the optimal positions for RIS installation.

4.1.1.4 RIS network energy efficiency perspective

There are several previous works that consider RIS to improve the energy efficiency of the network. The work in [53] explore resource allocation in a wireless communication network with distributed RIS. The network uses multiple RIS strategically placed to serve wireless users and aims to maximise energy efficiency. This is achieved by dynamically controlling the RISs' on-off status and optimising their reflection coefficients matrix while ensuring minimum user rate constraints. The problem is framed as a joint optimisation of transmit beamforming and RIS control. In [54], the use of RIS for improving downlink multi-user communication from a multi-antenna base station is investigated. The focus is on developing energy-efficient strategies for allocating transmit power and configuring phase shifts for the RIS elements, while ensuring individual link budget guarantees for mobile users. To solve the resulting non-convex optimisation problems, the paper proposes two computationally efficient approaches, involving alternating maximisation, gradient descent, and sequential fractional programming.

RIS have also been applied in mobile edge computing and Non-Orthogonal Multiple Access (NOMA) networks. For example, [55] investigates the use of RIS in a single-cell multi-user Multi-Access Edge Computing (MEC) system. In this setup, a RIS is deployed to enhance communication between a BS equipped with MEC servers and multiple single-antenna users. To efficiently utilise limited frequency resources, the paper assumes NOMA communication, where each user has computation tasks that can be computed locally or partially (or even fully) offloaded to the BS. The primary objective is to minimise the total energy consumption of all users, involving joint optimisation of passive phase shifters, transmission data size, transmission rate, power control, transmission time, and decoding order. Moreover, the authors in [56] explore the potential of combining mMIMO and NOMA techniques to enhance connectivity in future wireless networks. The integration of mMIMO and NOMA offers improved spectral efficiency and reduced communication latency. However, the inherent variability of wireless channels can still impact system performance. To address this challenge, RIS have emerged as a promising solution. RIS, which allows for software-controlled manipulation of propagation channels, offers benefits such as increased data rates, enhanced user fairness, and potentially higher energy efficiency in communication networks.

Regarding device-to-device (D2D) communications, [57] addresses the joint optimisation of power control for D2D users and passive beamforming of RIS in an RIS-assisted D2D communication network, with the goal of maximising energy efficiency. The optimisation problem at hand is non-convex and is divided into two separate subproblems: passive beamforming and power control, which are optimised alternately. The approach begins by decoupling the passive beamforming at the RIS using the Lagrangian dual transform. This problem is then solved through fractional programming. Next, the power control is optimised using the Dinkelbach method.

In Unmanned Aerial Vehicle (UAV) wireless networks, RIS can play an important role in achieving energy

efficiency. In particular, the authors in [58] present a novel framework for integrating RIS into UAV-enabled wireless networks. The RIS is deployed to enhance the service quality of the UAV, and NOMA technique is employed to boost spectrum efficiency. The scenario considers mobile users (MUs) in continuous motion. The primary objective is to minimize energy consumption by jointly optimising the UAV's movement, RIS phase shifts, power allocation from the UAV to MUs, and determining the dynamic decoding order.

Another example is [59], which focuses on enhancing the energy efficiency of UAV-enabled Wireless Power Transfer (WPT) systems, particularly when deploying multiple ground sensors. The challenge lies in optimising the energy consumption of the UAV while meeting the energy needs of each sensor. To address this, the article proposes the use of a RIS. The approach involves a "fly-hover-broadcast" (FHB) protocol where the UAV emits RF signals only at specific hovering locations. The optimisation problem is formulated to minimise the total energy consumption of the UAV, considering its trajectory, hovering time, and the RIS's reflection coefficients. Additionally, the article explores a more general scenario where RF signals are transmitted during the UAV's flight. The goal here is to minimise the UAV's total energy consumption by jointly optimising the UAV's trajectory, flight time, and the RIS's reflection coefficients.

In [60], the problem of maximising energy efficiency in an uplink wireless communication system aided by a RIS is addressed. In this system, a UAV equipped with a RIS serves as a mobile relay connecting a base station (BS) to a group of users. The focus is on optimising the system's secure energy efficiency by jointly optimising the UAV's trajectory, the RIS's phase shift, users' associations, and transmit power levels.

Finally, the authors in [61] address the problem of maximising energy efficiency in a downlink visible light communication (VLC) system aided by a RIS. The objective is to optimise time allocation, power control, and the phase shift matrix to achieve this goal, while considering unique power constraints specific to VLC. To tackle this non-convex energy efficiency maximisation problem, the paper first simplifies it into an equivalent problem with fewer variables. Then, it proposes an alternating algorithm with low complexity to obtain a suboptimal solution. This algorithm iteratively solves the joint time allocation and power control subproblem and the phase shift matrix adjustment subproblem.

4.1.2 BeGREEN approach on ISAC

The development of wireless communication systems was always focused towards maximizing the wireless channel capacity and minimizing transmitted power. Minimizing transmit power was mainly associated with the complexity and price of high-power transmitters. With the introduction of cellular networks and with the increase of their density, it became obvious that the radio segment of these networks started playing a significant role of the overall energy footprint of the overall network.

To reduce the energy footprint of the overall network, different strategies are pursued. One of the strategies that BeGREEN is pursuing is to use ISAC and RIS for improving the energy efficiency on the RAN as well as on the overall network.

BeGREEN will initially investigate the possibilities of using ISAC information for improving physical layer energy efficiency. This will address mainly optimization of the beam training algorithms. The primary goal will be to reduce the wireless medium usage during the beam training phase. The secondary goal will be to offer an efficient solution for beam training, which, in turn, will allow for higher usage of systems supporting beam forming and beam tracking. These systems generally need less transmit power for a given link budget, therefore increasing their energy efficiency.

Nowadays, different strategies for improving overall network energy efficiency are emerging. They are implemented at a network level and tend to optimally assign network resources where needed. Different approaches are used and many times they involve ML/AI based algorithms. All these approaches will require detailed information on the spatial user density to optimally assign network resources. BeGREEN will

investigate how ISAC can be used to provide user density information to the network. Additionally, the provision of other type of information deemed necessary for classification of the sensed objects, will be investigated. This should improve the network ability to perform optimization of the energy efficiency by optimal resource assignment of where needed.

Moreover, BeGREEN will investigate the use of RIS for improving the network energy efficiency. The RIS in combination with ISAC can be used to extend the network coverage and replace the use of additional RUs where the almost passive RIS can be deployed. BeGREEN investigates how sensing data acquired using ISAC can be used to optimise the network coverage using RIS, reducing at the same time the energy consumption.

Finally, BeGREEN will address improvements of the RU hardware which will enable better energy efficiency. In the past, energy efficiency of the RU was not a main concern. Therefore, in order to keep the price of the RUs competitive, no additional hardware was built-in for improving the energy efficiency. This limited the possibilities for energy efficiency optimisation of the RUs, but it kept their price low. Nowadays, the hardware becomes cheaper, and the energy prices steadily increase. This compels equipment manufactures to change strategies and to optimize for energy efficiency even if this demands for more complicated hardware. BeGREEN will address this problem and will investigate how additional energy saving at the RU can be secured, even if additional hardware is needed.

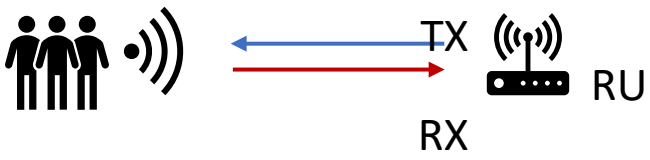
4.1.3 ISAC aided strategies for improvement of energy efficiency

As already discussed, future 6G networks should implement sensing functionality within the same RUs used for data transmission. This sensing functionality should give the network a perception of the environment and will be an enabler for deploying different energy optimisation strategies. The ISAC functionality of the network will allow for sensing of the surrounding environment and, depending on the sensing approach, different features of the environment, i.e. the objects in it, can be detected. This information can be used to optimise some of the functions of the network to improve its energy efficiency or, alternatively, they can be used by third parties for energy efficiency optimisation of other systems, i.e., traffic optimisation based on expected load.

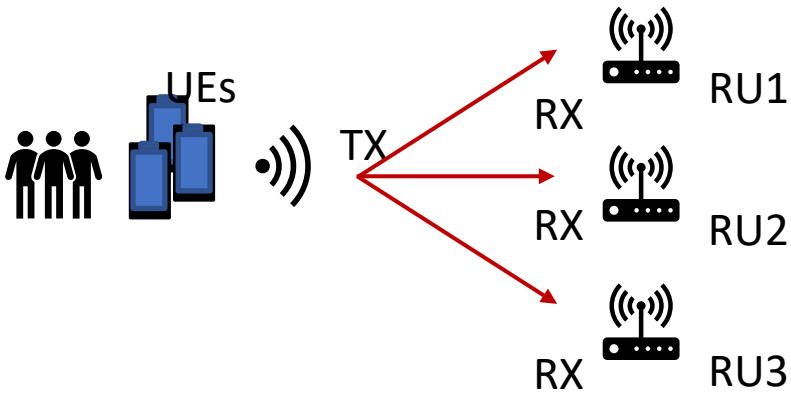
In the framework of BeGREEN, a few different aspects of the ISAC functionality will be investigated. First, the ISAC PHY-layer supporting ISAC, namely the waveforms that can be used for sensing, the possibilities for multiplexing of sensing and data transmission signals, as well as the optimal sensing schemes will be investigated, proposed, and evaluated. Second, improvement of the energy efficiency of some of the existing PHY-layer functions will be investigated. The focus will be towards improving the beam training and beam tracking functionality of phased antenna array data communication systems. These functions can be relatively complex, and their potential is, to date, rarely fully utilised. Finally, the creation of a digital model, i.e. digital twin, of the real-world environment will be pursued. The main idea is to create a real-time model of the environment that can feed different algorithms to perform network wide energy efficiency optimisation.

4.1.3.1 ISAC waveforms, multiplexing and sensing schemes

The sensing functionality of future mobile networks can be performed in a few different ways. One of the approaches is to sense the environment in a similar fashion as radar does. This means that the reflections of the transmitted signals are used to detect different objects, obstacles, and persons. Different features of the received reflections can be used to detect the objects' type, size, position, velocity, etc. This is usually not straightforward, but different ML/AI algorithms are lately being investigated to facilitate precise classification of the object based on the received signal properties. This is usually used in different radar sensing approaches.



a. Deviceless sensing



b. Device based sensing

Figure 4-1 a. Deviceless sensing; and b. Device based sensing

The second approach is to sense the incoming signals being transmitted from the UE. This approach is usually called device-based sensing [79]. With this approach, the positions of the UEs can be easily estimated, and they can be used for beam management or for optimisation of the network coverage/capacity, leading to increasing energy efficiency. Both approaches are shown in Figure 4-1.

Different waveforms can be used for ISAC. A logical approach is to use a radar waveform, like Frequency Modulated Continuous Wave (FMCW) or similar. These waveforms are usually used in radars to simplify the radar hardware and to lower the production costs. Nevertheless, a wireless communication system has all the necessary hardware to generate different waveforms. Therefore, within BeGREEN a few different waveforms will be evaluated and their applicability and suitability for sensing will be investigated. Additionally, suitability for sensing of transmission frames used in different wireless technologies will be investigated. These frames usually have a preamble, which has a predetermined waveform and a data field that is different for every frame. Both can be used for sensing. The preamble is usually designed with a good autocorrelation function, consisted of a strong peak for $\tau=0$ and low sidelobes for $\tau\neq0$. These makes the preamble an ideal waveform for sensing. Its main disadvantage is that it is relatively short, making it not the best candidate in low SNR scenarios. The data field, on the other hand, is much longer than the preamble, meaning that the autocorrelation peak at $\tau=0$ can be larger, hence more suitable for low SNR scenarios. This will enable sensing with wider radiation patterns (i.e. beams) having lower EIRP compared to narrow radiation patterns (i.e., beams). This is important since using wider radiation patterns for sensing will allow to sense a larger portion of the environment eliminating the need of scanning the environment with a small beam.

To accurately sense the environment and to minimise the artefacts produced by the sensing, it is important to have low side lobes in the autocorrelation function of the waveform. Low sidelobes means that the waveform is completely random, i.e. no correlation between the different samples of the waveform exists. For the preamble of the wireless frame, this can be assumed to be the case in many systems. Nevertheless, for the data field this must not be the case. In a more complex and optimised wireless systems, the source encoder has the function to remove the redundancy from the source signal, and to uniformly distribute the bits and symbols obtained from the data to be transmitted. Nevertheless, the channel coder in these wireless

systems adds redundancy, needed for forward error correction, which will make this waveform non-optimal for sensing. Anyway, even not optimal, this waveform can still be used for sensing.

BeGREEN will investigate the properties of these waveforms and their suitability for sensing. Additionally, PHY-layer multiplexing schemes will be investigated as well as their suitability for sensing. Finally, new multiplexing and sensing schemes enabling ISAC will be proposed.

4.1.3.2 PHY-layer function optimisation

The sensing function of a system supporting ISAC can bring a vast amount of information about the radio propagation properties of the surrounding environment. This information can be used to predict different parameters like the available link budget from different RU to the UE, how the link budget will change over time, to predict the coverage of the network based on the expected demand, etc.

BeGREEN will address some of these aspects and will propose solutions for improvement of the PHY-layer functions that will, in turn, enhance the energy efficiency of the RUs.

One of the aspects that BeGREEN will address is beam training (i.e. beam search) and beam tracking in systems having multiple antennas, i.e. phased antenna arrays, and supporting beamforming. Beamforming is a technique commonly used in mmWave systems due to the sparsity of the channel, i.e. when few or no multipath components are present. This is the only option for increasing the channel capacity, by increasing the EIRP and, at the same time, allowing for large coverage by electronically pointing the beam in different directions. Using MIMO for mmWave systems is usually avoided, since minimal channel capacity increase is expected due to the sparsity of the channel.

In the sub-6 GHz bands, a multipath rich environment is expected, especially indoors and in densely populated areas. In this band, MIMO is the perfect candidate for increasing the channel capacity of the system. Nevertheless, for mobile networks, beamforming is still of interest, since sectors of coverage can be defined and the capacity per unit of area can be controlled. This means, that in areas where high capacity per unit of area is needed, many RU can be deployed covering only smaller areas. Having same capacity per RU and smaller coverage, will naturally increase the capacity per unit of area. Additionally, the possibility for beamforming at the RU allows for dynamic change of the coverage and optimal adaptation of it depending on the current demand. Furthermore, this can be used to minimise interference with other RUs, which will additionally improve the energy efficiency.

