Beyond 5G Artificial Intelligence Assisted Energy Efficient Open Radio Access Network

BeGREEN Reference Architecture

[Deliverable D2.1]

July 2023

Co-funded by
the European Union

6GSNS

Contractual Date of Delivery:	**June 30, 2023**
Actual Date of Delivery:	**July 31, 2023**
Work Package:	**WP2**
Dissemination Level:	**Public**
Editor(s):	**Guillermo Bielsa, Juan Francisco Esteban Rivas (TSA)**
Author(s):	**Guillermo Bielsa, Juan Francisco Esteban Rivas (TSA)** **German Castellanos, Revaz Berozashvii, Simon Pryor (ACC)** **Jordi Pérez-Romero, Oriol Sallent, Juan Sánchez-González, Anna Umbert (UPC)** **Miguel Catalán-Cid, Esteban Municio (I2CAT)** **Keith Briggs (BT)** **Israel Koffman (REL), Joss Armstrong (LMI)** **Ory Eger (PW), Allaukik Abhishek (ARM)** **Josep Xavier Salvat Lozano, Jose Ayala Romero (NEC)** **Vladica Sark, Jesús Gutiérrez (IHP)** **Mir Ghoraishi (GIGASYS)**

This document has been produced in the course of SNS-JU BeGREEN Project. The research leading to these results received funding from the European Commission Horizon Europe Programme under grant agreement No. 101097083. All information in this document is provided “as is", there is no guarantee that the information is fit for any particular purpose. The user thereof uses the information at its own risk and liability. For the avoidance of all doubts, the European Commission has no liability in respect of this document, which is merely representing the authors view.

Revision History

Revision	Date	Editor / Commentator	Description of Edits
0.1	2023-01-10	Mir Ghoraishi (Gigasys)	Template made ready
0.2	2023-01-19	Guillermo Bielsa (TSA)	Table of contents/work distribution
0.3	2023-03-20	All	First round of contributions
0.4	2023-04-28	Keith Briggs (BT)	Reviewed sections with BT responsibility.
	2023-05-23	German Castellanos (ACC)	Revision of Ch3, and completitions of ACC sections in Ch4 and Ch5. Updated the List of acronyms Reviewing of Ch3, including merging of coverage extensions in RIS and Relays. Description of the Telemetry framework 6g Evolution.
0.5	2023-05-30	All	Second round of contributions
0.6	2023-06-08	Josep Xavier Salvat, Jose Ayala Romero (NEC)	Review of chapter 2
0.65	2023-06-12	Josep Xavier Salvat, Jose Ayala Romero (NEC)	Review of chapters 4 and 5
0.66	2023-06-22	Israel Koffman, Baruch Globen -(RunEL)	Top to bottom review
0.7	2023-06-22	TSA	Updated acronyms list (in word and excel)
0.8	2023-06-29	Keith Briggs (BT)	Style, grammar, spelling review.
0.9	2023-07-30	Jesús Gutiérrez (IHP)	Final Review
0.91	2023-07-31	Guillermo Bielsa (TSA), Josep Xavier Salvat (NEC), German Castellanos (ACC), Jesús Gutiérrez (IHP), Mir Ghoraishi (Gigasys)	Latest small changes to the document
1.00	2023-07-31	Simon Pryor (ACC)	Deliverable submission to the EC

List of Figures

List of Tables

List of Acronyms

3GPP	3rd Generation Partnership Project
4G	4th generation
5GC	5G Core
5GNR	5G New Radio
6G	6th generation
AC	Alternating Current
ACPI	Advanced Configuration and Power Interface
AF	Application Function
AI	Artificial Intelligence
ANN	Artificial Neural Networks
AoA	Angle-of-Arrival
AP	Access Point
API	Application Programming Interface
AR	Augmented Reality
ASM	Advanced Sleep Modes
B5G	Beyond 5G
BAP	Backhaul Adaptation Protocol
BBU	Baseband Unit
BS	Base Station
BW	Bandwidth
CAPEX	Capital Expenditure
CCO	Coverage Capacity and Optimisation
CF	Cell-Free
CF-mMIMO	Cell-Free Massive MIMO
CNF	Cloud-Native Network Function
CoMP	Coordinated MultiPoint
COTS	Commercial Off-The-Shelf
CP	Control Plane
CPCC	Collaborative Processor Performance Control
CPE	Customer-premises Equipment
CPU	Central Procesor Unit
C-RAN	Cloud Radio Access Network
CSI	Channel State Information
CU	Central Unit
CUPS	Control-User Plane Separation
D2D	Device to Device
DC	Direct Current
DL	Down Link
DME	Data Management and Exposure
D-MIMO	Distributed MIMO
DNN	Deep Neural Networks
DoW	Description of Work
DPD	Digital Predistortion

DQN	Deep Q Network
DRL	Deep Reinforcement Learning
DU	Distributed Unit
DV	Data Volume
DVFS	Dynamic Voltage and Frequency Scaling
E2E	End-to-End
E2SM	E2 Service Model
EC	Energy Consumption
EE	Energy Efficiency
EI	Enrichment Information
ES	Energy Saving
ET	Envelope Tracking
FOCOM	Federated O-Cloud Orchestration and Management
FR	Frequency Range
GUI	Graphical User Interface
GPU	Graphics Processing Unit
H-ARQ	Hybrid automatic repeat request
H-CRAN	Heterogeneous Cloud Radio Access Network
HetNet	Heterogeneous Network
HPN	High Power Node
HW	Hardware
IAB	Integrated Access and Backhaul
ICI	Inter-cell Interference
IoT	Internet of Things
IP	Internet Protocol
ISAC	Integrated Sensing and Communication
ISM	Industrial Scientific and Medical
JCAS	Joint Communication and Sensing
KPI	Key Performance Indicator
L2	Layer 2
LDPC	Low Density Parity Check
LESS	low-energy scheduler solution (LESS)
LLR	Log Likelyhood Ratio
LoS	Line-of-Sight
LPN	Low Power Nodes
LTE	Long Term Evolution
MAC	Medium Access layer
MANO	Management and Orchestration
MBS	Macro Base Station
MEC	Multi-access Edge Computing
MIMO	Multiple-Inputs Multiple-Outputs
ML	Machine Learning
mMIMO	Massive MIMO
mmWave	Millimetre Wave
MNO	Mobile Network Operators

MT	Mobile Termination
MU MIMO	Multiple User MIMO
MVA	Mobile Video Analytics
NDT	Network Digital Twin
Near-RT	Near Real Time
NEF	Network Exposure Function
NFO	Network Function Orchestration
NFV	Network Function Virtualization
Ng-eNB	Next generation evolved node B
nGRG	next Generation Research Group
NIC	Network Interface Card
NLoS	Non-Line-of-Sight
NMS	Network Management System
Non-RT	Non Real Time
NR	New Radio
NWDAF	Network Data Analytics Function
O-Cloud	Open Cloud
O-eNB	Open evolved Node B
OPEX	Operational Expenditure
O-RAN	Open RAN
O-RU	Open-Radio Unit
OS	Operating System
PA	Power Amplifier
PDCP	Packet Data Convergence Protocol
PEE	Power, Energy and Environmental
PHY	Physical layer
PoC	Proof-of-Concept
ProSe	Proximity based Services
QoS	Quality of Service
RAIE	RAN Analytics Information Exposure
RAN	Radio Access Network
rApps	radio access network Applications
REB	Resource Element Block
RF	Radio Frecuency
RIC	RAN Interface Controller
RIS	Reconfigurable intelligent surfaces
RL	Reinforcement Learning
RNIS	Radio Network Information Service
RRH	Remote Radio Head
RRM	Radio Resource Management
RS	Relay Station
RSS	Received Signal Strength
RT	Real Time
RU	Radio Unit
RUE	Relay UE

RX	Receiver
SDAP	Service Data Adaptation Protocol
SCB	Small Cell Base station
SDN	Software Defined Networks
SDR	Software Defined Radio
SLA	Service Level Assurance
SME	Service Management and Exposure
SMO	Service Management and Orchestrator
SNS	Smart Networks and Services
SOC	System on Chip
SON	Self-Organising Networks
SotA	State of the Art
SST-CP	Speed Select Technology Core Power
SU MIMO	Single User MIMO
SW	Software
TDD	Time Division Duplex
TR	Technical Recomnadation
TS	Technical Specification
TX	Transmiter
Tx/Rx	Transceiver
U2N	UE-to-Network
UE	User Equipment
UL	Up Link
UP	User Plane
UPF	User Plane Function
vBS	Virtualised Base Station
VNF	Virtualised Network Functions
VR	Virtual Reality
V-RAN	Virtualised RAN
xAPPs	Cross-Functional Applications
WD-MIMO	Widely-Distributed Multiple-Input-Multiple-Output
WP	Work Package
XDP	eXpress Data Path

Executive Summary

BeGREEN proposes an evolved radio network that targets not only improved wireless communication's efficiency by maximising performance but also reducing energy consumption. It aims to go beyond the current 5G wireless system specifications to accommodate increasing traffic and services but also consider power consumption as a factor. Determining the metrics by which power consumption should be included is a key feature which will be studied as first stage of the project. This will include not only the cost of the energy but also societal factors.

This deliverable delves into describing current energy efficient frameworks and strategies that aim to model the radio access network (RAN) that serve for identifying and quantifying the parts of this network segment that can allow for improved energy efficiency and reduced energy consumption.

It also provides an initial description of the proposed use cases and their specific requirements, focusing on the scenarios that are subject to study and the reference Key Performance Indicators (KPIs) that will be later evaluated and captured in subsequent BeGREEN documents.

Additionally, this deliverable presents a high-level analysis of the BeGREEN RAN architecture along with the optimization strategies that can be applied to the RAN components, supported by those additional components that can are being currently considered in 3GPP latest releases (e.g. relays) and others that offer improved wireless communication at low energy cost.

Finally, it provides the energy control mechanisms that can be applied in a 5G disaggregated RAN and mechanisms enhanced by artificial intelligence and/or machine learning (AI/ML).

BeGREEN D2.1 "BeGREEN Reference Architecture" will serve as input for Task 2.2 and 2.3, as well as for work packages (WP) 3, 4 and 5.

1 Introduction

BeGREEN is a Smart Network and Services Joint Undertaking (SNS-JU) project with European funding [1]. SNS BeGREEN is one of the projects from the first SNS call in 2022, and kicked off on January 1, 2023. BeGREEN's focus is on introducing innovations to improve radio access network (RAN) energy efficiency.

The technical works are organised in three technical work packages (WPs), where the outline of the project's proposed system architecture, and requirements and key performance indicators (KPIs) as the criteria for evaluation of the work is aimed to be sketched in WP2, for which D2.1 is the first technical deliverable of the project. It reports the state-of-the-art for the RAN architectures from the energy efficiency point of view, while an initial analysis of the proposed solutions for RAN, at hardware, link, and system levels are presented. Additionally, it provides an introduction towards the overall framework for the project, in terms of the architecture, KPIs, and specific technical targets. Proposing a base architecture, and determining the metrics by which power consumption should be included are key initial steps for the project. BeGREEN WP2 aims to take these steps and facilitate the technical innovations planned for the technical work packages of the project.

1.1 Motivation

The demand on mobile networks will continue to grow and, without action, energy utilization and related emissions is also expected to grow. To reach the *net zero* it is important to reduce energy consumption and break the energy curve [2]. Energy consumption in mobile networks has increased over time, although throughput and capacity have increased more, by far. About the same increase in energy consumption is seen with each new mobile generation, as shown in Figure 1-1. The way mobile newtworks are planned, deployed and operated needs to be enhanced to break the increasing trajectory of energy consumption.

Next-generation networks, i.e. B5G and 6G, are introducing architectural transformations that originate 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. Energy consumption of telecommunication networks is increasing as the technologies evolve, as it is observed by comparing the 4th generation (4G) (long-term evolution, LTE), 5G (new radio, NR), and B5G networks energy performances. Existing methods and techniques are focused on energy savings with specific network configurations to yield the best results.

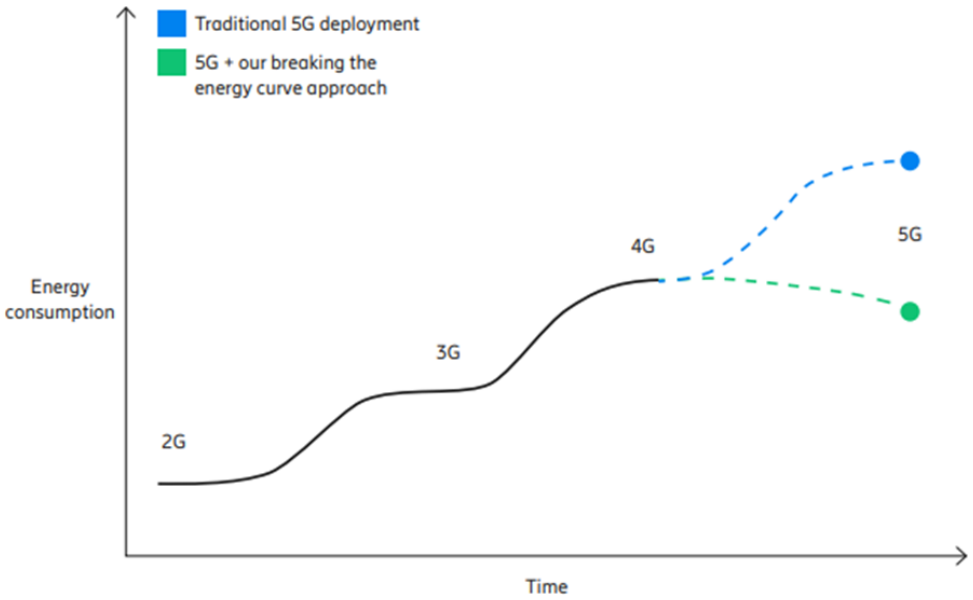


Figure 1-1 Breaking the energy curve [2]

With 73% of the network's total energy consumption occurring in the RAN [3], it is crucial to focus on RAN energy performance improvements to be able break the curve of energy consumption. The most straighforward techniques would be to avoid energy consumption when it is not required. Here the target is to automatically deactivate unrequired capacity during low and medium traffic loads. By performing advanced measurements that predict traffic patterns, traffic load, and end-user needs, from the cell down to subframe levels, it is possible to dynamically activate RAN processing and radio equipment to achieve the lowest possible energy consumption with maintained network performance. By using Artificial Intelligence (AI), service providers can operate the infrastructure more efficiently. For example, Ericsson's portfolio offers tools to control passive equipment, and enable predictive maintenance and no-touch problem-solving to reduce costs, site energy usage and site visits. Customer case-studies show that Mobile Network Operators (MNOs) have reduced site energy consumption by up to 15% through intelligent site control solutions [4].