Beamforming methods are already well known, and the current hardware can easily support them. Nevertheless, their use in data communication systems is relatively rare, especially in the sub-6 GHz systems. The main limiting factor is the complexity of the beam training procedure. Namely, the beam training in 2-dimensional (2D) case has a complexity of n^2 and in 3-dimensional (3D) space (i.e. beam steering can be performed in azimuth and elevation) a complexity of n^4. This is if exhaustive search is performed. In case of hierarchical beam search, the complexity will be $(log(n))^2$ in 2D case and $(log(n))^4$ in 3D case.

This means that many combinations of transmit and receive beams should be tested in order to find the most optimal one. Using digital beamforming, this procedure can be parallelised, but it will face additional issues. Therefore, for sub-6 GHz applications MIMO is the usual choice, especially in Wi-Fi applications where the cells are typically within the boundaries of a single house or apartment. Since the Wi-Fi access points (APs) have large enough capacity, and their duty cycle is relatively low, not a huge interference is expected between neighboring devices.

In the case of a mobile network, the RUs are more expensive and more effort is put on optimisation and minimising the number of deployed devices. Additionally, the interference in this case will play significant role. Therefore, different techniques are used to minimise the interference between the neighboring cells.

One approach for addressing the interference problem and for maximising the energy efficiency is to use beamforming. With beamforming, narrow beams can be used to minimise interference to the other users and, at the same time, due to the high gain of the phased antenna, low transmit power can be used to achieve high EIRP and large coverage. This is shown in Figure 4-2. As can be noticed, instead of having one wide beam covering a large area, one can have multiple narrower beams with the same EIRP, which also means less transmit power. It can be assumed that the area of the beam is proportional to the transmit power. Nevertheless, for the link budget, only the EIRP is important and not the beam width. Additionally, these narrower beams can be multiplexed in time, which will introduce additional energy savings.

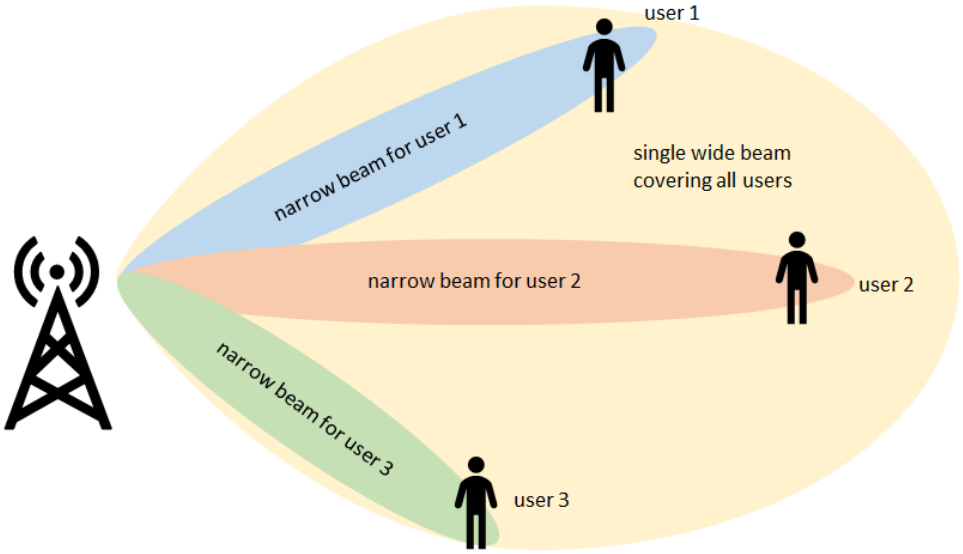


Figure 4-2 Comparison of single beam and multi beam coverage

Beamforming in sub-6 GHz is usually not used massively due to the complex beam training process, large antenna array size, etc. Additionally, the beam training process becomes complicated in mobile scenarios, where the users are constantly moving, and a mechanism for tracking the users must be used.

BeGREEN will address these issues using ISAC. With the help of ISAC, the positions of the users will be estimated and this information will be used to optimise the beam training process as well as to optimise the user tracking. The solutions that will be developed in BeGREEN target more efficient beam training and beam tracking processes, which will increase the EIRP of the transmitter, leading to less transmit power and lower power consumption. This will be a tradeoff between power consumption and the complexity of the system.

It is assumed that the added complexity of the system will not increase the overall power consumption, because for the integrated sensing functionality the existing wireless communications hardware will be used.

It can be argued that, using MIMO, the same link budget can be achieved using the same transmit power. Nevertheless, this is only possible if a rich scattering environment is available. Additionally, not all UE are large enough to be fitted with multiple antennas spaced apart at a meaningful distance. Therefore, even the MIMO approach can give some comparable results to the approach proposed here, it has strict requirements, not easily attainable in every environment.

4.1.3.3 Digital twin of the environment

Digital twin is a digital representation of an object, system, process or of an environment, created and used for modelling their properties, simulation, planning, etc. The sensing functionality of a network supporting ISAC can be used to reconstruct a digital twin of the environment. This digital twin must not represent only the physical properties of the environment, e.g. buildings and objects with their exact physical dimensions, but it can also represent the electromagnetic RF propagation properties. Using the digital twin, an RF

propagation model of the system can be built and later used for different purposes.

In BeGREEN, the creation of a digital twin based on the sensing data acquired from the sensing network will be investigated. The focus will be a digital twin representing the radio propagation properties of the environment. The spatial distribution of these properties should also be estimated to perform optimal network coverage for improving the energy efficiency of the radio network. In BeGREEN the precise geometrical properties of the environment are not of interest. Nevertheless, having in mind that these are in correlation with the RF propagation, some of them can be estimated with the sensing data obtained. Reconstructing the geometrical properties of the environment will require a bit different approach, because some of the objects are transparent for radio waves and, additionally, large bandwidths are needed for precise detection and reconstruction of the features of the objects.

4.1.4 Self-configuring RIS

RISs enable a fully controllable and flexible smart propagation environment by means of meta-surfaces composed of multiple patch antennas that alter the radio propagation properties of the impinging signals in favour of specific directions. However, in many RIS envisioned use cases an ad-hoc control channel is essential to dynamically reconfigure the propagation of radio waves based on current network dynamics. Additionally, due to the absence of signal processing units on such nearly passive surfaces, channel estimation and acquisition are usually performed over the entire transmitter-RIS-receiver path, which may prevent the agile deployment of the surfaces.

Imposing stringent requirement on these nearly passive envisioned devices would hinder its deployment and its ability to be seamlessly integrated into existing networks and standards [68]. An agile deployment and easy configuration of RISs are essential features for the openness of the RAN [67]. Therefore, we seek to develop reconfigurable devices that can be seamlessly plugged into the existing network infrastructure without requiring sophisticated installation procedures and that can autonomously play to enhance communications KPIs.

In what follows, we present 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.

We consider an HRIS [69] comprising an array of hybrid meta-atoms, which can simultaneously reflect and absorb (i.e., sense the power of) incident signals. In the considered architecture (see Figure 4-3), each metasurface element is coupled with a sampling waveguide that propagates the absorbed (i.e., sensed) power of the incident electromagnetic (EM) waves towards some downstream RF hardware for enabling signal processing.

To reduce the complexity and cost of the required hardware, the proposed plug-and-play (P&P) HRISs are not equipped with fully-fledged RF chains but only with an RF power detector, thereby eliminating the need for a receiver. As shown in Figure 4-3, the signals sensed by each metasurface element are summed together by RF combiners, which may be easily implemented as lumped components throughout the metasurface RF circuit [70]. The resulting signal is fed into an RF power detector that converts the RF power into a measurable direct current (DC) or a low frequency (LF) signal, which is, e.g., made of a thermistor or a diode detector [71][72]. In the considered hardware architecture, the reflected and absorbed signals are subject to the same phase shifts applied by the metasurface elements, which are tuned to simultaneously control the signal reflection and power absorption properties of the HRIS. Nonetheless, it is possible to enable the fully absorption operating mode by deactivating the reflection of signals by means of simple varactor diodes [73].

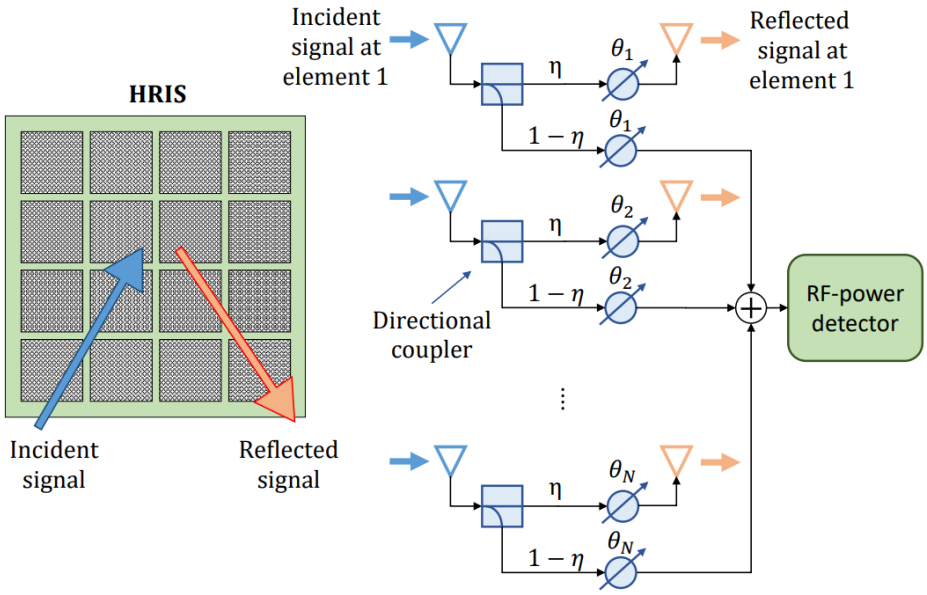


Figure 4-3 Reference diagram of a hybrid reconfigurable intelligent surface

As shown in Figure 4-3, the proposed HRIS design includes only one RF power detector, which can only measure the power of all the incident signals at every HRIS meta-atom. Most available angle of arrival (AoA) estimation techniques necessitate the signal samples at each receive antenna. Conversely, we make the most out of our proposed hardware design and perform an indirect estimation of the AoA, by optimising the phase shifts applied to the absorbed (sensed) signals at every meta-atom so as to maximise the power sensed by the detector. As adjusting the phase shifts applied by the meta-atoms is equivalent to realising a virtual (passive) beamformer towards specified AoA of the incident signals with respect to the HRIS surface, we can take advantage of the power sensing capability of the HRISs for estimating the BS-HRIS and HRIS-UE channels with very little local information.

We anticipate that any algorithmic solutions for optimising the HRIS require the estimation of the CSI of the BS-HRIS and HRIS-UE channels. This results in a chicken-egg problem that needs to be tackled. To this end, we devise an online optimisation approach that relies upon a finite set of HRIS configurations, namely a codebook that can be iteratively tested for probing a finite set of predefined AoAs. It is worth mentioning that this operation may have a disruptive impact on the network operation: a given HRIS configuration may be in use for assisting, through smart reflections, the data transmission of some BSs and UEs. Therefore, changing the HRIS configuration for sensing may negatively affect the communication performance. MARISA addresses this issue by means of a simultaneous hybrid probing and communication scheme.

4.1.5 PHY layer and RIS/ISAC aided network level energy consumption optimisation

In an O-RAN framework, one of the challenging roles that the CU/RIC environment possess is the ability to collect information from lower layers and convert to manageable data in intelligent layers. ISAC and RIC are new developments whose interfaces are not standardised at the moment, hence there is no clear interfaces within the O-RAN architecture that indicate how the information is transmitted to higher layers. With this in mind, new standardised messages based on the new capabilities imposed by the ISAC/RIC radio environment are needed to control and enhance the radio environment. This is the case of mMIMO, where in signal processing data needed for the proper MIMO codebook configuration, need to be compressed and exchanged to the RIC. This highlights the challenges in optimising beam parameters and codebooks, emphasising the need for data-driven solutions. In addition, it is important to establish that determining optimal offsets for mobility configuration and in grouping users for mMIMO in ISAC/RIC is needed for optimal performance and low energy consumption. It is needed to include potential issues related to policy guidance

and data enrichment for user grouping needed in ML/AI based xApps. Other open issue is the relatively unexplored dynamic, data-driven reconfiguration of beamforming with RICs in O-RAN literature. These challenges underscore the complexity of implementing RIC-controlled beamforming in O-RAN networks.

PHY layer radio enhancements from the CU perspective play a pivotal role in orchestrating radio resources efficiently and can be summarised into the topics described next. Dynamic Power Control (DPC) algorithms can be harnessed within the CU's capabilities to intelligently adjust transmit power levels based on prevailing channel conditions. By analysing and processing real-time data, the CU optimally allocates resources to ensure that the minimum necessary power is employed for reliable communications. Moreover, the CU's ability to implement adaptive transmission modes is crucial. These modes dynamically adjust modulation and coding schemes in response to changing channel conditions, thereby promoting energy efficiency. With this adaptive approach, the CU maximizes spectral efficiency, allowing for higher data rates while minimising the energy expended per transmitted bit. Additionally, the CU should possess the capability to select an optimal sleep mode and utilize power scaling techniques, enabling components to operate in low-power states during periods of low activity. By efficiently managing power utilisation, the CU significantly contributes to enhancing energy efficiency at the PHY layer.

For the RIC xApps and rApps, optimising energy efficiency at the PHY layer hinges on dynamic resource allocation and advanced scheduling. Leveraging the RIC's intelligence, xApps and rApps can employ dynamic resource management strategies based on traffic demand and channel conditions. This ensures that resources are allocated judiciously, minimising unnecessary power consumption during low-traffic intervals. Additionally, energy-aware scheduling algorithms can play a pivotal role. By prioritising scheduling for users and services with higher energy efficiency requirements, xApps and rApps can fine-tune resource allocation. This not only enhances energy efficiency but also guarantees that critical services receive the necessary resources. Harnessing ML/AI, xApps and rApps can predict traffic patterns and make real-time adjustments in transmission parameters. This proactive approach ensures that the network optimally utilizes its resources for energy-efficient communication. By collaborating closely with the CU, xApps and rApps contribute significantly to achieving enhanced energy efficiency in the PHY layer, ultimately leading to a more sustainable and high-performing wireless network.

4.2 RU power optimisation

Optimising power consumption in the RU is one of the key areas to research in O-RAN networks. Most of the deployed RUs are usually working at peak output power. O-RAN proposes now mechanisms to accommodate the power to the demand without impacting service level agreements. In addition, there exist other PHY layer enhancements that can lead to power reduction applicable to RUs.