Other common energy saving methods for wireless telecommunication equipment deployed in mobile networks are designed by using energy efficient hardware that can activate energy saving functions in RAN, core, and data centers. Activating energy saving functions is a widely used approach where autonomous network functions are used across a mobile network to enhance the energy efficiency. Among energy saving techniques one can mention *microsleep for transmitters*, *low-energy scheduler solution* (LESS), MIMO sleep mode, and cell sleep mode. To acquire the best potential for energy savings, suitable network function selection strategies (for example enabling selected energy savings functions at the appropriate sites), and specific network configurations are required in mobile networks structure.

1.2 Vision

BeGREEN will take a holistic view to provide evolving radio networks that not only accommodate increasing traffic and service levels but also consider power consumption as a factor. Determining the metrics by which power consumption should be included is a key feature of the project. An obvious first stage will be to consider the cost of the energy but also societal factors, linked to the necessary reduction in global emissions will also be considered.

The different mechanisms by which power consumption could be reduced will be evaluated. At a **hardware level**, techniques will include improvement of the energy efficiency of power amplifiers and the appropriate use of accelerators to reduce power consumption relative to the use of generic compute platforms when performing network function virtualisation. At the **link level**, techniques to provide a better estimate of the impact of the radio channel will be considered and the resulting improvement in spectral efficiency balanced against the increased power consumption associated with the resulting calculations. At a **system level**, new nodes and architectures are considered that move away from the assumption that additional capacity requires installation of new base-stations. These architectural considerations will include Reconfigurable Intelligent Surfaces (RIS) and cell-free architectures that could provide a better distribution of radio power around an area when compared to the centralised cellular approach.

To achieve the expected hardware, link-level and system-level benefits, research into some fundamental capabilities are required. The use of AI/ML techniques provides a solution to reduce the number of calculations required when compared to a more traditional approach. AI/ML can also be used to recognise patterns in the system level data associated with the behaviour of the user base and to learn the most appropriate response to this behaviour in terms of both network performance and also energy consumption. The location at which these AI/ML operations are carried out within the network will also have an impact on the performance of the approach, the consumption of power and the ability to share resources between different operations. The movement of data around the network to the appropriate location for calculations to be carried out also requires the definitions of new interfaces and protocols to enable the data to be processed in an open architecture of virtualised network functions (VNFs). For this reason, the project will assume that the emerging O-RAN standard and its evolution is the baseline architecture. The disaggregation,

virtualisation and network and service management capabilities inherent in O-RAN provide the mechanisms to realise many of the infrastructure changes and techniques for energy optimisation discussed above. BeGREEN will evaluate different mechanisms by which power consumption could be reduced based on the following pillars:

- At the architecture level, planning and evaluation of a RAN design to achieve flexible and energy efficient connectivity considering spectrum utilisation, interference mitigation and architecture/processing complexity.
- At the hardware and infrastructure level, radio-unit controlling schemes are used in power amplifiers energy optimisation. Also, an offloading engine for hardware acceleration will be employed to achieve energy efficiency when performing radio access functions and network function virtualisation.
- At the link level, the integrated sensing techniques are used to provide a better estimate of the impact of the radio channel toward improvements in spectral efficiency against the increased power consumption associated with the resulting calculations.
- At the system level, the project pursues the development and evaluation of AI-based procedures to adapt the energy consumption of softwarised network functions, aiming to minimize the overall consumed energy according to the utilisation patterns of network. BeGREEN proposes an "Intelligent Plane", as an additional plane along with user plane and data plane, that allows the data, model and inference to be seamlessly exchanged between network functions.

BeGREEN will use O-RAN as the baseline architecture, due to new/suitable interfaces and protocols it can provide, to be used for the movement of data around the network to the appropriate location for performance and efficiency assessments. In addition, the disaggregation, virtualisation and network and service management capabilities inherent in O-RAN provide the mechanisms to realise many of the infrastructure changes and techniques for energy optimisation pursued in BeGREEN.

BeGREEN will use AI and machine learning ML techniques to provide solutions for reducing the required calculations and to recognise patterns in the system level data associated with the behaviour of the user-base and to learn the most appropriate response to this behaviour in terms of both network performance and energy consumption. In this scheme, impact of the location of the AI/ML operations within the network on the performance of the approach, the consumption of power and the ability to share resources between different operations will be considered.

At the end of the project, BeGREEN technologies will be showcased in several in-lab or on outdoor testbed demonstrations. The technical innovations of the project will be worked out in WP3, and WP4, while the details of the demonstrations, planning, integration, and execution of the demos will be perfoemed in WP5.

1.3 Structure of the Document

BeGREEN deliverable D2.1, contains, based on the tasks defined for BeGREEN T2.1, a survey of the current state-of-the-art (SotA) energy efficiency methodologies which can serve as input for T2.2, T2.3, as well as for WP3, WP4, and WP5 works. BeGREEN D2.1 focuses on network architectures from the energy efficiecy perspective, energy-aware control mechanisms and AI or ML enhanced mechanisms. The document is structured as follows:

Chapter 2 presents energy efficient design choices and models and a high-level analysis of the RAN architecture.

Chapter 3 presents the reference scenarios and reference use cases followed by the definition of the preliminary KPIs.

Chapter 4 discusses BeGREEN system architecture, where a SotA analysis is performed, focusing on cellular mMIMO systems and distributed CF-mMIMO systems, the extension of coverage augmented by relays and RIS, and propose enhancements to existing O-RAN components, interfaces and functions related to RAN Intelligent Controller (RIC) to incorporate energy-efficiency awareness. Then, based on the previous analysis, an evolved architecture is presented for further detailed research in the BeGREEN project.

Chapter 5 introduces various energy-saving technologies considering the hardware and software aspects, with a special insight into the BeGREEN focus. BeGREEN will specifically work on AI-based modules of the radio unit (RU). Hardware acceleration will be another power-saving strategy which can contribute significantly, especially when computationally heavy processing is used. An analysis of the most suitable processes and platforms is introduced, and several strategies to reduce energy consumption are discussed. Finally, link-quality enhancements by incorporating integrated sensing and communication (ISAC) concept and relays in the RAN are introduced and the potential to improve energy efficiency in the RAN is discussed. Chapter 4 is finalised by introducing BeGREEN's proposed development on each of the discussed items.

Chapter 6 covers energy usage calculation, dynamic adaptation of energy consumption of software-based user plane network functions, energy efficiency (EE) aware coverage and capacity optimization in a beyond-5G (B5G) RAN, enhanced with relay nodes with focus on relay placement, energy efficient resource orchestration in virtualized RANs with shared computing infrastructure and AI edge services, service-aware energy-efficient RU control and means to allow the implementation of AI/ML native algorithms focused on the optimization of RAN functions to provide cloud native solutions involving RAN intelligent controller.

Chapter 7 finalizes this document providing a summary concluding the work done in WP2 so far.

2 Energy Efficiency Frameworks and Strategies for RAN

Energy efficiency in networks becomes a necessity rather than an optional goal. Around 70% of the overall network energy utilization in the network is in the RAN [5]. While the BeGREEN architecture focuses on enablers for the energy saving enhancements, it needs to be mindful of broader 6G evolution. In B5G/6G, energy saving will become a vital integrated feature, that will become essential elements in all market-leading automated intelligent networks, both for private and public service provision. This section presents energy efficient design choices and models and a high-level analysis of the BeGREEN RAN architecture.

2.1 General optimization strategies

Optimization of every aspect of network operations is a very large and complex study field. As a general principle, exact optimization is impossible, and heuristics are the only reasonable alternative. The following list summarizes the basic properties of the various approaches. These will be specialized in later chapters.

1. **Exact optimization** - for this approach we need to be able to write down a single global objective function (normally to be minimized), and a set of constraints, which could be simple bounds, or linear or nonlinear equality or (more commonly) inequality constraints. Unless the problem is convex (meaning that the objective function is convex and the feasible set is convex), it is unlikely that a good algorithm is available for the solution. Some radio engineering problems are convex, but large-scale optimizations such as those envisaged in BeGREEN are not convex. Moreover, in a real system it is unlikely that all the mathematical model input parameters are available at one instant.
2. **Local heuristics** - These are nearly always of the distributed asynchronous type. This means that many uncoordinated local heuristics are running, each locally optimizing some small part of the network operations. These can be thought as small negative feedback loops which stabilize some local aspect of the whole system. Typically, they compare the current value of a performance metric against a target value and adjust some system parameter to tune the performance closer to the target. When several such local heuristics are running simultaneously, there is a risk that they do not cooperate.
3. **Optimization with ML** – This is just another heuristic method that has been investigated in detail, for example in the recently completed AIMM project[1] [6]. The AIMM simulator is available for testing such methods [7]. However, a general conclusion from that project was that at the current SotA, ML does not perform better than local heuristics. Moreover, the computational load of ML is massively higher than local heuristics. There is a risk, when managing a network with the intention of minimizing energy use via ML, that the ML layer will dissipate more energy than it saves during the training of those technologies. This is yet to be determined according to the SotA; we present our work in this direction in section 6 of this document.

2.2 Energy-efficient design choices at RAN segment

The O-RAN 5G space is still in its infancy. There is a chance to make ground-up design choices. Energy consumption is becoming a key variable in the decision-making process for a telco deployment. Hence it is imperative that the right choices are made to keep the energy footprint of the solutions to a minimum.

Energy savings can be achieved in one of two ways:

[1] AIMM (AI-enabled Massive MIMO) is a CELTIC-NEXT European collaborative research and development project. The AIMM consortium targets radical performance improvements and efficiency dividends for 5G and beyond Radio Access Network (RAN), through advanced antenna array (Massive MIMO) and Reconfigurable Intelligent Surface (RIS) technologies, powered through and managed by the latest advancements in Artificial Intelligence (AI) and Machine Learning (ML).

- **Consolidation** – Adding more and more features in a single form factor to reduce the total power budget. Running in a single consolidated platform such as a server or a machine, versus multiple units, directly translates into energy savings. This way, no pressure is put on the computing efficiency of the machine, but the platform needs to be able to handle multiple workloads. Single-threaded performance of the compute core needs to be maximized, which translates writing efficient code and utilizing all the knobs that can be used in Hardware(HW)-Software(SW) to reduce footprint.
- **Energy-Efficient Hardware** – The energy footprint of the HW should be a primary design consideration. A choice needs to be made to ensure the right balance of performance and energy efficiency, for example choosing a highly optimal but high energy-consuming HW can lead to a higher energy budget.

The efficiency knobs, that is measurement and control, need to be part of the design to provide feedback and the right level of control to the applications to modify the solution energy consumption envelope . There are industry standard methods of power telemetry collection, for example getting data from sensors, register to measure power consumption, etc.; and control, for example changing the setting to modify power profile at server and system-on-chip (SoC) level as defined by Advanced Configuration and Power Interface (ACPI) standards [8] illustrated in Figure 2-1. The Operating System (OS) and application have clear Application Programming Interfaces (APIs) to access the sensors telemetry feedback and defined methodology to control power states.

The state transitions for Idle power management and Collaborative Processor Performance Control (CPCC) which is Dynamic Voltage and Frequency Working Power state management as stated by the ACPI standard.