4.2.1 State-of-the-art status in RU power optimisation techniques

RF Power Amplifiers (PAs) are a key part in wireless communication systems, and it is crucial improving their efficiency and performance. In the Digital Predistortion (DPD) technique, a pre-distortion process precedes the PA, and the combined configuration of both exhibits improved linearity of signal gain. With pre-distortion the PA can run up to its saturation region while maintaining its linearity and improving its power efficiency. DPD operates in the digital baseband domain, yielding lower implementation complexity compared with alternative techniques. DPD provides cost-efficient way for nonlinear PAs to operate at higher output powers coupled with minimised distortion, resulting in greater power efficiency.

Articles discussing integrating AI with DPD were published several years ago. A few examples are listed below. DPD for 5G RF transmitters using machine learning was introduced [74]. Driven by the shift from single antenna to the multiple-input multiple-output (MIMO) phased array RF and signal bandwidth increase, significant challenges in managing power consumption and meeting linearity requirement of wireless

transmitters for such architectures were addressed. Machine learning techniques were used to resolve some of the issues in linearizing 5G MIMO systems.

The publication [75] presents an auto-tuning approach for dual-input PAs using a combination of global optimisation search algorithms and adaptive linearization in the optimisation of a multiple-input PA. By using heuristic search global optimisation algorithms, it is possible to find the best parameter configuration for PA biasing, signal calibration, and digital predistortion linearization to help mitigating the inherent trade-off between linearity and power efficiency. Machine learning strategies were applied to a dual-input PA to optimize configuration parameters, such as: bias voltages, input signal phases, and power splitting ratios considering a user-defined cost function. Another work [76] discusses using artificial intelligence (AI) providing opportunities to enable high-efficiency wireless communication to dynamically adapt to the local environments and user demands. The paper discusses a continual learning DPD algorithm proposing to linearize an RF PA in 6G AI-empowered wireless communication.

In [77], the wideband signal increases the memory effect in PAs, which requires a better DPD algorithms that can capture the temporal dependency of the PA. Deep neural networks (DNNs), such as recurrent neural networks (RNNs) and convolutional neural networks (CNNs), are promising candidates to enhance the performance of DPD at a lower cost compared to conventional memory polynomial methods. As for envelope tracking, RF PA envelope tracking is a power supply technique for improving the energy efficiency of RF PAs by tracking the power demand as opposed to today's fixed power systems.

To the best of authors knowledge combing of Envelope Tracking with AI was not investigated in the past. Therefore, it is assumed that the research presented in the following paragraphs is pioneering work in this field.

4.2.2 Problem statement

4.2.2.1 OFDM Signal Transmission and PAPR

OFDM symbol is a time domain (TD) signal generated using IFFT of many independent frequency domain (FD) sub-carriers. This is equivalent to summation of time domain wave functions. The result is a wideband time domain signal which is a summation of orthogonal frequencies. Time domain symbols are prefixed with cyclic prefix, concatenated, and filtered to produce the transmit signal (TxSig). This signal fluctuates in time and exhibits high peak values in random locations. Peak-to-Average Power Ratio (PAPR) is defined as the ratio between the peak value and the average of the signal.

$$PAPR = \frac{max\{|x_n|^2\}}{E\{|x_n|^2\}}$$

Statistically, PAPR increases with the number of active FD sub-carriers. For 4K IFFT with 3300 sub-carriers PAPR is about 12 dB.

Main sources to signal degradation are fixed point quantisation, symbol concatenation and non-linearity of the analogue components. This results in reduced capacity of the channel and interference to neighbouring RF channels (both in-band and out-of-band RF channels). The first two are handled digitally by adjusting the number of bits in digital computation and by either signal shaping or filtering the signal. Analog nonlinearity is more challenging and requires the amplifier to work in a linear region which is usually power inefficient. A DPD is proposed as a solution, however a DPD is more useful for narrow band signals with low PAPR than for wideband OFDM like signals with high PAPR.

The following text describes this issue in detail. What can be achieved with PAPR reduction methods which improve the operation point of the power amplifier with a DPD implemented using a DNN to restore linearity of the transmitted signal. Further, an envelope tracking scheme is proposed.

4.2.2.2 OFDM transmission

OFDM is a complex signal. It is modulated into RF signal using QAM, which is a real signal, then amplified. The power amplifier is not perfect, it shows nonlinearity, memory (hysteresis) effects and saturation (clipping) at max value. Nonlinearity of the power amplifier destroys the sub-carrier orthogonality, also known as inter-modulation products, and distribute its energy to neighbouring frequencies in the form of interference.

In QAM each component of the complex OFDM signal is modulated by a carrier wave and summed into RF signal. Each carrier wave has the same frequency but phase difference of 90°. The PA is therefore influenced by the absolute value of the complex OFDM signal.

4.2.2.3 Power amplifier

PA efficiency requires operating near saturation point where nonlinearity is high. Also, PAPR is statistical and high peaks may also appear and be clipped. Adjusting the working point is a trade-off between PA efficiency and signal integrity.

Figure 4-4 shows the effect of lowering the saturation point using Wiener model, where random noise of frequency domain constellation points and out of band emission in the spectrum are observed. Using Wiener model, memory effects can be simulated, where hysteresis as widening of the PA curve is observed, that results in noise distribution which is not random but like multipath as shown in Figure 4-5. With a perfect predistortion, there is no nonlinearity in the curve, but peak levels over the saturation point are clipped as depicted in Figure 4-6. I tis case out of band emission exist, and therefore, DPD is not effective (because of clipping). As shown in Figure 4-7 DPD is only effective if PAPR probability of clipping is effectively zero, which means high enough power back-off or reducing PAPR levels.

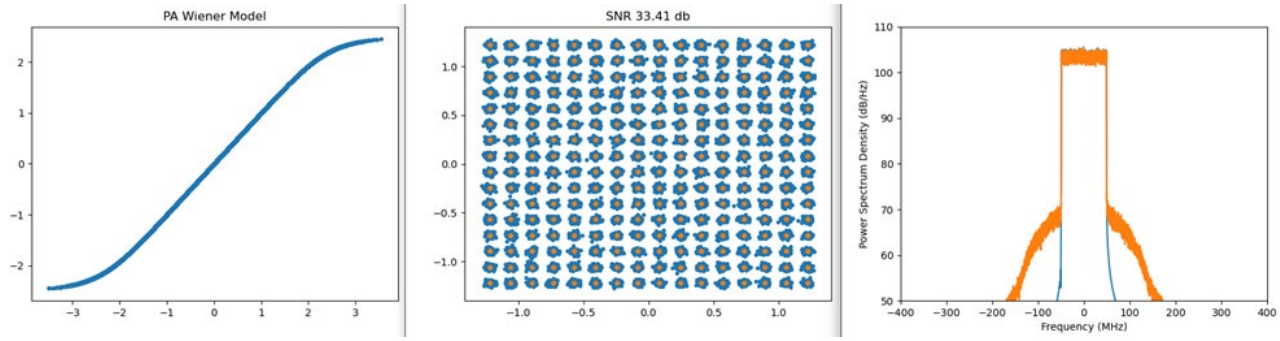


Figure 4-4 Performance of Wiener PA

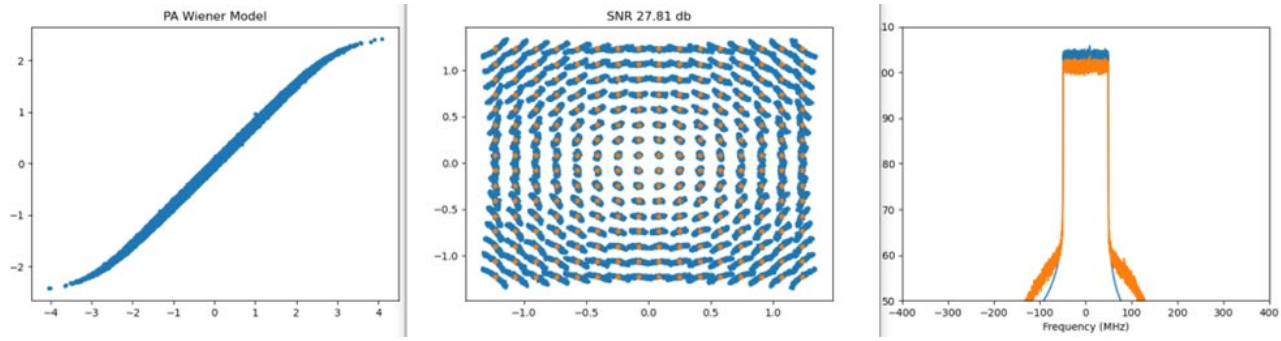


Figure 4-5 Performance of Wiener PA with memory

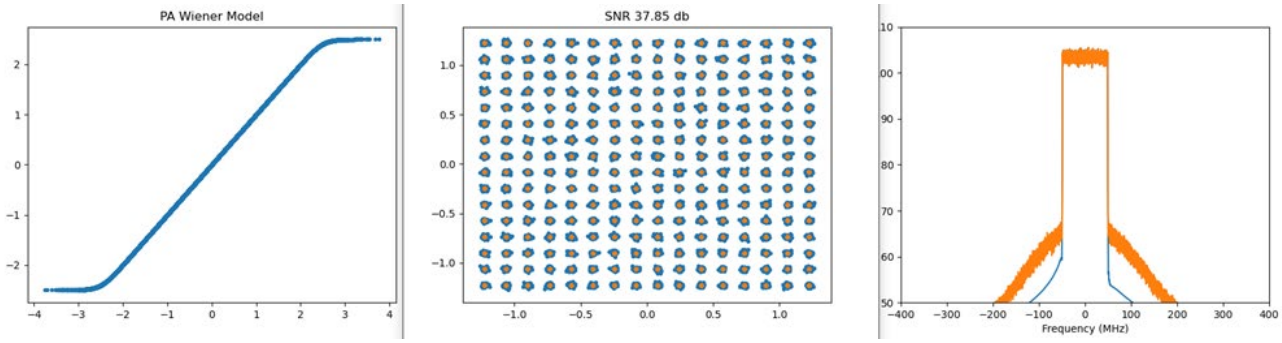


Figure 4-6 Performance of saturated PA

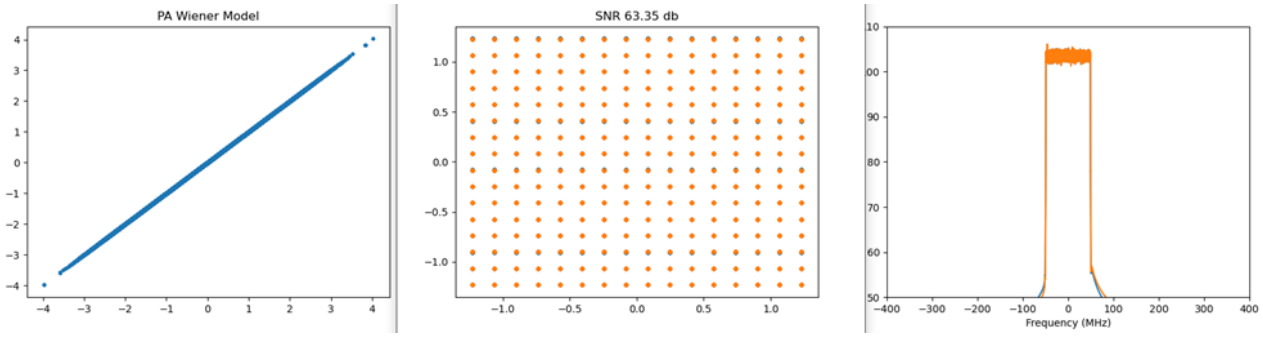


Figure 4-7 Performance of perfect PA

In conclusion, nonlinearity results in inter-modulation products that shows up as wideband noise. One can see this noise as out of band emission and noise around the constellation points. Nonlinearity may result from imperfection in the signal generation or PA nonlinearity and saturation. DPD cannot handle clipping. In OFDM a peak is not a rare phenomenon that can be ignored. Any solution is a trade-off between lowering PAPR, adjusting the saturation point and correcting nonlinearity with a DPD. Each one is either limited in performance or cost in power consumption. Power envelope tracking is proposed as a solution to reduce dependency on PAPR and saturation.

4.2.3 Proposed improvement strategies in RU power optimisation

4.2.3.1 PAPR reduction methods

The motivation is to lower the PAPR in a controlled way in order to reduce power back off requirements, with managed degradation to signal integrity. PAPR can be handled in time domain, frequency domain or both, or even by several time and frequency iterations. PAPR value is also statistical, so trying several symbols with the same information and choosing the best is another way around. Pure frequency domain methods are methods that reduce PAPR by manipulating the FD sub-carriers. These methods will preserve spectrum, for example by allocating dummy sub-carriers (cost in bandwidth) with values that lower the PAPR. In general, it is hard to predict the PAPR from FD. Therefore, several FD to TD transforms is needed until a reasonable PAPR is achieved.

Pure time domain methods reduce PAPR in time domain, but their effect on the frequency domain is arbitrary and generally produce inter modulations. Combined, time and frequency domain methods with iterations in between is also possible. This is achieved by hanging both time and frequency properties in several iterations, until properties in both domains are satisfied. This comes with the cost of many transforms' computation and time delay. It is reasonable to expect about 3dB of PAPR reduction in most methods, with a reasonable trade off.

4.2.3.2 Digital Pre-distortion (DPD)

DPD is a technique that compensates for RF PA's nonlinearity, allowing it to operate in its nonlinear region for maximum power efficiency. PAs are essential components in wireless communication systems, but they are inherently nonlinear. This nonlinearity generates spectral regrowth, leading to unwanted radiation and adjacent-channel interference. It also causes in-band distortion, resulting in degradation of its error vector magnitude (EVM) performance. To improve EVM, a PA's operating point needs to be backed off far from its saturation point, leading to very low power efficiency, typically less than 10%. DPD provides an effective method to linearize PAs. DPD lets cost-efficient nonlinear PAs run in their nonlinear regions with minimised distortions, resulting in higher output power and greater power efficiency. The concept is based on inserting a nonlinear function (the inverse function of the amplifier) between the input signal and the amplifier to produce a linear output. The DPD must adapt to variations in PA nonlinearity with time, temperature, and use of different operating channels.

A DPD is expected to inverse the imperfection introduced by the PA so that the overall signal is closer to the perfect signal. The DPD is a digital time domain function. Inputs and outputs are complex numbers. The PA input on the other hand is QAM modulated real signal. And the PA transfer function is also real. Distortion of the PA is a wideband modulated signal with envelope proportional to the complex time domain signal. The DPD can therefore only correct the envelope's amplitude. This is a limitation that prevents achieving perfect DPD. A DPD need to learn the PA behaviour for all input signals. The PA characteristics may also change in time as result of e.g. changing temperature. Based on feedback from the PA, an inverse function may be calculated and implemented in different ways (e.g., using polynomials). A new approach is to use a DNN (Deep Neural Network).

In both cases the learning process function is defined as follows:

x is the source signal

y is the feedback signal, y = f_PA(x)

f is the DPD inverse function, f(y) = x

and the inverse function is implemented by the DNN.

A DPD training needs samples from the input of the PA and from the output. Both inputs and output samples are complex and digital. Input to the PA is therefore, the transmitted digital samples. Output from the PA is a feedback which is demodulated and converted back, as a received signal, to the base-band back to a complex digital signal (see Figure 4-8).

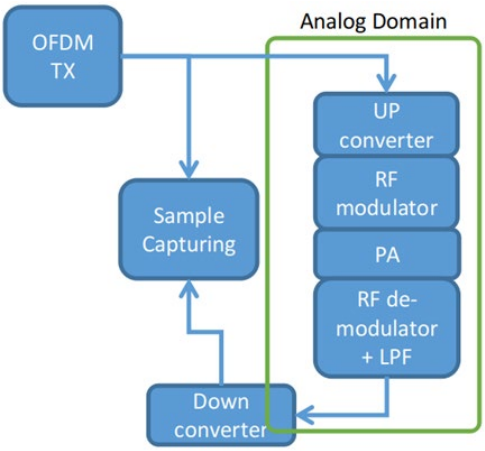


Figure 4-8 DPD training setup

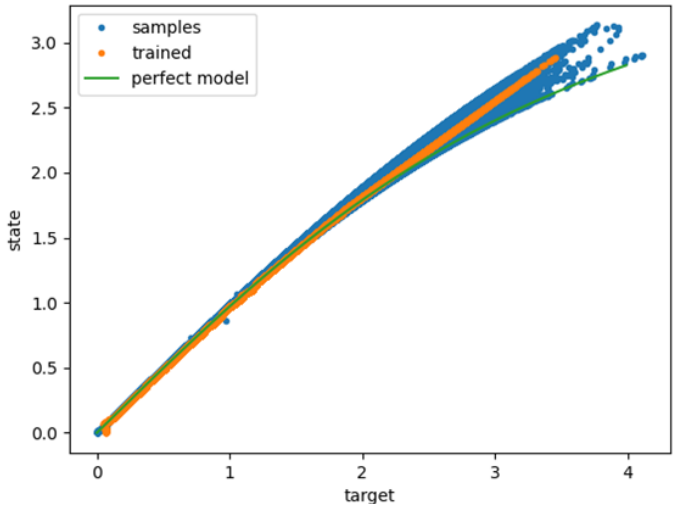


Figure 4-9 State-target training samples versus perfect and trained model

The sampling process involves matching between Tx and Rx samples. However, Rx samples are not unique in the sense that the same Tx sample may generate different Rx samples. This results from the phase of the carrier wave which is random at the sample instance.

The PA is working with continuous real analogue signal, where the carrier wave oscillates subject to the signal power envelope. The PA is due to both the signal amplitude and the carrier amplitude at each instance. This results in widening the curve with correlation to the PA distortion at that amplitude as shown in Figure 4-9. The figure also shows that simple training with all samples will produce some average curve for the state, while the perfect model follows the lower state samples of each input (target in the figure). This means that training is less effective for the upper part of the curve.

One should also note that digital signal bandwidth become wider at the DPD output and filtering may reverse the DPD operation. The DPD is working better with high sampling rate, meaning that it is better to up convert the signal before DPD than after. This requirement is limited by the digital capabilities of the system.

Following is an example of training a DNN to act as a DPD. This simulation uses a Rapp model with $p = 1$, and $A_{sat} = 5$ *(14 dB)* as the PA model and an OFDM signal with mean value of. The DNN is implemented using 2 hidden layers of 16 neurons in each layer and RelU activation.

Where Rapp model is:

$$F_a(|x(t)|) = \frac{|x(t)|}{[1 + (|x(t)|/A_0)^{2p}]^{\frac{1}{2p}}}$$

And A_{sat} is A_0 in the formula.

Interference is translated to random noise around constellation points, and out of band emission in the spectrum. Using the trained model has reduced the interference (see figures below). Out of band emission in the spectrum is reduced by about 6 dB. SNR of the constellation points is also improved similarly.

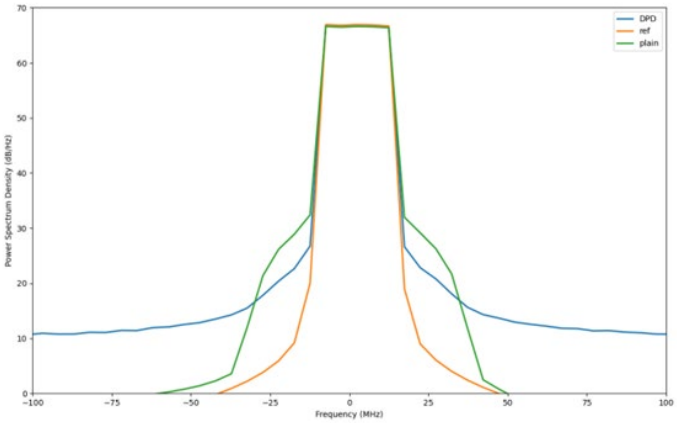


Figure 4-10 Signal spectrum for Asat = 5 of perfect reference signal compared with plain PA output and DPD output

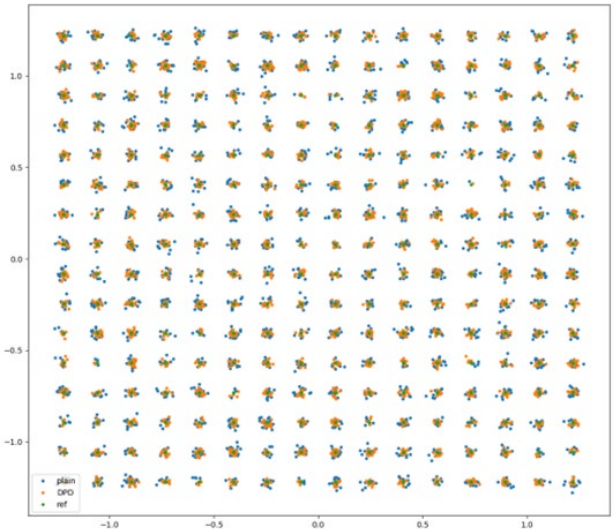


Figure 4-11 Constellation points for Asat=5 of reference, plain and DPD

Another example shows the effect of clipping. By changing A_{sat} to 3 (9.5 dB) the saturation point becomes lower, and clipping is more pronounced. Here the DPD is less effective and improvement to interference products (out of band emission and constellation SNR) is only about 3 dB less.

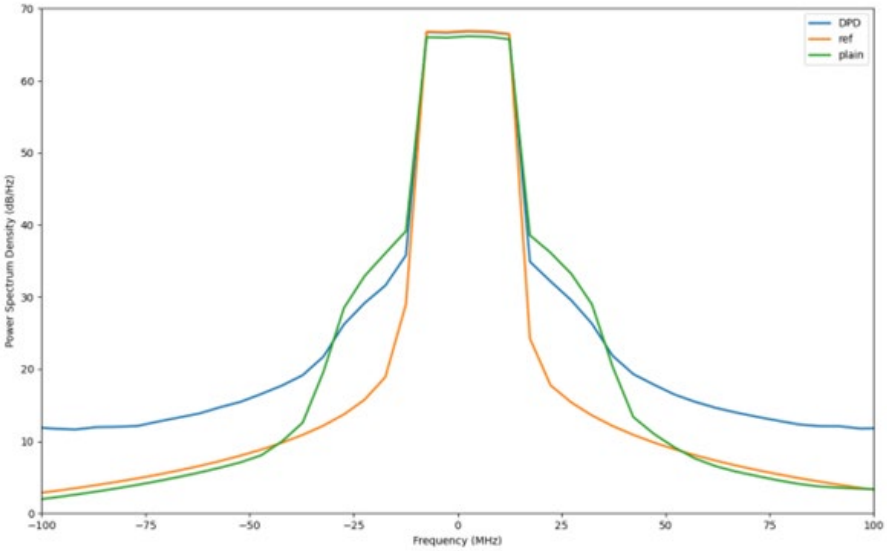


Figure 4-12 Signal spectrum for A_{sat} = 3 of perfect reference signal compared with plain PA output and DPD output

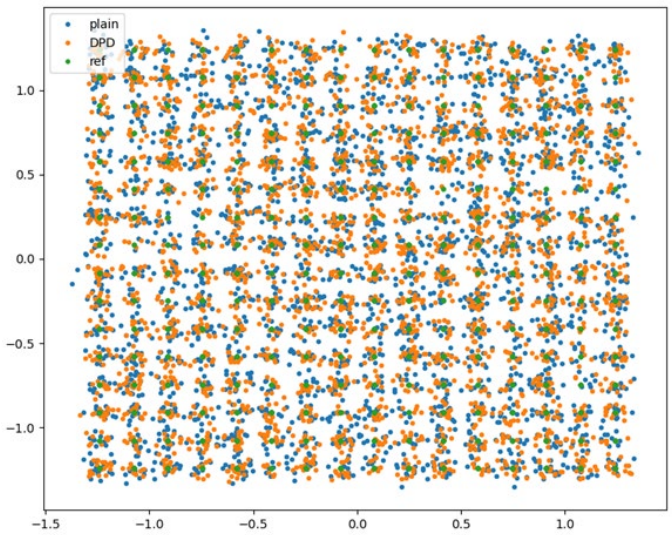


Figure 4-13 Constellation points for $A_{sat}=3$ of reference, plain and DPD

Conclusion is that DPD can be implemented with a DNN with only about 16 neurons and 2 layers which may be considered small. However, when approaching the saturation point, the random carrier phase effect is more pronounced and reduces the PDP performance. Changing the size of the DNN doesn't bring any improvements. It is also noted that a DPD amplifies the signal and therefore the DNN implementation will need more digital resources, i.e., larger multipliers, wider buffers, etc.

4.2.3.3 Power envelope tracking

So far, we have shown that PAPR reduction can improve PA efficiency. A DPD can only restore linearity. Clipping will be also minimised. For 5G OFDM signal with about 3300 sub-carriers the PAPR is about 12 dB which can be reduced with a good PAPR reduction method to about 9dB. The exact target PAPR may also depend on the level of signal degradation which is applicable in specification. Also, some guard from saturation is needed to minimize non-linear effects that cannot be reversed by DPD.

When power tracking is used, the power feed to the PA is not constant but varies with respect to the signal envelope. The signal power envelope in its simple form is the absolute of the time domain OFDM signal. Convolution of the envelope with shaping filter is also possible. The envelope is calculated for each Tx antenna independently. With a perfect power tracking and a perfect DPD it is expected to achieve PAPR = 1. In practice, for the same reasons as before, a guard from signal saturation is needed. This has the potential to save up to 10dB of transmitter power consumption. OFDM RF signal and its power envelope are depicted in Figure 4-14.

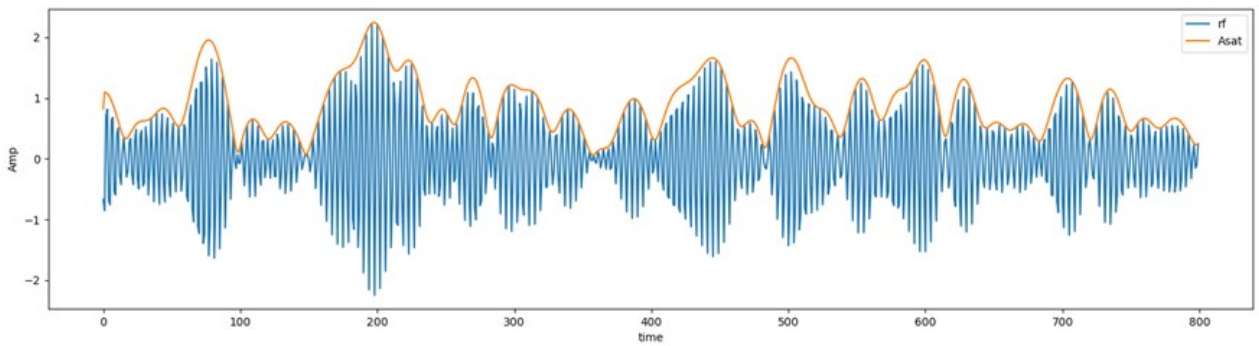


Figure 4-14 RF signal and power envelope

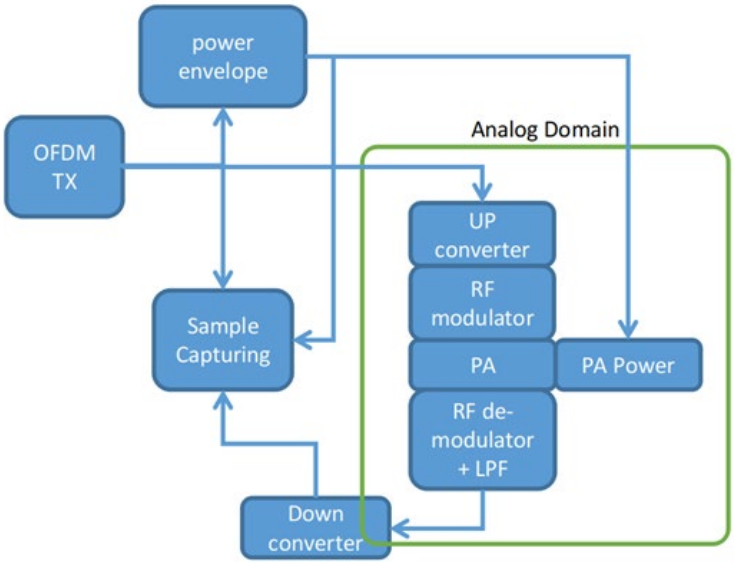


Figure 4-15 Modified training setup to support power envelope

Power envelope has a potential to save PA power but also introduces new challenges. These include:

i. The power feed becomes a parameter.
ii. PA parameters may change with envelope.
iii. The envelope signal may be designed to change slower than samples.
iiii. PA works even closer to saturation at every sample, which is more challenging for a DPD.
iv. The power envelope is additional input to the DPD or the DPD calculated the power envelope as well.

Training a DPD need the power feed as additional input. However, since response to envelope changes in the analogue circuit is different from the response to the signal itself, memory like effects are expected. Training will consequently have to consider past samples as well. Modification to the sample capturing setup is depicted below.