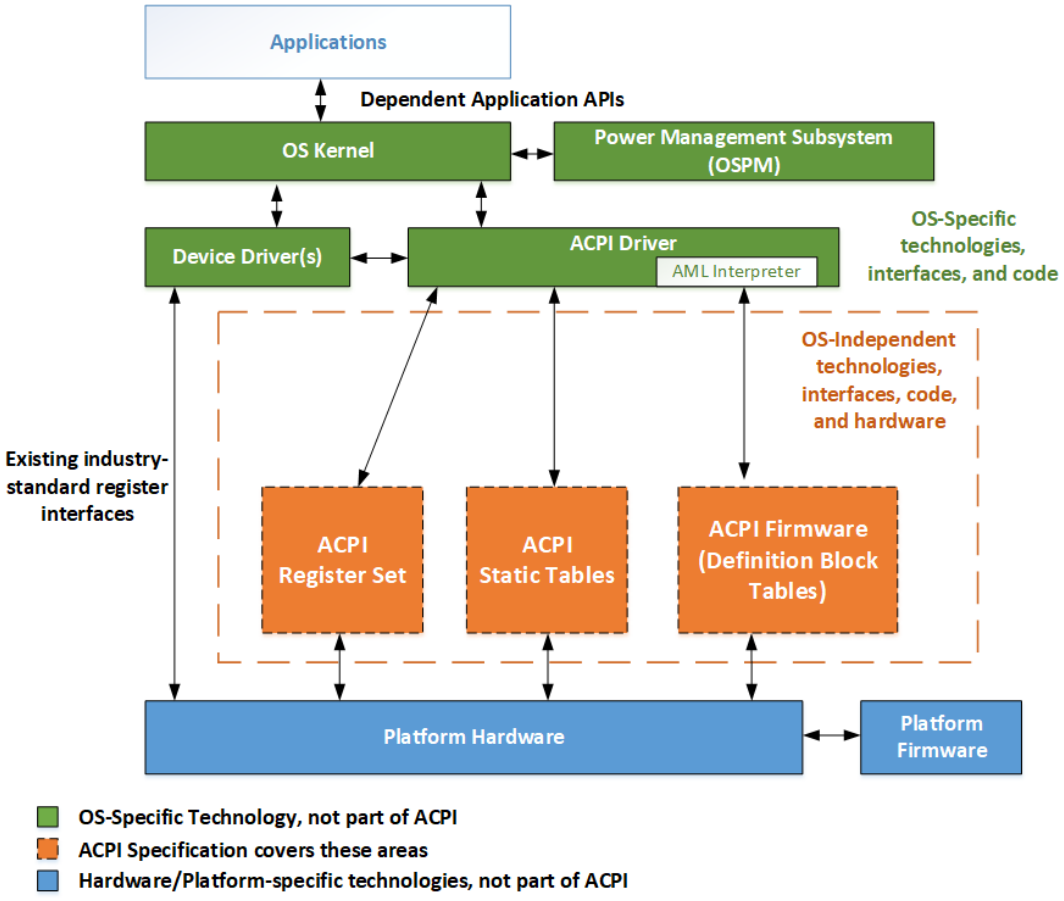


Figure 2-1 Power telemetry collection as defined by ACPI standards

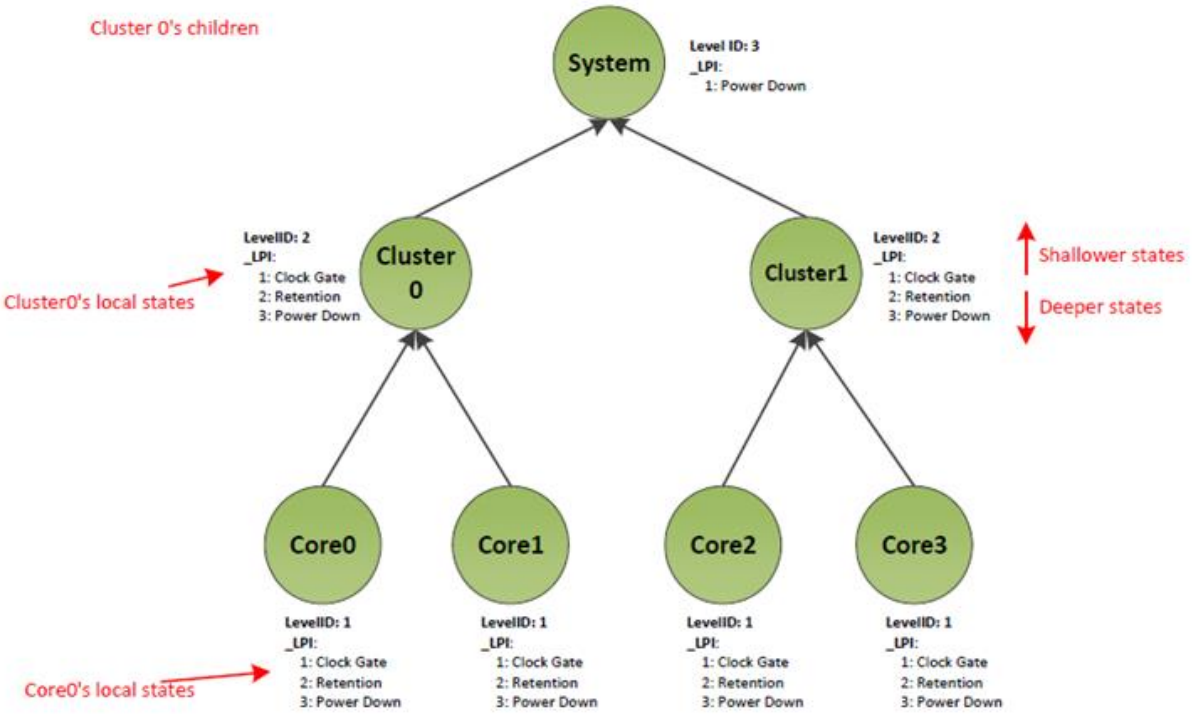


Figure 2-2: Idle power state management

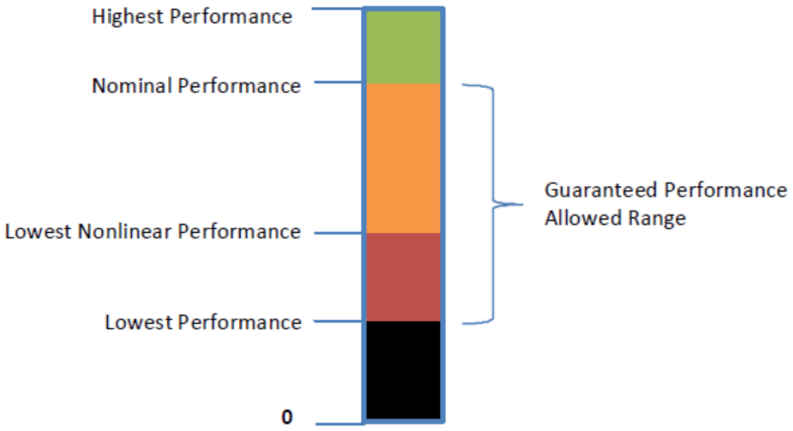


Figure 2-3: Performance power control threshold

The idle-power management diagram shown in Figure 2-2 describes the procedure by which a SoC with large number of core clusters are systematically set to various low power levels. The power states are chosen by the sleep level the system should be set to, which is a function of how quickly the workload (for example traffic) is expected to emerge out of idle mode.

The performance power control method, as shown in Figure 2-3, manages the power characteristics of the system by modifying the frequency and voltage. Depending on the load of the system and the charactersistics chosen (as shown in the diagram) for the workload, the system will try to settle down to the most optimal power profile. The power manager (user, application, use case) can select thresholds which describe performance level (high, nominal and low) and the power control mechanism sets the energy consumption to the right level for the current state.

2.3 RAN architectural starting point for BeGREEN

In the BeGREEN project, we based our proposed architecture on the O-RAN Alliance architecture [18], where

we seek to make all our technological enablers and components to be compatible with the O-RAN philosophy. In this context, it is important to discuss the four main principles of O-RAN:

- Openness: The interfaces shared by the different components must be compatible with all the various vendors supporting each function. This allows the integration of functionalities developed by companies/research institution to be implemented in any vendor device.
- Virtualization: The network functions integrated in proprietary hardware in legacy systems must be migrated to cloud platforms running in virtualized or containerized instances. This allows for hardware and software disaggregation where different commercial off-the-shelf (COTS) hardware can be selected, based on use case requirements and central processing unit (CPU) requirements.
- Intelligence: In order to optimize the network, the O-RAN has three control loops with different timings (to be introduced later in this section). These allow the application of cross-layer ML/AI algorithms to optimize radio resource management at the granularity of the user equipment (UE).
- Programmability: Based on continuous monitoring of the network interfaces, pre-trained AI/ML mechanisms can be programmatically configured to adapt to schedules, traffic loads or policies, among others.

Once we have learned the objectives of the O-RAN Alliance, we can show how the different components and interfaces are introduced in their system. The O-RAN architecture, including components and interfaces, is depicted in Figure 2-4.

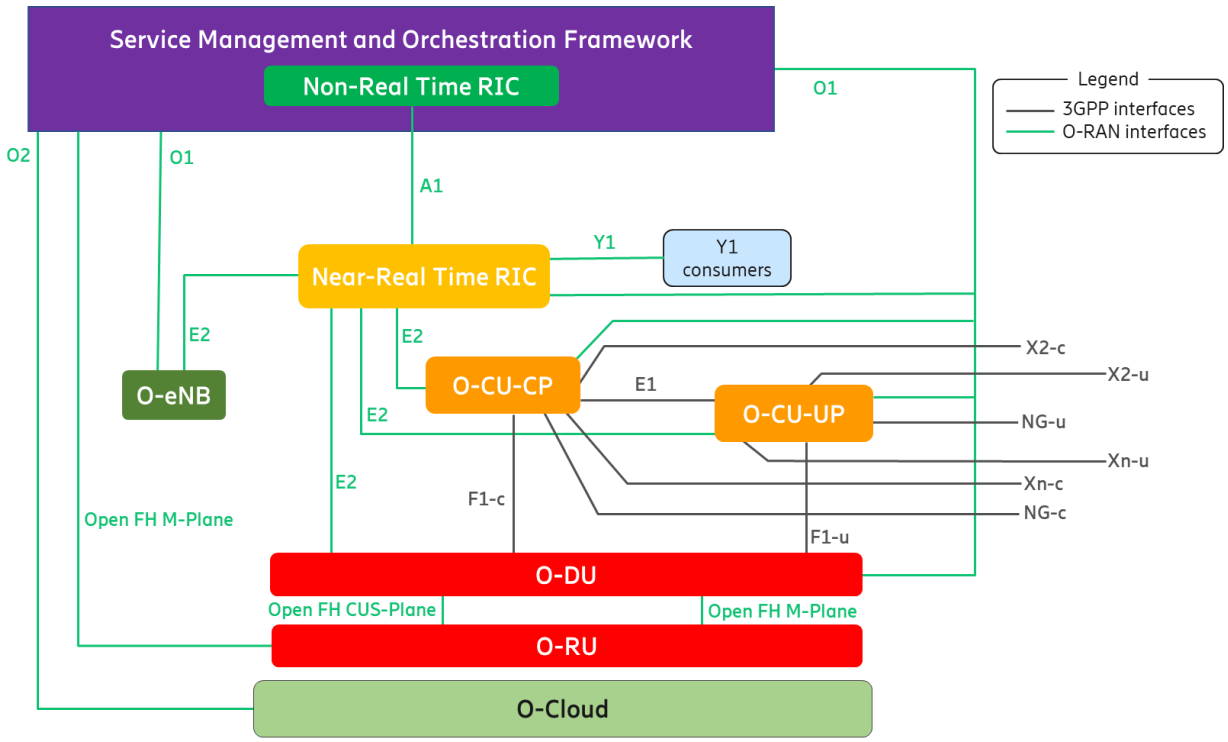


Figure 2-4 O-RAN logical architecture and control loops [9]

Following the description from [9], O-RAN components can be described as follows:

- Near-real-time RAN intelligent controller (**Near-RT RIC**)- This is a logical function which enables near real-time control and optimization of RAN elements and resources through fine-grained data collection and actions via the E2 interface. It can incorporate AI/ML workflow including model training, inference, and updates. It provides a platform for implementing new network functions and services, including traffic management, network slicing, and radio resource management. The near real-time aspect of the RIC refers to its ability to process and analyse data quickly, allowing for rapid

decision-making and network optimization [10]. In their latests specifications, the O-RAN Alliance has incorporated the Y1 interface, which will be used to expose RAN analytics produced in the near-RT RIC to external components such as the service management & orchestrator (SMO), the 5G Core or the radio network information service (RNIS) in multi-access edge computing (MEC) architectures.

- Cross functional Application (**xApp**)- This is an application designed to run on the near-RT RIC. Such an application is likely to consist of one or more microservices and, at the time of on-boarding, it will identify which data it consumes and which data it provides. The application is independent of the near-RT RIC and may be provided by any third party. The E2 provides a direct link between the xApp and the RAN functionality. xApps are developed using open APIs and are designed to be easily deployed and managed. They can be developed by network operators, third-party developers, or system integrators. The O-RAN Alliance promotes the development of a wide range of xApps to encourage innovation and drive the adoption of O-RAN solutions [10].
- **Non-RT RIC** (Non-real-time RAN intelligent controller)- This is a logical function within the SMO that controls the content carried over the A1 interface. It complements the near-RT RIC by providing long-term optimization and planning functions for the RAN. It enables network operators to manage network resources over longer periods, such as weeks or months, and provides insights into network performance and capacity planning. The non-RT RIC consists of the non-RT RIC framework and the non-RT RIC applications (rApps) [11].
- **Non-RT RIC Framework-** It involves the functionality internal to the SMO that logically terminates the A1 interface to the Near-RT RIC and exposes to rApps, via its R1 interface, the set of internal SMO services needed for their runtime processing. The Non-RT RIC Framework functionality within the Non-RT RIC provides AI/ML workflow including model training, inference and updates needed for rApp. The framework provides a standardized platform for implementing non-real-time RIC functions in the radio with a set of common interfaces and protocols that allow RAN components to communicate and exchange data. It also includes a common data model for representing network topology, resource allocation, and other relevant information [12].
- **rApps-** rApps are modular applications that leverage the functionality exposed via the Non-RT RIC Framework's R1 interface to provide added value services relative to RAN operation, such as driving the A1 interface, recommending values and actions that may be subsequently applied over the O1/O2 interface and generating "enrichment information" for the use of other rApp. The rApp functionality within the Non-RT RIC enables non-real-time control and optimization of RAN elements and resources and policy-based guidance to the applications and features in Near-RT RIC, enabling network operators to perform long-term network planning, optimization and management functions [12].
- **NMS-** A network management system (NMS) for the O-RU as specified in [13] to support legacy Open Fronthaul M-Plane deployments. It is a software application that enables network operators to manage and monitor the RAN. NMS provides a unified view of the network and its components, enabling operators to monitor network performance, identify and diagnose faults, and perform configuration and maintenance tasks. NMS typically includes features such as performance monitoring, fault management, configuration management, and security management. It may also include tools for capacity planning, traffic engineering, and service provisioning. The Open RAN Alliance recognizes the importance of NMS in managing and optimizing the radio access network, and promotes the development of open and interoperable NMS solutions to support the adoption of open RAN architectures [13].
- **O-Cloud-** That is a cloud computing platform comprising a collection of physical infrastructure nodes that meet O-RAN requirements to host the relevant O-RAN functions (such as Near-RT RIC, O-CU-CP, O-CU-UP, and O-DU), the supporting software components (such as Operating System, Virtual

Machine Monitor, Container Runtime, etc.) and the appropriate management and orchestration (MANO) functions. This is, the cloud infrastructure that supports O-RAN architectures, allowing operators to virtualize RAN functions and deploy them as software applications in the cloud [14].

- **O-CU-CP** (O-RAN Central Unit)- Control Plane is a logical node hosting the radio resource control and the control plane part of the Packet Data Convergence Protocol (PDCP). The control plane functions of the O-RU are responsible for managing the configuration, control, and coordination of the O-RU with other RAN components, such as the distributed unit (DU) and centralized unit (CU). This includes functions such as synchronization, channel estimation, and uplink (UL) and downlink (DL) scheduling.
- **O-CU-UP** (O-RAN Central Unit)- User Plane is a logical node hosting the user plane part of the PDCP and the service data adaptation protocol (SDAP). The O-CU-UP is responsible for processing and forwarding user data between the core network and the RAN, and for performing functions such as packet routing, scheduling, and buffering.
- **O-DU**- O-RAN Distributed Unit is typically located close to the radio heads and is responsible for performing functions such as baseband processing, modulation and demodulation, and signal conditioning. It interfaces with the remote radio heads (RRHs) and the centralized unit (CU) to exchange control and user plane data. The O-DU is designed to be a modular and scalable component that can be easily deployed and configured in a variety of network environments, and can be used to support a wide range of radio access technologies, including 4G and 5G. The O-DU is a logical node hosting radio link control (RLC), medium access control (MAC) and high physical layer (PHY) based on a lower layer functional split
- **O-eNB-** An O-RAN evolved Node B is a eNB [15] or ng-eNB [16] that supports E2 interface. Being this element responsible for radio resource management, scheduling and coordination of wireless transmissions, and the management of the UE connection.
- **O-RU** (O-RAN Radio Unit)- This is a logical node hosting low-PHY layer and radio frequency (RF) processing based on a lower layer functional split. This is quite similar to Third Generation Partnership Project (3GPP) transmission/reception point (TRP) or remote radio head (RRH), but more specific in including the low-PHY layer. The O-RU is a key component of the open RAN architecture that interfaces with the antennas and performs RF signal processing functions. The O-RU is responsible for converting the digital baseband signals from the DU into analog RF signals that are transmitted over the air to the UE, and for receiving and converting analog RF signals from the UE into digital baseband signals that are sent to the DU for further processing.
- **SMO-** The SMO is a key component of the NMS that is responsible for managing and controlling the end-to-end services provided by the network. SMO can refer to the MANO functions that are responsible for configuring, deploying, and managing the various components of the RAN, including the DU, CU, RU, and the transport network. This may involve tasks such as resource allocation, network slicing, and service provisioning.
- **Telemetry exposure**- From the previous element descriptions and from Figure 2-4, we can understand that there are different levels of control loops depending on the timing requirements of the use cases. The three main loops (but not limited to them) are: non-real time (>1 second), near-real time (>10 ms), and real time (<10 ms) control loops. These control loops can interact with each other as required. The exposure of telemetry to consumers, like SMO data-lake, become increasingly important towards 6G, as in network digital twins (NDT), etc, but also in 5G such as for data-sets for AI/ML re-training for intelligent energy saving. The O-RAN Y1 interface from Near-RT RIC, and the O1 high-volume VNF Event Stream (HV-)VES are O-RAN ways of exporting telemetry, to complement the 3GPP 5GC network data analytics function (NWDAF), in a consolidated data-lake, as considered in BeGREEN. In order to tackle these, a telemetry framework must be designed to support the increasing demand of data traveling across RAN elements, wich is detailed in Section 2.3.2.

2.3.1 Additional BeGREEN network elements

Even though the BeGREEN project takes O-RAN as the reference baseline, other network technologies and components will be integrated looking for enhanced network performance and energy savings. This network elements are defined as follows:

- **Edge Computing-** That is a distributed computing paradigm that brings processing closer to the user rather than in centralised data centres. This requires the deployment of distributed data centres across the RAN, closer to the user, to reduce latency, improve response times and increase network efficiency.
- **MC-UE-** Multi-Connectivity User-Equipment refers to a UE that can connect to multiple base stations BSs simultaneously. This way can provide uninterrupted connectivity in scenarios that the network conditions are challenging, resulting to a better user experience improving the signal strength, network capacity and reducing latency.
- **Relays-** A relay is a distributed node that acts as a wireless repeater, retransmitting the wireless received signal. Relays are used to augment the coverage and strengthen signal quality where these are not optimum. Being more cost-effective than deploying new BSs .
- **RIS-** RIS is a planar surface consisting of a large number of passive elements that can manipulate the channel impulse response reflecting or manipulating electromagnetic waves in real-time. This technology can enhance wireless communications by controlling the propagation environment as desired, manipulating the radio waveforms to avoid interferences or extend communication links passively.

2.3.2 Telemetry framework supporting 6G evolution

One common subsystem of 6G networks will be the telemetry and data collection frameworks, ingesting, processing and leveraging the 6G telemetry, where telemetry is used in this case to describe all sorts of 'big data' measurements and sensing derived state, meta-data, etc. These will be needed to enable the 6G RAN intelligence, for (re-)training and inference, to allow AI/ML to integrate and automate these zero-touch networks, optimizing for energy saving but also providing a future 6G evolution pathway to additionally optimize for electromagnetic field exposure, and to integrate 6G technology enablers like ISAC, RIS, cell-free, dynamic spectrum re-farming, NDT, among others, as described in Chapter 3.

A key driver for 5G O-RAN was to introduce intelligence into the RAN through the O-RAN RIC, with both the Near-RT and Non-RT RIC network functions, extensible through xApps/rApps. How an evolved 6G O-RAN will be better integrated into 3GPP 6G networks is currently beyond SotA and beyond BeGREEN scope. 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, in follow-up SNS actions and complementary 6G research projects, and later towards a higher Technology Readiness Level (TRL), as a commercial market success.

Figure 2-5 describes the general high level O-RAN telemetry framework for future 6G networks. As described below, the 6G Technology Enablers will provide a vast amount of information in the southbound direction, which can be connected to the native AI/ML Near-RT RIC to provide control and sensing for the xApp repository. Accordingly, this sensing data must be exposed to the BeGREEN intelligence datalake where integrated with other non-3GPP telemetry will complement the needed information for the proposed framework.

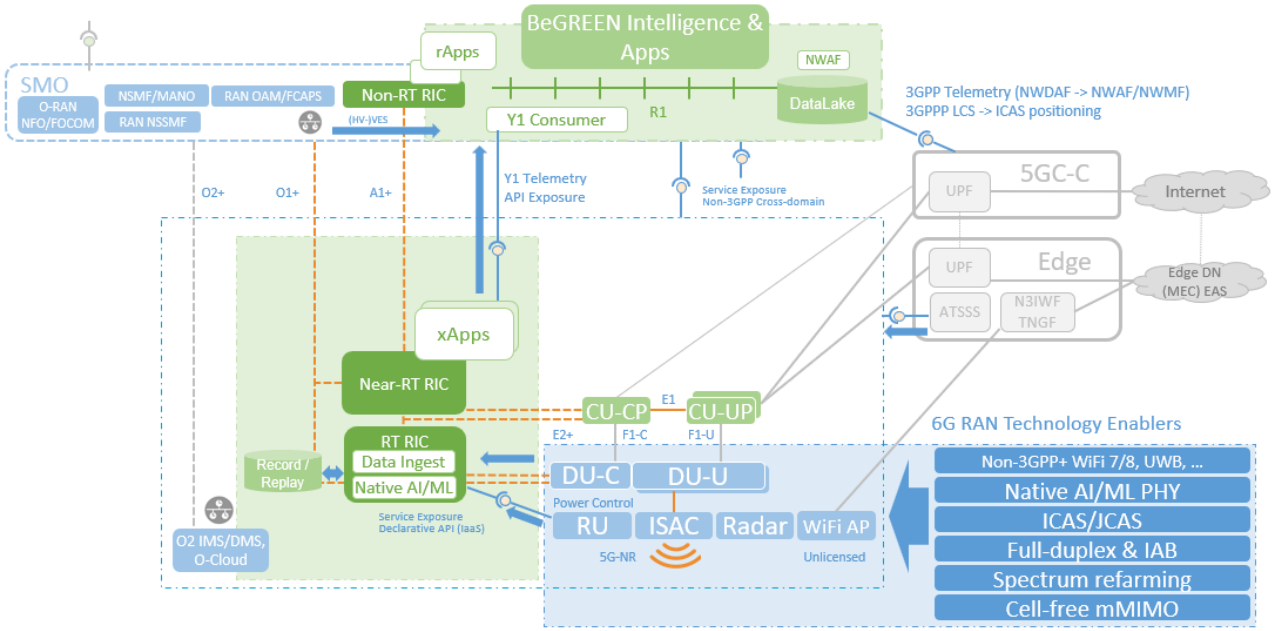


Figure 2-5 BeGREEN high-level Open RAN telemetry framework ready for future 6G exploitation

Whilst 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 multiradio access technology like Wi-Fi 7 or 8, but also ultra-wide band and LIDAR technologies. While in 3GPP 5G the focus was on UP integration towards the user plane function (UPF) , it is expected that the positioning capabilities will increasingly be injected into the RIC to provide enhanced cross-domain functionality, beyond SotA 3GPP (See Section 4.1.1).
- **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, the real intelligence will become a reality if it is deeply embedded into the 6G native AI/ML. This is the scope of the identified but unspecified real-time RIC (RT RIC), which is outside BeGREEN scope but will be needed for future 6G capabilities.
- **ISAC-** The addition of radar-like sensing, integrated into the communication radio units (see Section 5.3.2) will require processing and usage of this sensing information, to merge with other higher-level telemetry for advanced functionality.
- **Full-duplex and integrated access and backhaul (IAB)-** These technologies will extend 6G RAN coverage and will add additional interference management and bandwidth scheduling that will result in increased telemetry to be ingested and utilized (See Section 4.1.2).
- **Dynamic spectrum re-farming-** It will require full sensing of the channel-state information (CSI), and neighbour interference, to ingest and process the interference extracting valuable telemetry for optimization and co-existence instead of being discarded as pure noise.
- **Cell-free mMIMO-** This technology requires close coordination of DU and RU, with dynamic reconfiguration of the user-centric Access Point (AP) associations, for coherent multipoint transmit/receive and synchronization. Early research results show the need for RIC-like intelligence [17].

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. 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 next Generation Research Group (nGRG) [18] and is like Infrastructure as a Service (IaaS) design patterns.

All these telemetries need consolidation and 'northbound' egress towards the SMO, with datalakes and other types of distributed event store, telemetry unification and storage units, to provide the O-RAN R1 interfaces towards higher level intelligent agents and 6G NDTs, running in the cloud-native intent-based intelligence plane.

2.4 RAN energy efficiency modelling

BeGREEN system architecture will follow O-RAN concepts. However, previous pre-O-RAN based architectures, needs to be considered and reviewed. Recently, significant attention has been put on reducing the energy consumed by the BSs and, as a result, various BS power consumption models have been proposed and investigated.

In general, the initial analysis follows the formulation of two important works that is Auer et al. [19], and Israr et al. [20]. In effect, this section forms a literature review. One aim is to evaluate existing models for implementation on the AIMM simulator[2] [7]. The AIMM simulator at present does not support standardized O-RAN interfaces, though it does implement a RIC module which can host xApps through a non-standard interface. The present work is part of an evaluation of which features need to be added. Other useful recent surveys on energy use in networks are [21][22][23].

2.4.1 The model of Auer et al.

We summarize here the mathematical models for energy consumption proposed in [19]. These are intended as a basis for further development in the BeGREEN project. An architecture diagram outlining what Auer et al. call the EARTH energy efficiency evaluation framework is shown in Figure 2-6. It contains, in the lower layers, power models which feed reports into higher-layer global metrics.