Any solution to PAPR is a trade-off which depends on requirements. For example, QPSK can suffer more in-band noise than QAM256, so one can deliberately introduce artificial noise in a QPSK allocation as part of PAPR reduction scheme. Out of band emission is subject to regulation as emission mask. Therefore, signal integrity may also be compromised as long as regulation is not violated.

Envelope tracking is a promising technology to reduce PA power consumption. However, a good performance is expected only with a DPD that is also aware of the power supply input.

4.3 Interference management in relay-enhanced scenarios

4.3.1 State-of-the-art and problem statement

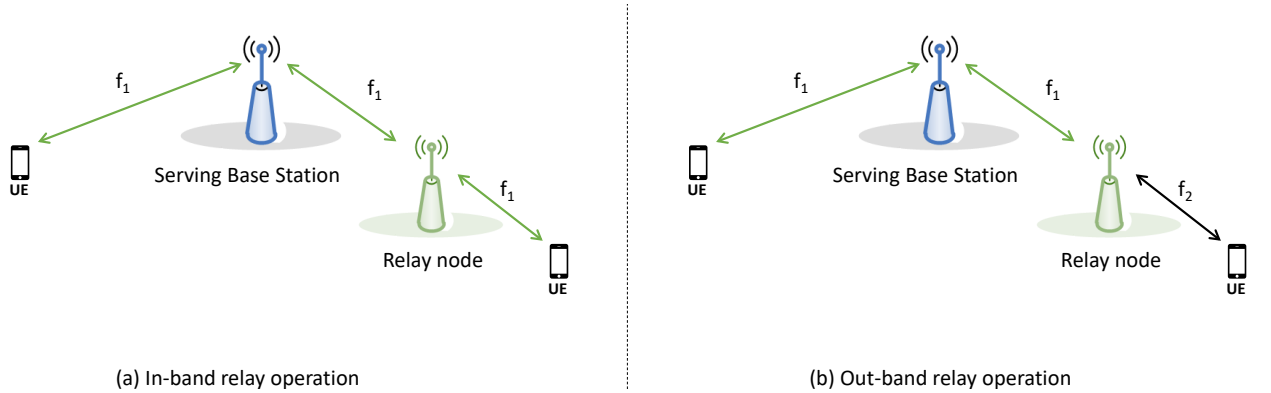
As presented in BeGREEN D2.1 ([1] Section 4.1.2), a RAN enhanced with relay nodes can reduce energy consumption while maintaining service requirements, thus providing a more energy efficient network. However, the use of relay nodes in the same coverage area of the cellular BSs causes different types of interference between cellular users connected directly to the BS, relay nodes, and users served by the relay. These interferences will be relevant depending on how radio resources are managed, e.g., reuse of cellular resources in the relay, allocation of separate resources, etc.

Based on spectrum utilisation, two operation modes are considered for relaying: in-band and out-band, as

depicted in Figure 4-16 [78]. 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. 3GPP describes in-band backhauling for Integrated Access and Backhaul (IAB) [80] when access and backhaul links operate in the same frequency, or at least exists a partial overlap in frequency, and out-band backhauling when they operate in different frequencies. The use of the same frequency creates half-duplexing or interference constraints, which imply that the IAB node cannot transmit and receive simultaneously on both links.

In [81], 3GPP defines the operating bands and bandwidths for New Radio (NR) IABs. IAB can operate in both NR FR1 and FR2, but only in Time Division Duplex (TDD) mode.

The in-band mode enables relaying without requiring additional spectrum. However, interference management becomes a challenge to avoid interferences between cellular and relay users, and self-interference in the relay nodes, see Figure 4-17.



(a) In-band relay operation

(b) Out-band relay operation

Figure 4-16 Relay operation modes

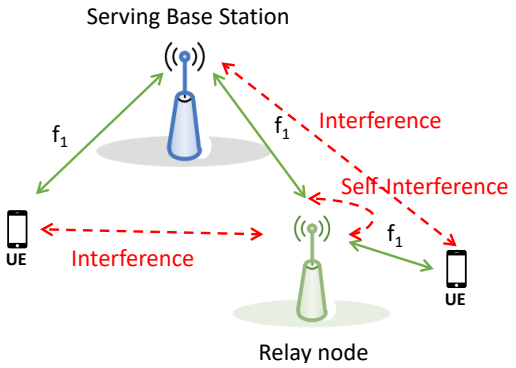
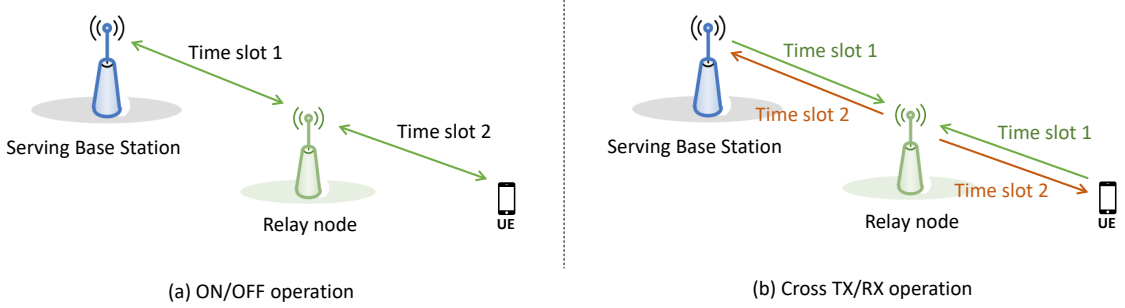


Figure 4-17 Interferences in in-band relay mode



(a) ON/OFF operation

(b) Cross TX/RX operation

Figure 4-18 Relay transmission and reception with TDM

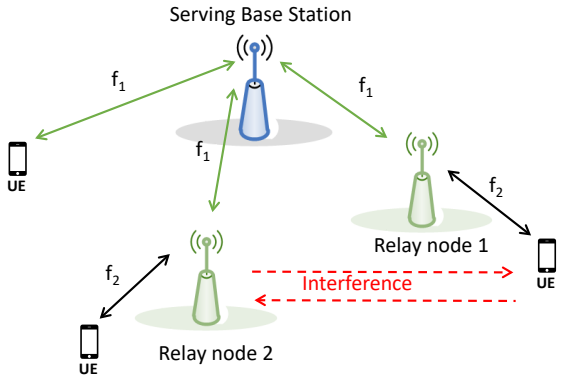


Figure 4-19 Interferences in out-band relay mode with FDD or fixed TDD

A solution to avoid self-interference in in-band mode is to use Time Domain multiplexing (TDM) between the backhaul and access links, this can be done either by forcing the two links of the relay node to be never active at the same time [78] (see Figure 4-18.a), or configuring the reception from both links at the relay at the same time, and the transmission in both links at the relay in another time [78][82] (see Figure 4-18 b).

In out-band mode, there are no interferences between cellular links (to cellular users and to relay nodes) and relay access links. However, interferences between relay nodes and users connected to the relays, must be considered if there is no coordination between relays. For Frequency Division Duplex (FDD), the interferences will be those of the relays operating at the same frequency, and the users connected to them. In Figure 4-19, we draw the interference from a user connected to relay node 1 towards relay node 2, and vice versa. Note that also relay node 1 and the user connected to relay node 2 will have interference. Thus, frequency planning at relay nodes plays an important role in reducing the existing interferences in this scenario.

When fixed TDD configuration is used between relay nodes, and the uplink and downlink resources of different relay nodes are synchronised in time, the same type of interferences shown in Figure 4-20 will also occur between relay nodes and users connected to other relay nodes.

In turn, when a dynamic TDD configuration is used, additional types of interference will occur. When one relay operates in downlink (DL), and the other in uplink (UL), as illustrated in Figure 4-20 for time slot 1 (T1), the signal transmitted by the relay 1 will interfere with the UL receiver at the relay 2, as well as for time slots 3 and 4 (T3 & T4) in reverse. Similarly, users connected to different relays can also mutually interfere when a relay operates in UL and the other in DL. This is illustrated in Figure 4-20, for time slot 1 (T1), where the signal transmitted by UE2 will interfere with the reception at UE1, and for time slots 3 and 4 (T3 & T4), where a UE1 transmitting in the UL to relay 1 generates interference to the DL receiver at the UE2 connected to relay 2. Cross interference between relays and UEs connected to other relays will also occur when relays transmit or receive simultaneously, as in time slot 2 (T2).

Since fixed relays and mobile relays are supposed to be owned and controlled by the mobile operators, they will operate in licensed bands. However, when considering the possibility of using relay UEs with Device-to-device (D2D) communications, out-band mode in an unlicensed band is also a possibility to be considered, e.g., in the industrial, scientific, and medical (ISM) spectrum. This requires UEs to have an additional interface to support Wi-Fi, Bluetooth, or Zig-Bee, which most devices on the market already have.

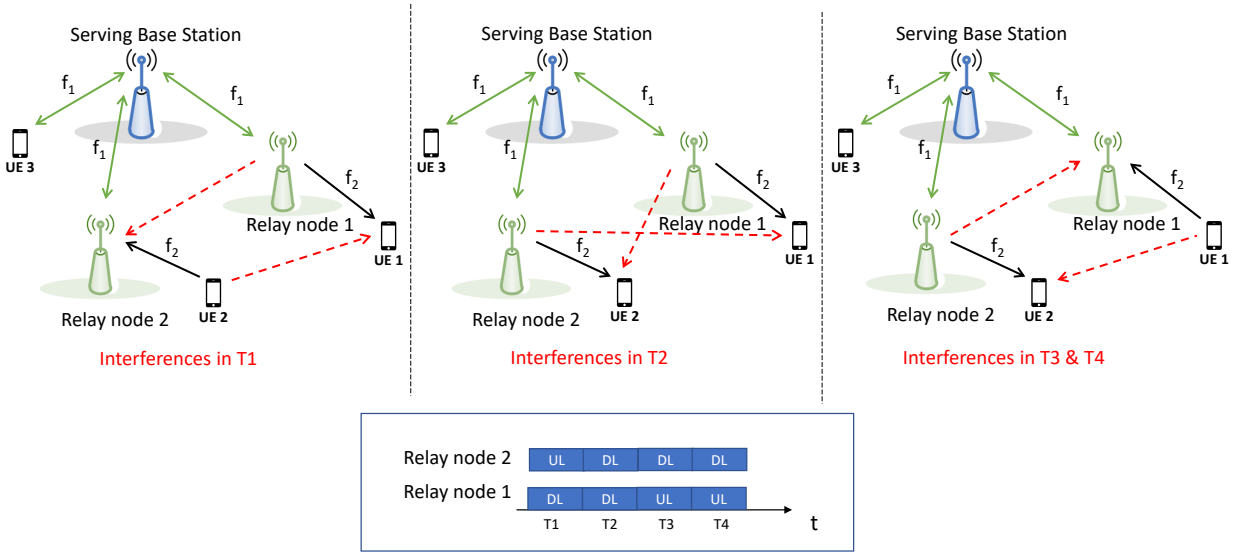


Figure 4-20 Interferences in out-band relay mode with dynamic TDD configuration

In that context, a specific classification for D2D communications can be found in the literature [83][84][85][7]. This classification distinguishes:

- Relays using licensed spectrum. It can be further classified in:
  - Overlay mode: When relaying links use different radio resources, e.g., different physical resource blocks (PRBs) than cellular users.
  - Underlay mode: When relaying links reuse the resources of the cellular users.
- Relays using unlicensed spectrum such industrial, scientific, and medical (ISM).

D2D communications in licensed underlay mode, improve spectrum reuse but there are more interferences to consider. In D2D communications using unlicensed spectrum, there is no control from mobile network operators to access to this spectrum and quality of service (QoS) is generally not guaranteed. In turn, in D2D communications using licensed spectrum with overlay mode, there are no interferences between relay links and cellular users, only interferences between relays exist, however more resources are required.

4.3.2 Interference management strategies

This section presents an overview of the interference management strategies identified in the relay-enhanced scenarios. Following a similar approach as in [84][86], we will classify the different schemes or techniques into three main groups, namely relay planning, radio resource management (RRM) techniques, and advanced signal processing techniques.

The first group includes the deployment of relays and spectrum splitting solutions, carried out in the frequency planning stage. The second group includes different RRM techniques such as: mode selection, relay selection, power allocation, radio resource allocation (RA), and joint use of them. And finally, the third group contains different signal processing techniques involving advanced antenna techniques and reception schemes that help to cancel interference signals, and thus improve signal decoding performance.

4.3.2.1 Relay planning

Relay planning includes both the deployment of relays, which consists of deciding when it is necessary to

[7] These references use in-band/out-band to refer to licensed/unlicensed, but we do not consider this terminology to avoid confusion with the general nomenclature for relays.

deploy a relay, where it is located, etc., and the frequency planning to decide at which frequency the relay works.

The deployment of relays has been studied for IAB nodes in [87], where authors propose geographical and inter-node distance constraints to decide the proper IAB network planning. In this work, the authors consider an in-band setup with a flexible allocation of resources to avoid interference between access and backhaul links, and they focus only on studying the IAB nodes' placement in order to optimize coverage and limit interference. Results demonstrate that proper network planning can boost the coverage of the IAB networks significantly.

Frequency planning strategies allocate a part of the spectrum band for the operation of relays, usually different from the part allocated to conventional cellular users. These schemes resolve cross-tier interference, i.e., interference from the cellular network to relay communications and vice versa, at the expense of worsening spectral efficiency.

Considering fixed relays, the problem of relay placement and spectrum allocation among the relays and the base station has been studied in [88]. Specifically, they analyse the deployment of three fixed relay nodes per BS in cellular systems employing soft frequency reuse (SFR) and fractional frequency reuse (FFR) schemes. Then, they propose two strategies for user association to BS or relay, a static assignment depending on their location, and a dynamic assignment depending on the strongest received signal strength. Finally, the authors propose two strategies to assign the total band for various links in each cell: a non-uniform in-band resource allocation and a uniform in-band resource allocation. Both strategies consider a dedicated band for the BS-relay link which is not directly used to serve any user. The performance of the proposed strategies has been analysed through simulations for a 19 BS system, and the results prove that the relay-aided schemes improve cellular systems in terms of coverage and throughput.

Also, for fixed relays, in [89], in-band and out-band schemes for spectrum allocation to the backhaul and access links have been proposed, and the performance of both schemes has been simulated and compared. In the proposed out-band scheme, access and backhaul use orthogonal spectrum bands, and the assigned spectrum to each small base station (a.k.a. relay) is further partitioned for access and backhaul transmissions. In the in-band scheme, access and backhaul use the same spectrum band, and the small base station alternates transmission and reception.