This model aims to capture the entire energy consumption of all network components, but recognizes that about 80% of energy use is consumed at BS sites. The authors state that their principal enhancements over previous models are:

1. A power model that maps the RF output power radiated at the antenna elements, P_{out}, to the total supply power of a BS site, P_{in}. The power model defines the interface between the component and system levels.
2. Long-term traffic models that describe load fluctuations over a day and complement statistical short-term traffic models.
3. Large-scale deployment models that extend existing small-scale deployment scenarios to large geographical areas.

The first equation relates total power amplifier (PA) consumption (P_{PA}) to actual power output (P_{out}) via an efficiency factor η (typical value 30%) and a feeder loss σ_{feed}.

[2] The AIMM simulator emulates a cellular radio system roughly following 5G concepts and channel models. The intention is to have an easy-to-use and fast system written in pure Python with minimal dependencies. It is especially designed to be suitable for interfacing to AI engines.

$$P_{\mathrm{PA}} = \frac{P_{\mathrm{out}}}{\eta \cdot (1 - \sigma_{\mathrm{feed}})}$$

Auer et al. next assume that losses incurred by DC-DC power supply, mains supply, and active cooling, scale linearly with the power consumption of the other components, and may be approximated by the loss factors σ_{DC}, σ_{MS}, and σ_{cool}, respectively. This gives the next equation. The σ factors are the losses due to DC power conversion, mains supply, and cooling respectively. The overall factor N_{TRX} is the number of transceiver chains. P_{BB} is the power used for baseband processing. See Figure 2-7 for typical values for these parameters. Further requirement are user mobility models and traffic models, ideally parameterized from real system data.

$$P_{\mathrm{in}} = N_{\mathrm{TRX}} \cdot \frac{\frac{P_{\mathrm{out}}}{\eta \cdot (1-\sigma_{\mathrm{feed}})} + P_{\mathrm{RF}} + P_{\mathrm{BB}}}{(1 - \sigma_{\mathrm{DC}})(1 - \sigma_{MS})(1 - \sigma_{\mathrm{cool}})}$$

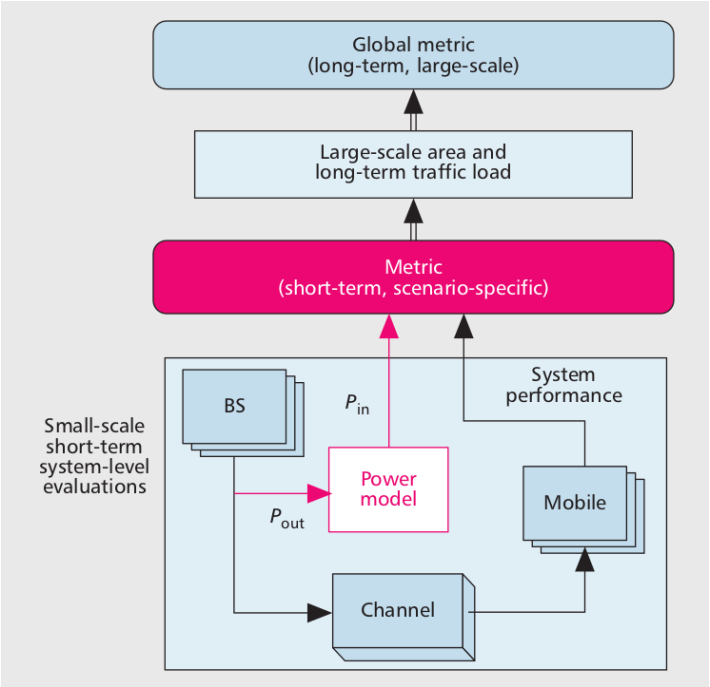


Figure 2-6 EARTH energy efficiency evaluation framework ([19], Figure 1)

			Macro	RRH	Micro	Pico	Femto
BS	Max Tx power	[dBm]	43.0	43.0	38.0	21.0	17.0
	(average) P_{max}	[W]	20.0	20.0	6.3	0.13	0.05
	Feeder loss σ_{feed}	[dB]	–3	0	0	0	0
PA	Back-off	[dB]	8.0	8.0	8.0	12.0	12.0
	Max PA out (peak)	[dBm]	54.0	51.0	46.0	33.0	29.0
	PA eff. η_{PA}	[%]	31.1	31.1	22.8	6.7	4.4
	Total PA, $\frac{P_{max}}{\eta_{PA} \cdot (1 - \sigma_{feed})}$	**[W]**	**128.2**	**64.4**	**27.7**	**1.9**	**1.1**
RF	P_{TX}	[W]	6.8	6.8	3.4	0.4	0.2
	P_{RX}	[W]	6.1	6.1	3.1	0.4	0.3
	Total RF, P_{RF}	**[W]**	**12.9**	**12.9**	**6.5**	**1.0**	**0.6**
BB	Radio (inner Rx/Tx)	[W]	10.8	10.8	9.1	1.2	1.0
	Turbo code (outer Rx/Tx)	[W]	8.8	8.8	8.1	1.4	1.2
	Processors	[W]	10.0	10.0	10.0	0.4	0.3
	Total BB, P_{BB}	**[W]**	**29.6**	**29.6**	**27.3**	**3.0**	**2.5**
DC-DC, σ_{DC}		[%]	7.5	7.5	7.5	9.0	9.0
Cooling, σ_{cool}		[%]	10.0	0.0	0.0	0.0	0.0
Mains supply, σ_{MS}		[%]	9.0	9.0	9.0	11.0	11.0
Total per TRX chain		**[W]**	**225.0**	**125.8**	**72.3**	**7.3**	**5.2**
# Sectors		#	3	3	1	1	1
# Antennas		#	2	2	2	2	2
# Carriers		#	1	1	1	1	1
Total N_{TRX} **chains,** P_{in}		**[W]**	**1350**	**754.8**	**144.6**	**14.7**	**10.4**

Figure 2-7 Typical system parameters from [19]

2.4.2 The model of Israr et al.

By contrast to the energy model of Auer [19], the model of Israr et al. [20] embraces a full 5G paradigm, with O-RAN-like features (though that term is not explicitly mentioned).

Israr et al. study the energy consumption in the 5G mobile network infrastructure. In particular, they are concerned about modelling the increased traffic demand from a large number of end-users with greater traffic volume and data rate. Appropriate propagation models are used. They also look at energy consumption in millimetre wave (mmWave) communication and mMIMO antennas. Israr et al. cover the following scenarios:

1. A multiband two-tier heterogeneous network (HetNet). This consists of a high power node (HPN) which is a macrocell, and several low power nodes (LPN), which are picocell BSs, is deployed. The HPN employs the traditional LTE frequency band and LPNs operate a the mmWave frequency band.
1. A cloud radio access network (C-RAN). Here baseband processing for all cells (pico and macro) is deferred to centralized BBU (baseband units), separated from remote radio heads.
2. A heterogeneous cloud radio access network (H-CRAN). Here centralized BBUs support oly the macrocell.

The factors considered by Israr et al. as inputs to the modelling process are:

- traffic load.
- number of antennas.

- bandwidth.
- density of LPNs.
- energy consumption of the RRHs.
- energy consumption of the BBU pool.
- energy consumption of fronthaul.
- energy consumption of macro BS (MBS).
- energy consumption of small cell BSs (SCBs).

The authors conclude that C-RAN is the most energy-efficient RAN architecture due to its cooperative processing and decreased cooling and site-support devices, and H-CRAN consumes most of the energy compared to other 5G RAN architectures mainly due to a high level of heterogeneity. The Israr model is very comprehensive, with a large number of parameters, as can be seen from the list in Figure 2-8.

It is intended that the BeGREEN project will implement the Israr model in full as Python code, and verify the results before building further upon these models. Note that such a project would be independent of any simulation-based approach. For makinge predications from the Israr model there is no need need for any simulation.

Notations	Description	Notations	Description
m	Number of HPNs	P_p	Transmission power of the pth base station
p	Number of LPNs	Λ_p	Antenna gain of transmitter
s	Number of sectors	Λ_i	Antenna gain of receiver
si	Number of users	N_o	Noise power density
k_m	Number of low bandwidth BBUs	NF	Noise figure
r_m	Number of low bandwidth RRHs	Φ	Co-channel interference
k_p	Number of high bandwidth BBUs	P^M_{max}	Maximum transmission power of HPN
r_p	Number of high bandwidth RRHs	P^P_{max}	Maximum transmission power of LPN
B_m	Bandwidth of HPNs	P_o	Static power
B_p	Bandwidth of LPNs	P_c	Total circuit power consumption
A_{HPN}	Number of HPN antennas	P_B	Baseband processing power
A_{LPN}	Number of LPN antennas	P_F	Fronthaul power
$L(d)$	Path loss in dB	P_{PA}	Total power of power amplifier
d	Distance between BS and mobile user in meters	P_{RF}	Total power of radio frequency
γ	Floating-intercept	RF_n	Fronthaul traffic between RRH and BBU.
α_m	Path loss exponent of the mm-wave link	C_n	Capacity of fronthaul link
α_u	Path loss exponent of the LTE link	ζ	Power consumption coefficient due to amplifier and feeder losses
λ_m	Carrier wavelength of mm-wave link	ξ	Traffic load scaling factor
λ_u	Carrier wavelength of LTE link	σ_{DC}	DC–DC conversion loss
Υ_m	Lognormal shadowing of mm-wave links	σ_{MS}	AC–DC main supply conversion loss
Υ_u	Lognormal shadowing of LTE links	σ_{cool}	Cooling power loss
σ^2	Variance of lognormal shadowing	β	Bandwidth factor
$h_{p,i}$	Small-scale channel fading		

Figure 2-8 Definition of the model parameters [16]

Path Loss Parameters			
$fc(mm-wave)$	28 GHz	$fc(LTE)$	2.5 GHz
α_m	2.92	α_u	3.3
σ_{γ_m}	8.7	σ_{γ_u}	7.38
λ_m	10.7 mm	λ_u	124.9 mm
γ	72		
System Parameters			
B_m	400 MHz	B_u	20 MHz
A_{LPN}	2	A_{HPN}	2 to 20
d_m	50 m	d_u	350 m
N_o	−174 dBm/Hz	NF	6 dB
Power Consumption Parameters			
P_{max}^{LPN}	30 dBm	P_{max}^{HPN}	43 dBm
P_B^{LPN}	3 W	P_B^{HPN}	29.4 W
P_{RF}^{LPN}	1 W	P_{RF}^{HPN}	12.9 W
σ_{DC}^{LPN}	0.09	σ_{DC}^{HPN}	0.075
σ_{MS}^{LPN}	0.11	σ_{MS}^{HPN}	0.09
σ_{cool}^{LPN}	0	σ_{cool}^{HPN}	0.1
η_{PA}^{LPN}	0.067	η_{PA}^{HPN}	0.31

Figure 2-9 Definition of the typical values for the system variables [16]

3 Reference Use Case Criteria

In this chapter, the reference use cases which will be considered in BeGREEN are presented. These use cases will be the reference points from which BeGREEN consortium will develop solutions related to energy efficiency. The aim is to provide the reader with a comprehensive insight into the challenges of BeGREEN's research by examining these reference use cases. The organisation of the chapter is as follows: firstly, a complete list of all the different use cases is offered and their challenges related to energy efficiency are introduced. The use cases are separated into two different groups, that is, i) reference use cases related to developing energy efficiency mechanisms at the physical layer (related to WP3 PHY Energy Efficient Solutions), and, ii) reference use cases related to the mechanisms for improving the energy consumption at the system level (related to WP4 AI-assisted O-RAN-based edge and NFV energy optimisation). Secondly, an overview of different reference scenarios, to be worked out in the project, are provided. This includes a description of the scenarios and their relevance. Further, these scenarios are mapped to use cases related to the PHY optimizations. Finally, a preliminary overview of the standard 5G KPIs is discussed. These will be used in upcoming deliverables to define the project KPIs.

3.1 Reference Use Cases

3.1.1 Physical layer reference use cases

In what follows, reference use cases related to energy optimization in PHY are provided.

3.1.1.1 B5G energy efficiency enhanced RAN through relay nodes

Energy-efficient RAN is an added challenge for mobile network operators (MNOs) already facing significant expenditures in deploying 5G RAN infrastructure to achieve the level of densification required for meeting the demands of new services. This will be exacerbated in B5G scenarios, where operation in high-frequency bands with poor propagation conditions is envisaged. The energy that these dense infrastructures will consume will dramatically increase the costs for the MNO [24].

In that context, the use of relay nodes in future B5G RAN is a cost-efficient option for energy saving and, consequently, reducing operational expenditure (OPEX) for MNOs through the reduction of transmit power consumption in the mobile networks. At the same time, UEs which use the relays consume less battery since they transmit less power when they are connected through the relay compared to when directly connected to the BS. The reduction in transmit power, and the consequent reduction in the power consumed, resulting from the use of relays, has been identified in some works as a means to improve energy efficiency in wireless networks thanks to better propagation conditions in the involved links. In [25] the authors carried out an analysis of a joint optimization of relay station (RS) placement and RS sleep/active probability strategy in order to successfully enhance the overall energy efficiency. Other works, such as [26], [108], [27] and [28] also highlight the improvement of energy efficiency in networks with relay nodes deployed. Similarly, works in [29] and [30] investigate the problem of user association to macro or RS to minimize the system total energy consumption. Furthermore, in [31] the authors carried out an assessment of the energy savings which can be achieved by means of relaying in a wireless system. In that context, the 3GPP has identified new requirements for relays for energy efficiency [32].