Similarly, for D2D communication, in [90], authors use spectrum splitting, where a part of the spectrum band is dedicated to cellular users, and the other to D2D users. They distribute the two fractions, depending on the D2D user density, to reduce signalling overhead between D2D users and their home BSs.

Finally, the frequency planning of the relay nodes in tactical communication scenarios is solved in [91] with a genetic algorithm that considers the interference of relay nodes.

4.3.2.2 Radio Resource Management (RRM)

The most relevant radio resource management techniques considered in the current state-of-the-art for relaying are mode selection, relay selection, power allocation, and resource allocation.

The *mode selection* technique decides whether a UE should be assisted by the relay or not, in order to optimize the network performance. For example, in the case of relay UEs, the authors in [92] propose an analytical approach for a mode selection scheme to minimize the transmission power consumption that at the same time reduces the outage probability. Similarly, in [93] a deep learning model is used to estimate the optimal mode selection in the case of blocking of mmWave transmission or low coverage area of mmWave communications. In this study, the potential transmitter decides to transmit the data either based on dedicated D2D communication or through the cellular uplink using a BS as a relay based on the throughput, the energy efficiency, and the coverage probability.

The *relay selection* consists of choosing the most appropriate relay to assist a UE. The techniques for choosing the relay consider different aspects such as power/energy consumption of all the elements in the relaying chain, the remaining battery state of the relay UE, its buffer state, etc. The authors in [94] propose, in the case of fixed relays, a relay selection algorithm that minimizes energy consumption in OFDMA networks considering the impact of the load of cells on transmission energy. Moreover, a distributed scheme for fixed relay selection in cooperative HARQ (C-HARQ) based transmission is introduced in [95] and focuses on the reduction of the outage probability and the increase the energy efficiency simultaneously. In this scheme, the different relays that will send the information packets are chosen to minimize the number of retransmissions, reducing the overall energy consumption of the system. Finally, in [96], a relay selection algorithm for fixed relays in cooperative multicast transmission has been proposed to maximize the minimum data rate.

The *power allocation* techniques manage the allocation of the transmission power at the transmitting nodes with the aim of improving the performance of all UEs (e.g., their throughput, power consumption, etc.) and, at the same time, mitigating interference between relays and cellular UEs. Therefore, power allocation is particularly critical when in-band operation is considered. In [97], an adaptive power allocation strategy based on two-hop rate matching of the LTE-A relay network is proposed. This strategy works in TDD and considers OFDM, and in the first time slot allocates resources to fixed relays and UEs connected directly to BS, and in the second time slot allocates resources to UEs connected to relays. Then, for the relay, it compares the rates of the first and second hops (links) and readjusts the power in order to maximize user rate and minimize the transmission power of the system. More recent works have studied power allocation strategies when using NOMA using fixed relays, such as [98].

The *resource allocation* determines which physical radio resources of the BS (e.g., channels or resource blocks in orthogonal frequency division multiple access) are allocated to the cellular and relay links. For the case of fixed relays in an indoor dense deployment, [99] consider in-band backhauling for 5G small cell networks using a TDD flexible frame format. This work identifies self-interference cancellation mechanisms in a full-duplex relay, and two strategies for direct link interference mitigation at the destination: sequential combination or time-reversed cancellation. Similarly, for IAB, in [100] a TDD frame with silent slots is proposed as a solution to reduce the impact of interference. Similarly, the authors in [101] consider an in-band IAB network and propose a double deep Q-learning-based resource allocation strategy for the backhaul link that maximizes user sum capacity.

An algorithm for resource allocation in the case of moving relays is proposed in [102], considering the interferences and maximising the ergodic rate. In that work, the authors consider moving relays in a vehicle (like a train or bus), named mobile cells, that serve the users traveling in the vehicle through an in-vehicle antenna, and that have a separate external antenna to connect to the nearest base station to backhaul data, and also to neighbouring mobile cells. Besides, they consider cellular users out-of-vehicle, transmitting to the base station. Resources must be shared between the cellular users, the links between neighbouring mobile cells, and the users in the vehicle connected to the mobile cell.

The authors in [103] propose an interference coordination scheme that utilizes prioritised scheduling for fixed relay nodes in the framework of relay-based LTE-Advanced networks. A more recent work [104] proposes a coordinated parallel resource allocation scheme for an IAB network, which consists of both a centralised scheme at the IAB donor for efficient coordination of resource allocation and a distributed scheme at each IAB node for quick response to short-term bursty traffic.

Other techniques for radio resource allocation have been proposed by the researcher community, in D2D communications, to mitigate interferences between relay and cellular networks. Some examples are fractional frequency reuse (FFR) as proposed in [105] and [106], and time-frequency hopping as proposed in [107] and [108]. Specifically, the work in [105] proposes a scheme with a central region and 3 sectored FFR,

and allocates resources first to cellular users, and then the transmitter D2D user chooses the resources that are not used in that cellular user region or sector. Similarly, the authors in [106] propose a novel D2D-aware dynamic FFR algorithm where D2D uses resources left unused by the cellular users that would have otherwise been wasted. On the other hand, a time hopping based radio resource allocation scheme operating in TDD mode and where D2D communications reuse uplink cellular resources is considered in [107]. Meanwhile, in [108], the D2D communications reuse downlink cellular resources, and a time-frequency hopping scheme is proposed for scheduling D2D links.

It is worth noting that the algorithms discussed in the above paragraph deal mainly with D2D communications in general, not necessarily related to D2D-based relays. In this respect, they can also be applied for the specific case in which D2D is used to support relay UEs.

Finally, the joint use of some of these RRM techniques is considered in different works to improve relay network optimisation. The resource and power allocation have been tackled in [109] for a high-speed railway system with moving relay nodes placed on each train wagon. Both backhaul and access links have been analysed, and finally, a planning method to determine the inter-BS distance to guarantee seamless handover between neighbour BSs serving multiple moving relays has been proposed. Similarly, in the case of fixed IAB, a scheme that combines node placement and resource allocation to maximize the downlink sum rate is developed in [110]. Furthermore, reference [111] employs simulated annealing algorithms to optimize joint scheduling and power allocation also in IAB networks. Similarly, the joint resource allocation and relay selection problem has been addressed in [112] to improve the achievable capacity for D2D communications with the help of network coding.

On the other hand, it is also important to highlight that different works have proposed different inter-cell interference coordination (ICIC) techniques in wireless communication systems when the same frequency is used in different cells, causing interference in neighbouring cells [113]. Some of the commonly used ICIC techniques include frequency reuse, power control, interference avoidance, resource block allocation, and Coordinated Multi-Point (CoMP) transmission and reception. For example, in [114], they propose to use ICIC in 3GPP LTE/LTE-A mobile networks for interference avoidance, and additionally, they provide a comparison of the state-of-the-art developments in ICIC for macro OFDMA systems. 3GPP Long Term Evolution (LTE) introduces two ICIC techniques in Rel-8, one based on the exchange of interference information between Base Stations through X2 interface, and the other based on FFR. Then, Rel-10 defines an Enhanced ICIC (eICIC) technique that improves the performance of pico cells affected by macro cells by stopping radio transmission in certain slots denoted as Almost Blank Subframes (ABS) [115], and finally, in Rel-11 the technique was evolved to further enhanced ICIC (FeICIC), to improve interference cancellation in the UE [116]. Although these ICIC techniques have not been defined for the case of relays, they could also be considered as a reference to mitigate interference in relay-based scenarios.

4.3.2.3 Advanced signal processing techniques

This last group of interference management strategies use advanced signal processing techniques to cancel interfering signals, either by reconstructing the received interfering signal and removing it, by sending the information from different points in a coordinated way, or by grouping the interferences at the receiver so that a part of the space of the received signal is free of interferences. Each of these strategies is briefly described below. As with RRM, the strategies presented with examples for D2D can be adapted to the case of relays.

Multiple-input multiple-output (MIMO): this strategy uses antenna arrays to serve multiple users in each time-frequency resource block. It can be used in different ways to avoid interferences in relay-enhanced RAN. Several works like [117] use MIMO precoding schemes, where the downlink cellular transmitter (i.e., the base station) and the D2D transmitter cooperate to choose the precoder pair that minimize interferences.

Others like [118] exploit mMIMO schemes under both perfect and imperfect channel state information (CSI) at the receivers that employ partial zero-forcing and demonstrate that mMIMO can efficiently handle the D2D-to-cellular interference.

Interference regeneration and cancellation: the basis of this technique is to regenerate the interfering signals at the receiver and subsequently cancel them from the desired signal. Different schemes have been proposed to cancel the interfering signal like the ones in [119] and [120]. In [119], an analytical framework for studying the performance of successive interference cancellation (SIC) in large-scale D2D-enabled cellular networks is presented using a stochastic equivalence of the interference to simplify the analysis. SIC technique is based on decoding the strongest interfering signal by considering other signals as noise, then regenerating the analogue signal and cancel it from the received composite signal, and then it repeats the procedure with the second strongest interfering signal from the remaining signal and so forth, until de desired signal can be decoded. On the other hand, the authors in [120] propose to track the near-far interference by monitoring common control channels and identify the interfering cellular users, and thus avoid their interference in D2D transmissions that reuse the uplink resources.

Beamforming: this technique uses multi-antennas in the BS to change the beam direction in a desired place, thus focusing the transmitted signal to the receiver position, and reducing interference to other locations. The authors in [121] exploit beamforming technique to design a D2D interference avoiding approach in which D2D share uplink resources with cellular UEs, and thanks to the coordination between them, the interfering CSI can be estimated and then derive the null space to the estimated channels to minimise the interference.

Cooperative Multi-Point (CoMP): this technique uses multiple transmit and receive antennas from multiple site locations, which may or may not belong to the same physical cell, to enhance the received signal quality as well as decrease the received spatial interference. In [122], they propose the use of CoMP to remove inter and intra interference between D2D and cellular users.

Coding schemes: these schemes use different coding techniques to mitigate interferences. In [123] the authors propose two superposition coding-based schemes for a cellular network with D2D that cooperate with the cellular link to compensate the interference generated by using the same radio resources. Instead, the authors in [124] propose a rate splitting technique in a D2D scenario, which consists in dividing the transmitted message into private and public parts. Only the intended receiver can decode the private part, while the public part can be decoded by all the receivers subject to the interference who can cancel the interference caused from the public part.

Interference Alignment (IA): this technique uses a precoding scheme which aligns the interfering signals at each receiver. Thus, the interference is concentrated into one part of the signal space at each receiver, leaving the other part available for the desired signal and free of interference. This technique is exploited in some previous research works like [125] and [126].

4.4 Telemetry subsystem

For BeGREEN and B5G networks, one common sub-system of these future networks will be the telemetry and data collection system. It will have the role of ingesting, processing and leveraging the 6G telemetry (where telemetry is used in this case to describe all sorts of 'big data' measurements, sensing, derived state, meta-data, etc) to enable the 6G RAN intelligence, for (re-)training and inference functionality, to allow AI/ML to integrate and automate these zero-touch networks, optimising for energy saving. It will also provide a future 6G evolution pathway to additionally optimise for EMF exposure, and to integrate 6G technology enablers like JCAS, RIS, cell-free (CF), dynamic spectrum re-farming, Network Digital Twins (NDTs), etc.

A key driver for 5G Open RAN was to bring intelligence into the RAN, through the O-RAN RIC, with both the Near-RT and Non-RT RIC network functions, extensible through xApps/rApps. As BeGREEN embraces the

Open RAN architecture, a 'Key Exploitable Result' of BeGREEN will be an enhancement of the current Open RAN telemetry framework, advanced to support the BeGREEN energy saving, but also to become exploitable for further evolution into an Open 6G architecture. While for BeGREEN, the implementation of the telemetry framework will be limited to the energy saving use-case and KPI fulfilment, the architecture and design will be mindful of a 6G telemetry sub-system, supporting the following 'southbound' technologies:

- Advanced non-3GPP: heterogeneous multi-RAT like Wi-Fi 7 or 8, but also UWB and Lidar technologies. While in 3GPP 5G, the focus was on user-plane integration of N3IWF/TNGF towards the UPF, it is expected that the Positioning and capabilities will increasingly be injected into the RIC to provide enhanced cross-domain functionality, beyond SotA 3GPP ATSSS.
- Native AI/ML in DU/RU: While the Near-RT RIC can ingest telemetry and control some aspects of the DU/RU, through the O-RAN E2 service modules, when the intelligence becomes deeply embedded in to the 6G Native AI/ML. This is the scope of the identified but unspecified Real time (RT) RIC. This is outside of BeGREEN scope but will be needed for future 6G capabilities.
- ISAC: The addition of radar-like sensing, integrated into the communication RUs will require processing and usage of this sensing information, to merge with other higher-level telemetry for advanced functionality.
- Full-duplex and IAB: will extend the 6G RAN coverage but will add additional control aspects of interference management and bandwidth scheduling, that will result in increased telemetry to be ingested and utilised.
- Dynamic Spectrum Re-farming: will require full sensing of the CSI (Channel Sate Information) and neighbour interference, to ingest and process the interference, extracting valuable telemetry for optimisation and co-existence, instead of discarding as pure noise.
- CF mMIMO: requires close co-ordination of DU and RU, with dynamic reconfiguration of the use-centric AP associations, for coherent multi-point Tx/Rx and synchronisation. Early research results show the need for RIC like intelligence.

While some of this 'telemetry' may be raw unprocessed 'big data' sensing information, requiring highly efficient telemetry processing towards RIC xApps/rApps, in other cases processing of these algorithms will be embedded, with only lower volume 'meta-data' control information ingested into the RIC. The BeGREEN telemetry framework will be mindful of this impact in design, even though many of the advanced 6G features will not be implemented.

Runtime dynamic service exposure of 6G technologies, using declarative APIs, will be a key enabler. This is being actively considered in O-RAN nGRG (next Generation Research Group) and is similar to IaaS (Infrastructure as a Service) design patterns.

All this telemetry needs consolidation and 'northbound' egress towards the SMO, with data lakes and other types of Apache Kafka event and telemetry unification and storage, to provide the O-RAN R1 interfaces towards higher level intelligent agents and 6G NDTs, running in the cloud-native intent-based intelligence plane.

5 Summary and Conclusions

BeGREEN aims at providing innovative solutions to improve the energy efficiency of 6G RANs. It builds up on 3GPP and O-RAN architectural frameworks and covers a wide range of mechanisms to reduce energy consumption at hardware, link, and system levels. This deliverable, BeGREEN D3.1, provides the SotA on PHY mechanisms and solutions to enhance the energy efficiency. Some of these solutions require a thorough design to be integrated in the 3GPP or O-RAN context, while others do provide improvements in specific parts of the RAN components.

Chapter 2 presented how the implementation of known 5G O-RAN components –particularly the CU and DU– can be realized in a different (and optimised) architecture, optionally offloading some of the functionalities to accelerators. The project is expecting a considerable improvement compared to the implementation in SotA architectures.

Chapter 3 tackles the energy efficiency improvement of the RU, both from the development of techniques for the optimization of the physical components of the RU and using the O-RAN capabilities, but also to leverage from the RICs to optimally manage the RU.

Chapter 4 presents the design and future implementation of the additional PHY components (and their optimisation) to improve energy efficiency and to reduce the energy consumption. External components to the 3GPP or O-RAN architectural framework such as the use of ISAC and the employment of RIS and relays do provide yet a plethora of possibilities to improve the energy efficiency of the RAN. This chapter also presents the envisioned initial design of these components and also other techniques for interference management in relay-enhanced scenarios.

The initial implementation and results of the abovementioned solutions will be reported in the next WP3 deliverable, BeGREEN D3.2.

6 Bibliography

[1] "BeGREEN D2.1, "BeGREEN Reference Architecture", June 2023. Available Online: https://www.sns-BeGREEN.com/deliverables

[2] M. Polese, L. Bonati, S. D'Oro, S. Basagni and T. Melodia, "Understanding O-RAN: Architecture, Interfaces, Algorithms, Security, and Research Challenges," in *IEEE Communications Surveys & Tutorials*, vol. 25, no. 2, pp. 1376-1411, Secondquarter 2023, doi: 10.1109/COMST.2023.3239220.

[3] G. Vallero, D. Renga et al., "Greener RAN Operation Through Machine Learning," IEEE TNSM, vol. 16, no. 3, pp. 896–908, 2019.

[4] M. Hoffmann and P. Kryszkiewicz, "Reinforcement Learning for Energy-Efficient 5G Massive MIMO: Intelligent Antenna Switching," IEEE Access, vol. 9, pp. 130 329–130 339, 2021.

[5] R. Verdecchia, J. Sallou, and L. Cruz, "A Systematic Review of Green AI," WIREs Data Mining and Knowledge Discovery, 2023.

[6] 3GPP, "Technical Specification TS 38.864 - Study on Network Energy Savings for NR," [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3 987

[7] Shakil, K. "5G Virtual DU - hardware acceleration is also about software" 10/10/2019. URL: https://www.linkedin.com/pulse/5g-virtual-du-hardware-acceleration-also-software-kashif-shakil/

[8] Parallel Wireless, "Everything you need to know about Open RAN", 30/11/2020. https://www.parallelwireless.com/blog/everything-you-need-to-know-about-open-ran-an-e-book/

[9] Stanley, S. "Why open RAN needs flexible hardware acceleration"", 06/04/2021 https://www.lightreading.com/open-ran/why-open-ran-needs-flexible-hardware-acceleration/a/d-id/769984

[10] NOKIA. "Cloud RAN: A guide to Acceleration options". 2023 https://onestore.nokia.com/asset/213050?_ga=2.238667584.1705283296.1689591088-1285650428.1689248342

[11] Stanley, S. "Accelerating Open RAN Platforms Operator Survey A Heavy Reading white paper produced for Qualcomm, Wind River, WWT, and Xilinx", May 2021.

[12] R. Misra, C. Dick, S. Velayutham and Y. Huang, "Enabling GPU Acceleration in Near-Realtime RAN Intelligent Controllers". 17/06/2021. https://developer.nvidia.com/blog/enabling-gpu-acceleration-in-near-realtime-ran-intelligent-controllers/

[13] Qualcomm® X100 5G RAN Accelerator Card

[14] Intel FlexRAN Reference Architecture

[15] Intel, "Virtual RAN (vRAN) with Hardware Accelleration – white paper", 2020. https://networkbuilders.intel.com/solutionslibrary/virtual-ran-vran-with-hardware-acceleration

[16] ARM RAN Acceleration Library

[17] Intel® vRAN Dedicated Accelerator ACC100

[18] NXP Layerscape®Access LA12xx Programmable Baseband Processor

[19] NVIDIA Aerial SDK

[20] Chance Tarver, Matthew Tonnemacher, Hao Chen, Jianzhong Zhang and Joseph R. Cavallaro, "GPU-Based, LDPC Decoding for 5G and Beyond" CAS – IEEE Open Journal on Circuits and Systems, 2021

[21] Dell technologies, "Dell Open RAN Accelerator Card." 2022, FY23Q1_prod_1649_dell_open_ran_accelerator_card_brief_021622 https://www.delltechnologies.com/asset/en-us/solutions/service-provider-solutions/briefs-summaries/dell-open-ran-accelerator-solution-brief.pdf

[22] Murad Qasaimeh, Kristof Denolfy, Jack Loy, Kees Vissersy, Joseph Zambreno, and Phillip H. Jones, "Comparing

Energy Efficiency of CPU, GPU and FPGA Implementations for Vision Kernels", IEEE International Conference on Embedded Software and Systems (ICESS), 2019.

[23] NVIDIA Jetson AGX Orin 64GB

[24] Dimitris Vordonis and Vassilis Paliouras, "Sphere Decoder for Massive MIMO Systems", IEEE Nordic Circuits and Systems Conference (NORCAS), 2019

[25] Wu, Y & McAllister, "Configurable Quasi-Optimal Sphere Decoding for Scalable MIMO Communications", IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 68, no. 6, pp. 2675- 2687, 2021.

[26] 3GPP TS 38.214, "NR; Physical layer procedures for data (Relase 16)", Spetember 2021.

[27] NVIDIA Jetson Stats package for monitoring and control

[28] E. Kolta, T. Hatt et al., "Going Green: Benchmarking the Energy Efficiency of Mobile," GSMA Intelligence, Tech. Rep., Jun. 2021.

[29] Small Cell Forum, 5G FAPI: PHY API Specification. Document number: SCF222, version 07.00 august 2023 http://scf.io/doc/222

[30] Small Cell Forum, 5G FAPI: P19 RF and Digital Frontend Control API. Document number: SCF223, version 05.00 August 2023, http://scf.io/doc/223.

[31] Small Cell Forum, 5G FAPI: Network Monitor Mode API. Document number: SCF224, version 02.00 August 2023, http://scf.io/doc/224

[32] O-RAN Alliance, O-RAN.WG7.NES.0-v1.00 - O-RAN Work Group 7 (White-box Hardware Workgroup) Network Energy Savings Procedures and Performance Metrics,

[33] O-RAN Alliance, O-RAN.WG1.Use-Cases-Analysis-Report-R003-v12.00 - O-RAN Work Group 1 (Use Cases and Overall Architecture) Use Cases Analysis Report, https://orandownloadsweb.azurewebsites.net/specifications

[34] 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

[35] O-RAN Alliance, O-RAN.WG4.CUS.0-R003-v13.00 - O-RAN Work Group 4 (Control, User and Synchronization Plane Specification), https://orandownloadsweb.azurewebsites.net/specifications

[36] O-RAN Alliance, O-RAN.WG4.MP.0-R003-v13.00 - O-RAN Work Group 4 (Open Fronthaul Interfaces Workgroup) Management Plane Specification, https://orandownloadsweb.azurewebsites.net/specifications

[37] O-RAN Alliance Working Group 4, O-RAN.WG4.MP.0-v06.00 Technical Specification, Management Plane Specification.

[38] BeGREEN D4.1, "State-of-the-Art on PHY Mechanisms Energy Consumption and Specification of Efficient Enhancement Solutions", December 2023. Available Online: https://www.sns-BeGREEN.com/deliverables

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

[40] F. Liu and C. Masouros, "Hybrid Beamforming with Sub-arrayed MIMO Radar: Enabling Joint Sensing and Communication at mmWave Band," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 2019.

[41] J. A. Zhang et al., "An Overview of Signal Processing Techniques for Joint Communication and Radar Sensing," in IEEE Journal of Selected Topics in Signal Processing, vol. 15, no. 6, pp. 1295-1315, Nov. 2021.

[42] C. Sturm and W. Wiesbeck, "Waveform design and signal processing aspects for fusion of wireless communications and radar sensing," Proc. IEEE, vol. 99, no. 7, pp. 1236–1259, Jul. 2011.

[43] Q. Zhang, H. Sun, X. Gao, X. Wang, and Z. Feng, "Time-division ISAC enabled connected automated vehicles cooperation algorithm design and performance evaluation," IEEE J. Sel. Areas Commun., vol. 40, no. 7, pp. 2206–2218, Jul. 2022.

[44] Renzo, Marco Di, et al. "Smart radio environments empowered by reconfigurable AI meta-surfaces: An idea whose time has come." EURASIP Journal on Wireless Communications and Networking 2019.1 (2019): 1-20.

[45] Asif Haider, Mirza and Zhang, Yimin D., "RIS-aided integrated sensing and communication: a mini-review", Frontiers in Signal Processing, vol. 3, 2023.

[46] R. S. Prasobh Sankar, B. Deepak and S. P. Chepuri, "Joint Communication and Radar Sensing with Reconfigurable Intelligent Surfaces," 2021 IEEE 22nd International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), Lucca, Italy, 2021.

[47] R. W. Heath, "Communications and Sensing: An Opportunity for Automotive Systems [From the Editor]," in IEEE Signal Processing Magazine, vol. 37, no. 4, pp. 3-13, July 2020.

[48] A. Klautau, N. González-Prelcic and R. W. Heath, "LIDAR Data for Deep Learning-Based mmWave Beam-Selection," in IEEE Wireless Communications Letters, vol. 8, no. 3, pp. 909-912, June 2019, doi: 10.1109/LWC.2019.2899571.

[49] S. Jiang and A. Alkhateeb, "Computer Vision Aided Beam Tracking in A Real-World Millimeter Wave Deployment," 2022 IEEE Globecom Workshops (GC Wkshps), Rio de Janeiro, Brazil, 2022, pp. 142-147, doi: 10.1109/GCWkshps56602.2022.10008648.

[50] Ahmed M. Nor, Simona Halunga, Octavian Fratu, Survey on positioning information assisted mmWave beamforming training, Ad Hoc Networks, Volume 135, 2022, 102947, ISSN 1570-8705.

[51] A. Zhang, M. L. Rahman, X. Huang, Y. J. Guo, S. Chen and R. W. Heath, "Perceptive Mobile Networks: Cellular Networks With Radio Vision via Joint Communication and Radar Sensing," in IEEE Vehicular Technology Magazine, vol. 16, no. 2, pp. 20-30, June 2021, doi: 10.1109/MVT.2020.3037430.

[52] J. Moghaddasi and K. Wu, "Multifunctional transceiver for future radar sensing and radio communicating data-fusion platform," IEEE Access, vol. 4, pp. 818–838, Feb. 2016.

[53] Yang, Zhaohui, et al. "Energy-efficient wireless communications with distributed reconfigurable intelligent surfaces." IEEE Transactions on Wireless Communications 21.1 (2021): 665-679.

[54] Huang, Chongwen, et al. "Reconfigurable intelligent surfaces for energy efficiency in wireless communication." IEEE transactions on wireless communications 18.8 (2019): 4157-4170.

[55] Li, Zhiyang, et al. "Energy efficient reconfigurable intelligent surface enabled mobile edge computing networks with NOMA." *IEEE Transactions on Cognitive Communications and Networking* 7.2 (2021): 427-440.

[56] De Sena, Arthur S., et al. "What role do intelligent reflecting surfaces play in multi-antenna non-orthogonal multiple access?." IEEE Wireless Communications 27.5 (2020): 24-31.

[57] Jia, Shuaiqi, Xiaojun Yuan, and Ying-Chang Liang. "Reconfigurable intelligent surfaces for energy efficiency in D2D communication network." *IEEE Wireless Communications Letters* 10.3 (2020): 683-687.

[58] Liu, Xiao, Yuanwei Liu, and Yue Chen. "Machine learning empowered trajectory and passive beamforming design in UAV-RIS wireless networks." IEEE Journal on Selected Areas in Communications 39.7 (2020): 2042-2055.

[59] Ren, Hong, et al. "Energy minimization in RIS-assisted UAV-enabled wireless power transfer systems." IEEE Internet of Things Journal 10.7 (2022): 5794-5809.

[60] Long, Hui, et al. "Reflections in the sky: Joint trajectory and passive beamforming design for secure UAV networks with reconfigurable intelligent surface." *arXiv preprint arXiv:2005.10559* (2020).

[61] Cao, Binghao, et al. "Reflecting the light: Energy efficient visible light communication with reconfigurable intelligent surface." *2020 IEEE 92nd Vehicular Technology Conference (VTC2020-Fall)*. IEEE, 2020.

[62] 3GPP TS 38.141-1, "NR; Base Station (BS) conformance testing Part 1: Conducted conformance testing (Release 16)", September 2021

[63] J. A. Zhang et al., "Enabling Joint Communication and Radar Sensing in Mobile Networks—A Survey," in IEEE Communications Surveys & Tutorials, vol. 24, no. 1, pp. 306-345, Firstquarter 2022, doi: 10.1109/COMST.2021.3122519.

[64] S. Häger, M. Haferkamp and C. Wietfeld, "Beam-based 6G Networked Sensing Architecture for Scalable Road Traffic Monitoring," 2023 IEEE International Systems Conference (SysCon), Vancouver, BC, Canada, 2023, pp. 1-8, doi: 10.1109/SysCon53073.2023.10131249.

[65] Y. Cui, W. Yuan, Z. Zhang, J. Mu and X. Li, "On the Physical Layer of Digital Twin: An Integrated Sensing and Communications Perspective," in IEEE Journal on Selected Areas in Communications, vol. 41, no. 11, pp. 3474-3490, Nov. 2023, doi: 10.1109/JSAC.2023.3314826.

[66] A. Nacef, M. Bagaa, Y. Aklouf, A. Kaci, D. L. C. Dutra and A. Ksentini, "Self-optimized network: When Machine Learning Meets Optimization," 2021 IEEE Global Communications Conference (GLOBECOM), Madrid, Spain, 2021, pp. 1-6, doi: 10.1109/GLOBECOM46510.2021.9685681.

[67] Garcia-Saavedra, Andres, and Xavier Costa-Perez. "O-RAN: Disrupting the virtualized RAN ecosystem." IEEE Communications Standards Magazine 5.4 (2021): 96-103.

[68] Albanese, Antonio, et al. "RIS-aware indoor network planning: The Rennes railway station case." ICC 2022-IEEE International Conference on Communications. IEEE, 2022.

[69] Alexandropoulos, George C., et al. "Hybrid reconfigurable intelligent metasurfaces: Enabling simultaneous tunable reflections and sensing for 6G wireless communications." arXiv preprint arXiv:2104.04690 (2021).