The deployment of relay nodes in wireless networks offers different approaches. The first approach involves the use of fixed relays, where the MNO chooses the position of the relay as an extension of infrastructure as described in [33]. The second approach consists of installing relays within a moving element (Moving Relay), such as a bus, a train and so on, as exemplified in [34]. Lastly, involves equipping UEs with relay functionality, taking advantage of the technological evolution of UEs, to turn them into relay nodes when needed (relay-UE) [35]. Standardization for relay support in 5G has been recently introduced by 3GPP in the so-called IAB

technology of Rel-16. This considers only fixed relays, while mobile IAB is currently under study in Rel-18 [36][37]. Also, efforts to define architecture enhancements for vehicle-mounted relays are being made in 3GPP [38].

With all the above, this use case assumes a RAN with different RSs that are used to enhance the coverage conditions and thus reduce the required transmission power and energy consumption. Figure 3-1 provides a visual representation of a BS within this RAN, illustrating the deployment of different relays to improve user experiences. For example, there is a fixed relay at the edge of the coverage area that provides signal to a user who is a little further away, thus extending the coverage area of the BS (yellow area). Another example is a relay on the train, which gives connectivity to the users who travel on it, improving the services they experience (for example improving throughput). Moreover, there is a user located behind a building that produces shadowing losses, so it connects to a fixed relay or a UE with relay capabilities that provides connectivity with much lower power consumption since it transforms one non-line-of-sight (NLoS) link into two shorter line-of-sight (LoS) links. And finally, there is a UE in a sporadic coverage hole (pink area) which maintains its connectivity thanks to another UE which has relay functionality and gives coverage to it. In all cases, the UEs which are connected to the relays transmit less power than if they were connected directly to the BS, and at the same time the BS transmits less power to reach the relays than the power it should transmit to reach the UEs directly.

The proposed use case location is the University Campus of UPC in Barcelona, which corresponds to an urban macro scenario. The considered environment is a 350 m x 125 m area with 25 buildings of 3 floors as depicted in Figure 3-2. 5G NR coverage on the Campus is provided by three outdoor macrocells of a public MNO in band n78 (3.3-3.8 GHz) [39].

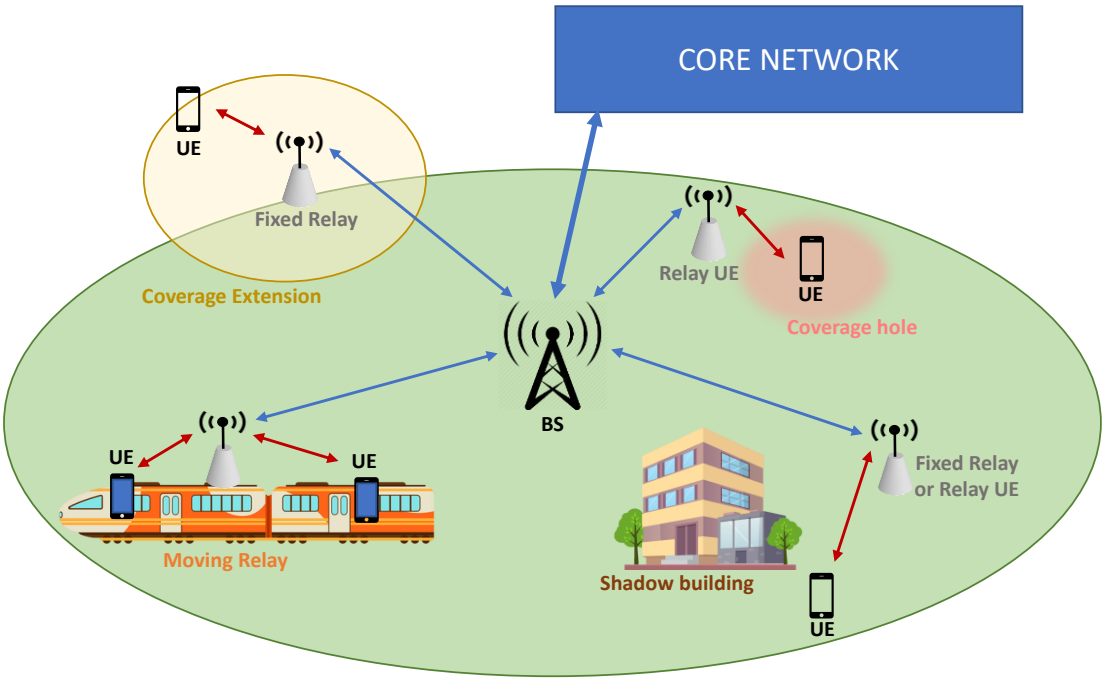


Figure 3-1 BS area with different relays deployed to enhance coverage conditions

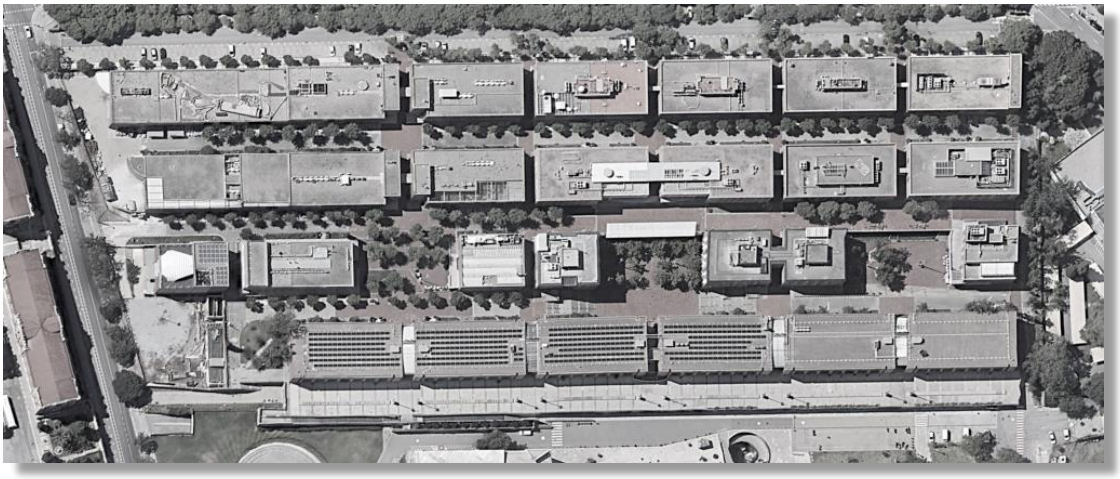


Figure 3-2 UPC University Campus

In this scenario, different types of relays (for example fixed or relay-UE) can be selectively activated or deactivated based on the total energy consumption of the RAN. Evaluations will be conducted to assess the impact of relay deployment on coverage (both indoor and outdoor) and spectral efficiency. Optimizing energy efficiency in the RAN involves considering multiple system parameters such as relay positions, operating frequencies, bit rate requirements, propagation, etc. Both the improvement in spectral efficiency and the increase in coverage that can be achieved by relaying will lead to an enhanced B5G energy-efficient RAN.

Initially, two different periods will be considered, that is one with high occupancy, when the campus is full on working days, and the other one with low occupancy, during weekends and nights when few people are on campus. Moreover, since vehicles are not allowed on the campus, users' mobility will be considered fixed (for users in class or in the office) or pedestrian (for users moving around the campus).

3.1.1.2 Densification of the radio access architecture

Distributed transmission is anticipated to become increasingly prevalent to establish more consistent quality and create non-intrusive, flexible, and robust networks [40]. Future networks will likely combine a range of RAN technologies from macro cells to small cells with very high-capacity short-range links. This calls for dense small cell deployments, especially for throughput-demanding use cases required simultaneously by high proportion of people in populated areas such as dense cities [41]. On the other hand, the technology evolution towards cost reduction and improved efficiency requires all future deployment scenarios rely on a superior transport network and network fabric that is flexible, scalable, and reliable to support demanding use cases and novel deployment options, such as a mixture of distributed RAN and centralised/cloud RAN enabled by artificial intelligence (AI)-powered programmability. The benefit of cell densification is to achieve certain area capacity using less complicated hardware at the expense of using more APs (adding to the infrastructure requirement) which then means more interference. Distributed MIMO (D-MIMO) is a technology that combines the best aspects of ultra-dense cellular networks with MIMO technology to enjoy the strength of both technologies [42]. At lower carrier frequencies (sub-6 GHz), where coherent transmission is possible, D-MIMO can be used to increase the system's spectral efficiency and, in principle, to avoid inter-cell interference. This is possible due to the deployment of service antennas spread out over a large area [41].

CF-mMIMO [43], [44] combines the elements from small cells [45], mMIMO [46], and UE-centric JT-CoMP [47]. In a CF context, the mMIMO regime is achieved by spreading many antenna elements across the network (even in the form of single-antenna APs [48]), which provides enhanced coverage and reduced pathloss. Moreover, a UE-centric coherent transmission extended to the whole network, where each UE is

served jointly by several BSs, allows to practically eliminate the interference, as shown in [49].

Such a large-scale D-MIMO system, that is CF-mMIMO, which can be thought of as the ultimate embodiment of concepts, such as network MIMO [49], multi-cell MIMO cooperative networks [50], virtual MIMO [51], ultra-dense networks and JT-CoMP, is now regarded as a potential physical-layer paradigm shift for 6G networks [52].

CF-mMIMO and D-MIMO communication are of great value in high-reliability scenarios. By distributing many nodes and bringing them closer to the users, path loss, coverage and shadowing are mitigated. At the same time, user mobility is no longer an issue as different UEs are connected to different nodes while maintaining connectivity and control. This technology is highly valuable for indoor and urban scenarios where UEs require very high reliability and low latency links, where forward error correction techniques such as hybrid automatic repeat request (H-ARQ) are not desired due to the latency generated.

The main challenge for large scale D-MIMO rollout is arguably the cost for installing many nodes in different places, each requiring fast and high speed fronthaul connections. D-MIMO installations should be easy to deploy, have a small and none-intrusive visual footprint, and are flexible to scale and extend.

As a first step, it would be necessary to understand the required distribution level, where is *the sweet spot* in terms of complexity versus robustness and performance considering the trade-off between distributed and co-located communications. In the context of BeGREEN's technical work, this shall include the power consumption of the RAN as well. Increasing the number of nodes translates to more equipment for RRH that consume energy. On the other hand, the total RF power consumption calculation requires to take a number of parameters into consideration, e.g. the total number of antennas, total number of PAs, etc. Then one needs to deal with the practical approaches to non-coherent operation in higher bands, and transport solutions satisfying the requirements. The optimum performance solution would be phase-coherent transmission with a single centralized processor, but it will be difficult to build and meet the feasibility requirements. The other extreme would be phase non-coherent phase transmission with duplicating every data in each AP and relying on single frequency network, but it will be inefficient. Further research for finding the balance, in terms of complexity, robustness, performance, and importantly, energy efficiency is necessary.

As we move towards higher frequencies, more available bandwidth comes into play and spectral efficiency is not necessarily the main concern anymore. Instead, reliability of the communication links becomes a priority. Reliability is impacted by the higher pathloss, lower available output power of semiconductors, narrower antennas beams and, most importantly, a higher level of signal blockage. On the other hand, the feasibility of practical implementation highly depends on the RF hardware capabilities, and other constraints, such as size, power source, and mobility. Moreover, the responses of different hardware components are influenced by the centre frequency, bandwidth, and waveform. Furthermore, the beamforming architecture and the possibility of exploiting spatial multiplexing depend on the radio channel characteristics, which needs extensive measurement and modelling.

3.1.1.3 General mMIMO incentives

There are several advantages in mMIMO. It can be used for enhancing the channel capacity, that is, improving spectral efficiency by maximizing the bits/s/Hz; it offers higher orders of Single User MIMO (SU-MIMO) by making it easier to reach higher ranks. For high traffic deployments, Multi User MIMO (MU-MIMO) will introduce more than three times capacity according to literature and demonstrations.

Another key usage of mMIMO is coverage enhancement. Specifically, it is a very good fit for the frequency ranges being used in 5G. C-band spectrum as an example, that is 3300-4200MHz, is the largest spectrum resource in FR1 allocated to 5G in most countries. As such, most 5G deployments (auctions and actual deployments) are seen to be focused on these bands. As frequencies are higher than most 3G/4G

deployments with two detrimental effects are introduced. Firstly, higher pathloss, equal to an additional (around) 5dB -compared to the traditional midbands, and then higher penetration loss as measurement campaigns show around 5dB-7dB additional loss for various types of buildings or materials. LTE usually operates in bands lower than 2.5 GHz, i.e. much lower than common 5G C-band with carrier frequencies up to 4 GHz.

A straightforward approach for solving the high pathloss problem is to deploy smaller cells and scale the cell with the frequency. However, the common industry approach suggests that the most cost-effective way for deployment of the C-band frequencies is to utilize the existing infrastructure, i.e. cell towers and backhaul links, and compensate for the additional losses through beamforming using a more antenna elements. In this way, the operators do not need to reduce cell size by adding new cell towers, and can place the new 5G RRHs on the same towers used for the legacy RATs. Reusing the same cell towers means also that there is potential for reusing the backhaul, being wireless or wired, assuming that it is able to manage the required throughput. This can result in capital expenditure (CAPEX) reduction.