[70] Lamminen, Antti EI, Jussi Saily, and Antti R. Vimpari. "60-GHz patch antennas and arrays on LTCC with embedded-cavity substrates." IEEE Transactions on Antennas and Propagation 56.9 (2008): 2865-2874.

[71] Li, Jian-Hua, Xiaoping Liao, and Chenlei Chu. "A novel thermistor-based RF power sensor with wheatstone bridge fabricating on MEMS membrane." Journal of Microelectromechanical Systems 29.5 (2020): 1314-1321.

[72] Yasir, M., et al. "Integration of antenna array and self-switching graphene diode for detection at 28 GHz." IEEE Electron Device Letters 40.4 (2019): 628-631.

[73] Bhattacharya, Ritabrata, et al. "An 8-channel varactor-less 28-GHz front end with 7-Bit resolution 340 RTPS for 5G RF beamformers." IEEE Transactions on Circuits and Systems II: Express Briefs 66.12 (2019): 1937-1941.

[74] Prof. Anding Zhu from the University College Dublin at the ACRC webinar Dec. 21 “Digital Predistortion for 5G MIMO Transmitters Using Machine Learning”.

[75] Teng Wang, Wantao Li , Roberto Quaglia and Pere L. Gilabert from April 21 “Machine-Learning Assisted Optimization of Free-Parameters of a Dual-Input Power Amplifier for Wideband Applications”

[76] Yucheng Yu, Peng Chen, Xiao-Wei Zhu, Jinfeng Zhai and Chao Yu all from State Key Laboratory of Millimeter Waves, Southeast University, Nanjing, China: “Continual Learning Digital Predistortion of RF Power Amplifier for 6G AI-Empowered Wireless Communication” , China. Paper published in: IEEE Transactions on Microwave Theory and Techniques (October 2022).

[77] Chang Gao from TU Delft , Holland “Hardware Accelerated AI for Digital Predistortion In 5G era [2023]

[78] J. Gora and S. Redana, "In-band and out-band relaying configurations for dual-carrier LTE-advanced system", 2011 IEEE 22nd International Symposium on Personal, Indoor and Mobile Radio Communications, Toronto, ON, Canada, 2011, pp. 1820-1824.

[79] D. K. Pin Tan et al., "Integrated Sensing and Communication in 6G: Motivations, Use Cases, Requirements, Challenges and Future Directions," 2021 1st IEEE International Online Symposium on Joint Communications & Sensing (JC&S), Dresden, Germany, 2021, pp. 1-6

[80] 3GPP TR 38.874 “Study on Integrated Access and Backhaul; (Release 16)”, December 2018.

[81] 3GPP TR 38.174 “Integrated access and backhaul radio transmission and reception (Release 18)”, March 2023.

[82] G. Noh, H. Chung, I. Kim, "Mobile Relay Technology for 5G", in IEEE Wireless Communications, June 2020.

[83] A. Asadi et al., "A Survey on Device-to-Device Communication in Cellular Networks“, in IEEE Communications Surveys & Tutorials, vol. 16, no. 4, pp. 1801-1819, Fourthquarter 2014.

[84] P. Mach and Z. Becvar, "Device-to-Device Relaying: Optimization, Performance Perspectives, and Open

Challenges Towards 6G Networks", in IEEE Communications Surveys & Tutorials, vol. 24, no. 3, pp. 1336-1393, third quarter 2022.

[85] G. Fodor, et al., "Design aspects of network assisted device-to-device communications", in IEEE Communications Magazine, vol. 50, no. 3, pp. 170-177, March 2012.

[86] M.Noura and R. Nordin, “A survey on interference management for Device-to-Device (D2D) communication and its challenges in 5G networks”, Journal of Network and Computer Applications, Volume 71, 2016, Pages 130-150, ISSN 1084-8045

[87] C. Madapatha, B. Makki, H. Guo and T. Svensson, "Constrained Deployment Optimization in Integrated Access and Backhaul Networks," 2023 IEEE Wireless Communications and Networking Conference (WCNC), Glasgow, United Kingdom, 2023, pp. 1-6.

[88] S. Ghosh and A. P. Kannu, "Relay placement and spectrum sharing strategies for soft and fractional frequency reuse schemes," 2015 Twenty First National Conference on Communications (NCC), Mumbai, India, 2015, pp. 1-6.

[89] U. Siddique, H. Tabassum and E. Hossain, "Downlink spectrum allocation for in-band and out-band wireless backhauling of full-duplex small cells," in IEEE Transactions on Communications, vol. 65, no. 8, pp. 3538-3554, Aug. 2017.

[90] B. Cho, K. Koufos and R. Jäntti, "Spectrum allocation and mode selection for overlay D2D using carrier sensing threshold," 2014 9th International Conference on Cognitive Radio Oriented Wireless Networks and Communications (CROWNCOM), Oulu, Finland, 2014, pp. 26-31.

[91] Y. Ma, Q. Shan, D. Ma, J. Diao and J. Xiong, "Frequency Planning for Non-dedicated Relay Node Relay Based on Genetic Algorithm," 2020 IEEE 3rd International Conference of Safe Production and Informatization (IICSPI), Chongqing City, China, 2020, pp. 61-66.

[92] J. Wang, X. Xu, X. Tang, S. Zhang and X. Tao, "Analytical Modeling of Mode Selection for UE-To-Network Relay Enabled Cellular Networks with Power Control," 2018 IEEE International Conference on Communications Workshops (ICC Workshops), Kansas City, MO, USA, 2018, pp. 1-6.

[93] A. Abdelreheem, A. S. A. Mubarak, O. A. Omer, H. Esmaiel and U. S. Mohamed, "Improved D2D Millimeter Wave Communications for 5G Networks Using Deep Learning," 2020 2nd International Conference on Computer and Information Sciences (ICCIS), Sakaka, Saudi Arabia, 2020, pp. 1-5.

[94] L. You, D. Yuan, N. Pappas and P. Värbrand, "Energy-Aware Wireless Relay Selection in Load-Coupled OFDMA Cellular Networks," in IEEE Communications Letters, vol. 21, no. 1, pp. 144-147, Jan. 2017.

[95] J. Kim, K. Kim and J. Lee, "Energy-Efficient Relay Selection of Cooperative HARQ Based on the Number of Transmissions Over Rayleigh Fading Channels," in IEEE Transactions on Vehicular Technology, vol. 66, no. 1, pp. 610-621, Jan. 2017.

[96] Y. Wang, W. Wang, L. Chen, P. Zhou and Z. Zhang, "Relay Selection for Multi-Channel Cooperative Multicast: Lexicographic Max-Min Optimization," in IEEE Transactions on Communications, vol. 66, no. 3, pp. 959-971, March 2018.

[97] X. Li and B. Liu, "Adaptive power allocation based on the two-hop rate matching of LTE-A relay network," Fifth International Conference on Computing, Communications and Networking Technologies (ICCCNT), Hefei, China, 2014, pp. 1-5.

[98] Y. Lee, B. -Y. Huang and S. -I. Sou, "An Efficient Algorithm for Joint Power Allocation in NOMA-Based Diamond Relay Networks," 2020 International Computer Symposium (ICS), Tainan, Taiwan, 2020, pp. 288-293.

[99] R. -A. Pitaval, O. Tirkkonen, R. Wichman, K. Pajukoski, E. Lahetkangas and E. Tiirola, "Full-duplex self-backhauling for small-cell 5G networks," in IEEE Wireless Communications, vol. 22, no. 5, pp. 83-89, October 2015.

[100] V. F. Monteiro et al., "TDD frame design for interference handling in mobile IAB networks," GLOBECOM 2022 - 2022 IEEE Global Communications Conference, Rio de Janeiro, Brazil, 2022, pp. 5153-5158.

[101] J. Park, H. Jin, J. Joo, G. Choi and S. C. Kim, "Double Deep Q-Learning based Backhaul Spectrum Allocation in Integrated Access and Backhaul Network," 2023 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), Bali, Indonesia, 2023, pp. 706-708.

[102] S. Jaffry, S. F. Hasan and X. Gui, "Efficient Resource-Sharing Algorithms for Mobile-Cell's Sidehaul and Access Links," in IEEE Networking Letters, vol. 1, no. 2, pp. 72-75, June 2019.

[103] Z. Ren, A. B. Saleh, Ö. Bulakci, S. Redana, B. Raaf and J. Hämäläinen, "Joint interference coordination and relay cell expansion in LTE-Advanced networks," 2012 IEEE Wireless Communications and Networking Conference (WCNC), Paris, France, 2012, pp. 2874-2878.

[104] M. Yu, Y. Pi, A. Tang and X. Wang, "Coordinated parallel resource allocation for integrated access and backhaul networks, " Computer Networks, vol. 222, p. 109533, 2023.

[105] S. Mumtaz, K. M. S. Huq, A. Radwan, J. Rodriguez and R. L. Aguiar, "Energy efficient interference-aware resource allocation in LTE-D2D communication," 2014 IEEE International Conference on Communications (ICC), Sydney, NSW, Australia, 2014, pp. 282-287.

[106] T. Bansal, K. Sundaresan, S. Rangarajan and P. Sinha, "R2D2: Embracing device-to-device communication in next generation cellular networks," IEEE INFOCOM 2014 - IEEE Conference on Computer Communications, Toronto, ON, Canada, 2014, pp. 1563-1571.

[107] T. Chen, G. Charbit and S. Hakola, "Time Hopping for Device-To-Device Communication in LTE Cellular System," 2010 IEEE Wireless Communication and Networking Conference, Sydney, NSW, Australia, 2010, pp. 1-6.

[108] Q. Ye, M. Al-Shalash, C. Caramanis and J. G. Andrews, "Resource Optimization in Device-to-Device Cellular Systems Using Time-Frequency Hopping," in IEEE Transactions on Wireless Communications, vol. 13, no. 10, pp. 5467-5480, Oct. 2014.

[109] H. Ghazzai, T. Bouchoucha, A. Alsharoa, E. Yaacoub, M. -S. Alouini and T. Y. Al-Naffouri, "Transmit Power Minimization and Base Station Planning for High-Speed Trains With Multiple Moving Relays in OFDMA Systems," in IEEE Transactions on Vehicular Technology, vol. 66, no. 1, pp. 175-187, Jan. 2017.

[110] J. Y. Lai, W. -H. Wu and Y. T. Su, "Resource Allocation and Node Placement in Multi-Hop Heterogeneous Integrated-Access-and-Backhaul Networks," in IEEE Access, vol. 8, pp. 122937-122958, 2020.

[111] F. Gómez-Cuba and M. Zorzi, "Twice Simulated Annealing Resource Allocation for mmWave Multi-hop Networks with Interference," ICC 2020 - 2020 IEEE International Conference on Communications (ICC), Dublin, Ireland, 2020, pp. 1-7.

[112] C. Gao, Y. Li, Y. Zhao and S. Chen, "A Two-Level Game Theory Approach for Joint Relay Selection and Resource Allocation in Network Coding Assisted D2D Communications," in IEEE Transactions on Mobile Computing, vol. 16, no. 10, pp. 2697-2711, 1 Oct. 2017.

[113] A. S. Hamza, S. S. Khalifa, H. S. Hamza and K. Elsayed, "A Survey on Inter-Cell Interference Coordination Techniques in OFDMA-Based Cellular Networks," in IEEE Communications Surveys & Tutorials, vol. 15, no. 4, pp. 1642-1670, Fourth Quarter 2013.

[114] C. Kosta, B. Hunt, A. U. Quddus and R. Tafazolli, "On Interference Avoidance Through Inter-Cell Interference Coordination (ICIC) Based on OFDMA Mobile Systems," in IEEE Communications Surveys & Tutorials, vol. 15, no. 3, pp. 973-995, Third Quarter 2013.

[115] 3GPP TS 36.300, "Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2".

[116] J. Acharya, L. Gao, S. Gaur, "Release 11 Further Enhanced ICIC: Transceiver Processing," in Heterogeneous Networks in LTE-Advanced, Wiley, 2014, pp.133-148.

[117] H. Tang, C. Zhu and Z. Ding, "Cooperative MIMO precoding for D2D underlay in cellular networks," 2013 IEEE International Conference on Communications (ICC), Budapest, Hungary, 2013, pp. 5517-5521.

[118] X. Lin, R. W. Heath and J. G. Andrews, "The Interplay Between Massive MIMO and Underlaid D2D Networking,"

in IEEE Transactions on Wireless Communications, vol. 14, no. 6, pp. 3337-3351, June 2015.

[119] C. Ma, W. Wu, Y. Cui and X. Wang, "On the performance of successive interference cancellation in D2D-enabled cellular networks," 2015 IEEE Conference on Computer Communications (INFOCOM), Hong Kong, China, 2015, pp. 37-45.

[120] S. Xu, H. Wang, T. Chen, Q. Huang and T. Peng, "Effective Interference Cancellation Scheme for Device-to-Device Communication Underlaying Cellular Networks," 2010 IEEE 72nd Vehicular Technology Conference - Fall, Ottawa, ON, Canada, 2010, pp. 1-5.

[121] Wei Fu, Ruochen Yao, Feifei Gao, J. C. F. Li and Ming Lei, "Robust null-space based interference avoiding scheme for D2D communication underlaying cellular networks," 2013 IEEE Wireless Communications and Networking Conference (WCNC), Shanghai, China, 2013, pp. 4158-4162.

[122] S. Mumtaz, K. M. S. Huq and J. Rodriguez, "Coordinated paradigm for D2D communications," 2014 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), Toronto, ON, Canada, 2014, pp. 718-723.

[123] Chuan Ma, Gaofei Sun, X. Tian, Kai Ying, Hui Yu and X. Wang, "Cooperative relaying schemes for device-to-device communication underlaying cellular networks," 2013 IEEE Global Communications Conference (GLOBECOM), Atlanta, GA, USA, 2013, pp. 3890-3895.

[124] C. -H. Yu and O. Tirkkonen, "Device-to-Device underlay cellular network based on rate splitting," 2012 IEEE Wireless Communications and Networking Conference (WCNC), Paris, France, 2012, pp. 262-266.

[125] H. E. Elkotby, K. M. F. Elsayed and M. H. Ismail, "Exploiting interference alignment for sum rate enhancement in D2D-enabled cellular networks," 2012 IEEE Wireless Communications and Networking Conference (WCNC), Paris, France, 2012, pp. 1624-1629.

[126] L. Yang, W. Zhang and S. Jin, "Interference Alignment in Device-to-Device LAN Underlaying Cellular Networks," in IEEE Transactions on Wireless Communications, vol. 14, no. 7, pp. 3715-3723, July 2015.