A codebook-based MIMO with large number of elements as a baseline is supposed to improve spectral efficiency alongside the coverage enhancement. Non-codebook beamforming will increase spectral efficiency even further at the cost of baseband computing resources. The reason is that codebook-based MIMO is limited to a finite set of precoding matrices defined in the 3GPP standard, where non-codebook can use any beamforming coefficients. Also, DL codebook-based MIMO relies on feedback from the UE to a large number of BS DL antenna ports, which results in a high overhead and throughput reduction. In case reciprocity is feasible, as in time division duplex (TDD) systems where the DL and UL share the same frequency band, non-codebook beamforming can be accomplished based on the UL sounding reference signal (SRS), which have much less overhead as the number of UE antenna ports is significantly lower than number of BS antenna ports.

There are two main mMIMO approaches being used by the industry:

- Approach 1: Using only mMIMO
- Approach 2: Heterogeneous network - Deploying a mixture of mMIMO for higher traffic areas and smaller RRHs for low traffic areas

Many of the leading vendors use the second approach. For example, Nokia's C-band portfolio support a range of deployment scenarios: 64TRX or 32TRX mMIMO RRHs for extreme capacity, 8T8R macro RRHs for coverage solutions, 4T4R micro RRHs for street level and venue deployments and indoor pico RRHs for extensive in-building coverage. A similar approach is used by Samsung as described in Figure 3-3 [53].

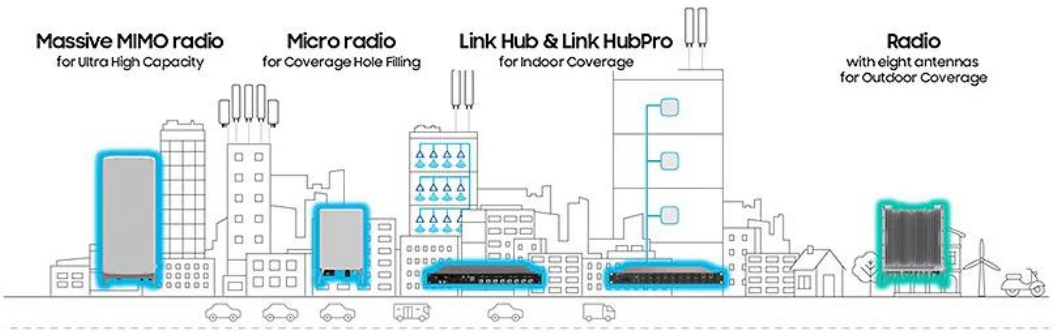


Figure 3-3 Samsung different types of RRHs [53]

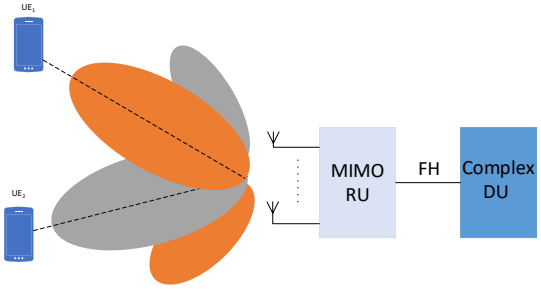
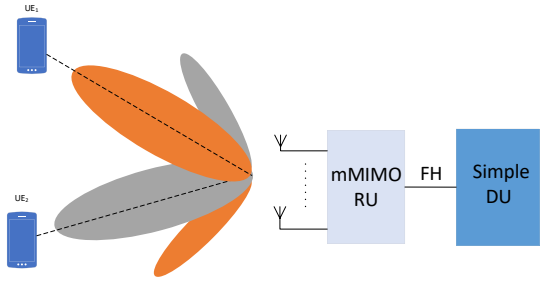
For the architectural exploration, it is recommended to define a heterogeneous network with different RRH types, that is include mMIMO, macrocell, microcell and picocells. This can offer the required flexibility in meeting power consumption requirements. Based on the spatial diversity introduced by the MIMO channel, the theoretical maximum number of layers that can be multiplexed over the same physical resources is

determined. However, the actual achievable number highly depends on the RRH type. We should analyze for each what is the maximal number of layers that can be used. Furthermore, even when mMIMO is available, it may not always be the most energy effective choice. For example, in areas where we do not expect many UEs sharing the same transmission time interval and we do not need to use MU-MIMO, the power consumption of the large antenna arrays do not provide a large return on investment.

On the other hand, different considerations are to be made for UL compared to DL. As in DL a higher level of SU-MIMO can be achieved compared to the UL. Hence, the power consumption, coverage, and spectral efficiency of several MIMO schemes for each component need to be compared. Usually a higher number of MIMO layers can be achieved in DL, as the number of transmit antennas in the BS is high. These layers can be divided between MU-MIMO and SU-MIMO. Depending on the number of served UEs per slot, and the DL throughput used for each UE, there will be cases where MU-MIMO is optimal, cases where SU-MIMO is optimal and cases where a combination of both is optimal. As for UL, SU-MIMO is limited by the relatively low number of Tx antennas. Hence, MU-MIMO will be deployed, and for each UE a maximum of two layers will be used. However, there can be cases where this number may be larger than two, especially since the number of antennas in UEs and customer premise equipments (CPE) is expected to increase in the future.

As for interworking with WP3, in WP3 different types of receivers and different types of processors to implement them on such as central processing unit (CPU) and graphical processing unit (GPU) are being evaluated. These evaluations can give insights to the overall mMIMO architecture being designed in WP2. The choice of which receiver to use can affect power consumption, coverage and spectral efficiency. For example, when using a sphere decoder (near ML receiver) in the BS we can handle higher-rank UL MIMO with a smaller number of antennas as in Table 3-1.

Table 3-1 Different RU/DU Types and their Implications

Regular MIMO RU	Massive MIMO RU
	
In this case a regular MIMO RU is used with a small number of antennas (as 4 or 8). On one hand the RU power consumption is relatively low as it is divided to a small number of ports. However, the DU will be required to implement a highly complicated near maximum likelihood receiver which will require a lot of processing and its power consume will be high. Also, with this configuration the spatial diversity of the channel will not necessarily will be fully utilized.	In this case a massive MIMO RU is used with many antennas (as 64). The RU will perform beamforming of the antenna array into a much lower number of virtual antennas to be sent to the DU over the FH. On one hand the RU power consumption will be high. However, the DU can use a reduced complexity linear receiver and its power consuption will be low. Also, this configuration has a potential for fully utilizing the spatial diversity of the channel.

3.1.1.4 Reconfigurable Intelligent Surfaces

RIS are widely seen as a critical technology for the next generation of mobile systems [54]. A RIS is a planar structure made up of reflective cells which can passively control the electromagnetic response of incoming RF signals by altering factors like phase, amplitude, or polarization. RISs represent a shift in paradigm by making the wireless channel an active element subject to optimization, rather than just a constraint around which to optimize. By doing so, they have the potential to increase the energy efficiency of mobile networks

by over 50% [55]. To meet this goal, RISs must meet several key requirements. Firstly, they should be able to steer RF signals with minimal power loss. Secondly, they must not utilize active RF components. Thirdly, they must minimize the energy required to reconfigure their reflective cells. Fourthly, they should be able to reconfigure themselves in real-time. Lastly, they should be capable of being produced at scale with low-cost methods.

The literature has several examples on implementing and using RIS using different technologies and configurations. [56] describes a 16x16 RIS that operates at 28 GHz. [57] introduces a 16x10 RIS that works at sub-6 GHz and uses an Arduino control unit, but only groups of elements can be configured. Both of these solutions use a 1-bit resolution phase shift with a RIS based on PIN diodes. In contrast, [58] achieves a 2-bit phase quantization by using 5 PIN diodes per RIS element, while [59] uses 3 PIN diodes to allow for 8 phase states. There are also RIS prototypes that use varactor diodes instead of PIN diodes. [60] presents a 14x14 RIS based on varactor diodes, which allows for continuous control of the phase shifts but requires a wide range in control voltages.

RF switch-based implementations are another option that can reduce costs compared to PIN diodes. [61] presents a RIS prototype with 14x16 reflectors at 60 GHz using RF switches. However, the unit elements are placed more than one wavelength away to reduce mutual coupling, which limits the maximum scanning angle. [62] describes a RIS with 40 reflectors mounted on boards that are one quarter of a wavelength tall and one tenth of a wavelength wide, with one tenth of a wavelength separation on both the x and y-axis. Finally, [63] allows for singular configurations of the elements, reducing the number of pins required at the controller side. This RIS is made of 4x4 patch antennas operating at 5 GHz and controlled with a 2-bit phase shifter using the transmission line method.

Overall, RISs have the potential to revolutionize the way we design and operate wireless communication networks. The most prominent application of RIS is coverage extension, where RIS can enhance signal quality and improve spectral efficiency of shadowed areas at a minimal cost as they are almost passive devices.Thus, a RIS can be configured to reflect radio waves towards a low coverage area so that we can improve the signal-to-noise ratio without deploying more BS. RIS offers a more flexible and cost-effective alternative to deploying more BS by allowing network operators to add them to strategically improve the coverage in well-known low coverage areas. Dense urban areas that have lots of shadowed spots that benfit from RIS deployments to increase coverage at a minimal cost. Furthermore, indoor environments such as smart homes, offices, or public spaces also benefits from RIS deployments as RIS can create adaptive and energy-efficient spaces by manipulating wireless signals. They can also improve indoor localization and tracking systems and enhance radar and sensing capabilities.

In certain scenarios where direct communication with a controller is challenging or impractical, a self-configuring RIS can prove invaluable. For instance, in very dense urban scenarios or harsh terrains where there are a lot of obstacles the best locations for setting up a RIS might not have direct sight with BS Self-configuring RIS units are equipped with power sensors that enable autonomous configuration and adaptation. A self-configuring RIS senses its wireless environment, assesses operational requirements, and adjusts its configuration accordingly. Through iterative adaptation and continuous monitoring, a self-configuring RIS units can optimize their configuration independently, without the need for direct communication with a centralized controller. This capability enhances flexibility, scalability, and adaptability in challenging environments, making RIS technology more practical and effective.

RIS can be integrated with ISAC data to enhance the overall performance of communication and sensing systems. The integration involves leveraging the capabilities of RIS to optimize wireless communication links and improve sensing capabilities simultaneously. One aspect of integration is utilizing RIS to enhance the quality of communication channels in ISAC systems (see Figure 3-4). By intelligently manipulating the reflection and refraction of electromagnetic waves, RIS can improve signal quality, increase coverage, and

mitigate interference. This leads to improved reliability and higher data rates in the communication link between sensors, devices, or networks involved in ISAC applications. Additionally, RIS can enhance the sensing capabilities of ISAC systems. By strategically modifying the propagation characteristics of electromagnetic waves, RIS can improve the accuracy, resolution, and range of sensing and detection systems. This can be particularly useful in scenarios where the quality of sensing signals is compromised due to obstacles, interference, or noise. RIS can optimize the sensing environment by redirecting or focusing signals to enhance target detection, imaging, and tracking. The integration of RIS with ISAC data involves a collaborative approach where the RIS, communication devices, and sensing systems work in conjunction to achieve optimal performance. The RIS dynamically adapts its properties based on the communication and sensing requirements, ensuring that both aspects are optimized simultaneously.

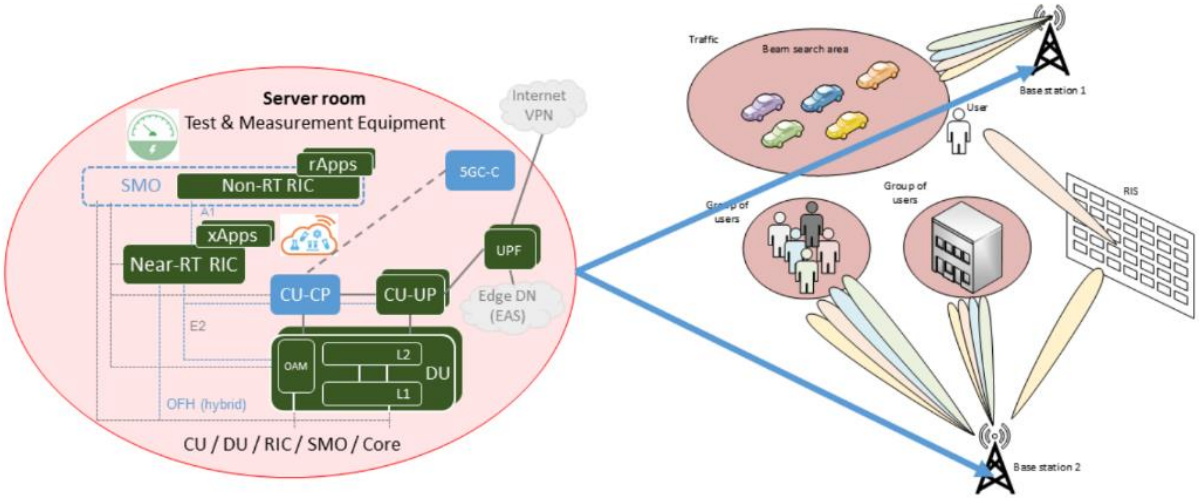


Figure 3-4 Sensing-assisted communications including ISAC and RIS

3.1.2 System-level reference use cases

In what follows we provide the reference use cases related to energy optimization in the system-level.

3.1.2.1 Energy efficiency in vRAN deployments with shared computing infrastructure

RAN virtualization is well-recognized as a key technology to increase cost-efficiency at the very edge of next-generation mobile systems [68]. The urge to increase the density of radio access points—yet preserve or even reduce costs—has attracted Network Function Virtualization (NFV) such as resource multiplexing by sharing infrastructure [71]. The idea of RAN pooling is not new: 71% of US operators indicated the intent to deploy RAN centralization by 2025 in a recent survey [72], for example NTT Docomo, Ericsson or AT&T are famously interested in this type of technologies [73][74][75]; and centralization is at the forefront of O-RAN [76]. However, the real-time impact of resource contention in shared RAN pooling platforms has not been studied sufficiently.

The success of Software-Defined Networking (SDN) and Network Function Virtualization in other domains has spurred the market to build virtual network functions (VNFs) such as firewalls, switches, virtual private networks (VPNs), etc., that provide carrier-grade performance. However, research has shown that resource contention caused by VNFs sharing common computing infrastructure may lead to up to 40% of performance degradation compared to dedicated platforms [77][78]. Although sharing resources the attention of industry in this direction; The O-RAN alliance advocates for this direction in its RAN specification architecture [69][70]. Virtualized RANs (vRAN) are expected to import the advantages of

through different VNFs tends to be more cost-efficient than using dedicated resources, the performance degradation leads to increased resource usage and the need to over-dimension the system, which renders

minimal benefit in sharing resources. The term *noisy neighbours problem* has been coined to refer to this issue, and has motivated substantial research over the years [78][79][80][81].

The virtualization of BSs (vBS) is not alien to this issue. vBSs' subframe processing must be completed as fast as possible, as they must comply with different tight time constraints, such a hybrid automatic repeat request (H-ARQ), which sets a 4 ms timeout. This makes vBSs very sensitive to low isolation as resource sharing leads to resource contentions, increasing the total processing time and energy consumption leading to user plane resets in the worst case. This fundamental tradeoff has been studied thourghoutly in the NFV literature. The more we share resources the more cost-effective we are but the less performance guarantees we have. Allocating and isolating the shared resources correctly from different deployed vBS is critical so that resource contention does not increase the total energy consumption and decrease the system overall energy efficiency .

3.1.2.2 Joint Orchestration of vRAN and Mobile Edge AI Services

There is a growing consensus that the next generation of mobile networks need to support AI and other intelligent services at the edge. These services typically require the collection, transfer, and processing of data flows in near-real time, with the aim to provide data operations, for example inferences, to end users such as small IoT devices, drones, or smartphones on the go. A representative example of these services are mobile video analytics (MVA), which are used in AR and VR services [83], cognitive assistance applications [82], surveillance systems [84], among other similar AI services. The core task of MVA is that user devices send video frames to the network, which needs to process them and transmit back accurately-detected depicted objects, or extract other important information [85].

While MVA-like services are already considered a utility that each user should be able to enjoy, their wide deployment requires a fundamental shift in the way we manage mobile networks. Namely, in these services, the network's role is not confined to transferring data from one point to another, nor even in processing the data. Instead, the network needs to directly optimize the service performance, which involves the criteria of accuracy (confident inferences), end-to-end latency (fast inferences), and task throughput (inferences/sec) in a resource-efficient fashion. This latter requirement is crucial since such services create voluminous data flows, involve heavy computations, and consume large amounts of energy [86]. In fact, energy consumption is not only one of the most prevalent operating expenditures for mobile networks [87] but has been also identified as the main blocker for the success of these inherently energy-demanding services. Besides, energy is the common resource consumed by all network operations (for example data transmissions, transfer, or computing) and its efficient management is imperative also when considering the performance point-of-view.

3.1.2.3 Traffic-aware management of NFV user-plane functions

Softwarization and virtualization of network functions has been one of the main technological pillars of 5G [89] and its relevance will continue to grow in B5G and 6G networks aiming for a full cloud-native architecture [90]. The implementation of network functions as VNFs enables key functionalities such as network slicing, control-user plane separation (CUPS), edge computing or disaggregated vRAN, enhancing the flexibility of the network to adapt to specific service requirements and dynamics. Examples of these functions are the 5GCore UPF and the CU-UP. While virtualization allows to run these user-plane functions on general purpose IT servers, this comes at a cost in terms of energy consumption due to an unefficient utilization of the CPU resources which are not adapted to the real traffic demands [91].

Modern CPU architectures allow to programmatically manage the pool of CPU resources by controlling CPU sleep modes/C-states/P-states [92][93]. C-states can achieve significant power by switching off certain parts of the CPU, while P-states allow changing the voltage and frequency of the CPU core. These methods permit to dynamically modify the frequency of the CPU or to apply micro-sleeps, which enables the intelligent

adaptation of CPU resources to the traffic demands. As described in [91], such adaptations can lead to significant power reductions in servers implementing the UPF, especially during idle and low load periods. While the actual contribution of the 5G Core to the total energy consumption of the mobile network is limited (13% of the total, according to [94]), its relevance is expected to increase in the near future. For instance, the adoption of edge computing and network slicing approaches will increase the number of UPFs in the mobile networks, which also will need to cope with intensive packet processing or with very low latency [91]. In the context of O-RAN, the intelligent control for managing the CPU resources assigned to the UPF could be implemented as an rApp joining RAN and 5GCore telemetry and control. In addition, the implementation of AI-driven algorithms will be a requirement to enhance decision-making, for instance by using traffic load predictors. A similar approach could be followed to enhance the energy efficiency of CU-UP instances in the O-Cloud [92][93].

3.1.2.4 RIC driven energy-efficient RU on/off control

The adoption of the RIC paradigm proposed by O-RAN as the de facto solution to control and manage RAN functions, has fueled the incorporation of data-driven analytics and AI/ML techniques to RAN procedures like Radio Resource Management (RRM) and Self-Organising Networks (SON) by means of the so called rApps and xApps [99]. Improving RAN energy saving is one of optimization targets that can be considered and, thus, it has been recently added to the use cases studied by O-RAN in [100] and further described in [93]. Due to its significant impact on the energy consumption of the mobile network [94], optimization efforts are usually focused on the intelligent control of the O-RU node through O1, E2 or the FH M-plane, with the objective of enhancing its energy efficiency according to traffic status and predictions.

One significant use case is the management of booster cells or carriers in areas which eventually require extra capacitiy. In such scenarios, RU energy consumption can be reduced during idle periods by switching off one or more carriers or cells of a given technology according to monitored, expected or predicted traffic load [93][100], leading to most of the RF and Digital Front End components to be set to sleep mode [92]. By implementing AI/ML mechanisms, rApps/xApps could provide traffic load or user mobility predictions to improve decision making and also provide a global view of the network to avoid local optimizations (for example switch off a certain cell) to negatively impact the overall energy consumption of the network (for example other cells become overloaded) [93][100].

These solutions will also require of adequate traffic steering or handover policies to avoid impairing service continuity and Quality of Service (QoS) [100]. Thus, rApps/xApps devoted to Energy Efficiency should interact with rApps and xApps managing these procedures (for example through the R1 or A1 interface). In addition, as described in [102], energy-aware and load-aware traffic steering and handover policies can provide indirect energy efficiency optimizations by efficiently balancing the load among available RAN resources.

3.2 Reference Scenarios

In this section we provide a description of the reference scenarios that we consider for the different use cases presented above. We follow the 3GPP TR 38.913 [88] document, which defines the different scenarios that will be considered in the BeGREEN project. Following we present which are the scenarios applicable to the different use cases:

Indoor:

An indoor hotspot scenario refers to a small indoor environment like a home or office. In this scenario, a small cell or microcell gives coverage to the indoor location boosting cellular signal strength and coverage. This would allow mobile devices within the hotspot to connect to the cellular network and access high-speed data services, such as streaming video, downloading large files, or using real-time video conferencing. In this scenario, mobile signals can be weakened by walls and structures, resulting in poor coverage and signal

strength indoors. Overcoming signal attenuation, ensuring reliable indoor coverage, and managing interference from multiple devices are key considerations in providing seamless connectivity within buildings.

Urban:

A urban scenario refers to a highly populated area, such as a city center, where there are numerous tall buildings and other structures that can cause signal blockage. This environment presents unique challenges for mobile communication due to the high concentration of people and the complex radio frequency environment. MNOs need to deploy a high density of BSs and antennas to ensure that there is sufficient coverage and capacity to serve the large number of people who require mobile connectivity.

Rural:

A rural scenario refers to an area with low population density and limited infrastructure. In such areas, there may be a scarcity of network coverage, and people may have to travel long distances to access a reliable network. Rural environments present unique challenges for mobile communication because the low population density and limited infrastructure can make it difficult and expensive to deploy mobile network infrastructure. MNOs may need to use innovative approaches like satellite or low-power wide-area networks to provide connectivity to rural areas. Mobile networks in rural settings need to extend coverage over larger distances, overcome signal attenuation due to distance and geographical barriers, and provide reliable connectivity to bridge the digital divide and enable access to essential service

Overall, these four mobile communication scenarios present unique challenges that require different approaches to ensure that mobile users have access to reliable connectivity. MNOs need to take into account the specific requirements of each environment when designing their networks to ensure that they provide the best possible service to their customers. The scenario for each use case is presented in Table 3-2.

Table 3-2 The Reference Scenarios and Use Cases

Use Case	Reference Scenario
Relay enchanced communication	Urban/Indoor
Cell Densification	Urban/Indoor
mMIMO	Urban/Rural
RIS	Urban/Indoor
vRANs with shared infraestructure	Indoor
AI Edge optimization	Indoor
UPF CPU management	Urban/Indoor
RU switch on/off control	Urban

3.3 General 5G Key Perfomance Indicators (KPIs)

This section overviews the standard network KPIs defined for 5G networks according to the document [96]. BeGREEN's future WP2 deliverables will use these KPIs' definitions and evaluation methods as input to define the project's KPIs. The KPIs will be defined in D2.2 with an initial evaluation of the performance results. The KPIs associated to BeGREEN use cases will be captured in deliverable D5.1.

In what follows, we provide a definition of some of the general KPIs that are considered in BeGREEN project.

3.3.1 Data Rate

The data-rate KPI measures the rate at which data (total bytes) is transferred over a mobile network. It quantifies the speed at which information can be sent (as UL) and received (as DL) by users, reflecting the

network's capacity to handle data traffic. The data rate is typically expressed in terms of bits per second (bps) or its multiples such as kilobits per second (Kbps), megabits per second (Mbps), or even gigabits per second (Gbps). It's important to note that the data rate KPI can vary depending on factors such as network congestion, signal strength, user location, device capabilities, and network configuration.

A higher data rate indicates a faster network that can deliver data-intensive services and applications more efficiently. It enables users to experience smooth multimedia streaming, quick file downloads, and responsive web browsing.

In this context, the peak data rate refers to the highest attainable data rate achievable under optimal conditions, measured in bits per second (bps). It represents the maximum number of data bits that can be received by a single mobile station, assuming a flawless transmission with no errors. This calculation takes into account the utilization of all available radio resources specifically assigned to the corresponding link direction.

3.3.2 Bandwidth

The bandwidth KPI in a mobile network measures the capacity or range of frequencies available for data transmission. It represents the amount of data that can be transmitted simultaneously over the network within a given time frame. Bandwidth is typically measured in Hertz (Hz) or its multiples such as kilohertz (kHz), Megahertz (MHz), or Gigahertz (GHz).

A higher bandwidth indicates a wider frequency range and thus a greater capacity for data transmission. This allows for the simultaneous transfer of larger amounts of data, enabling faster and more efficient communication between devices and applications. A wider bandwidth facilitates the support of bandwidth-intensive services such as high-definition video streaming, online gaming, and real-time multimedia communication.

3.3.3 Connection density

The connection density KPI in a mobile network measures the number of active connections or devices per unit area. It quantifies the volume of simultaneous connections within a given geographic region. Connection density is typically expressed as the number of active connections per square kilometer or square mile.

A higher connection density indicates that the network needs to support a large number of devices in close proximity. This is particularly important in densely populated areas, urban environments, or venues with high user concentrations such as stadiums or shopping malls.

3.3.4 User plane latency

The user plane latency KPI in a mobile network measures the time it takes for data to travel from a user's device to its intended destination and vice versa. It quantifies the delay experienced by user data during transport over the network. User-plane latency is typically measured in milliseconds (ms) and encompasses the time required for data to traverse various network elements and processes.

A lower user-plane latency indicates faster data transfer and responsiveness in the network. It is particularly crucial for real-time applications and services that require immediate interaction, such as online gaming, video conferencing, and augmented reality. A low latency network ensures minimal delay between user actions and the corresponding system responses, resulting in a seamless and immersive user experience.

3.3.5 Energy efficiency

The energy efficiency KPI in a mobile network measures the amount of energy consumed per unit of data transmitted or received. It quantifies the network's ability to optimize energy consumption while delivering

reliable connectivity and services to users. Energy efficiency is typically expressed as bits transported per unit of energy consumption (bits per Joules or bit/J).

A higher energy efficiency indicates that the network can deliver data and services while minimizing energy consumption. Energy-efficient networks are essential for reducing operational costs, minimizing carbon footprint, and prolonging the battery life of connected devices. By optimizing energy efficiency, network operators can contribute to environmental sustainability and achieve cost savings in energy consumption. In addition to measuring the amount of energy consumed per unit of data transmitted or received, energy efficiency in a mobile network can also be evaluated by considering other measures divided by energy. For example, we can measure the amount of energy per active user. These measures help assess the network's ability to maximize its performance while minimizing energy consumption.

4 System Architecture

General strategies for optimization have been mentioned in Section 2.1. This chapter will focus more precisely on mapping these general optimization strategies onto specific system architectures being proposed for study in the BeGREEN project. In each case, strategies will be required to be robust and scalable, which means in general that distributed heuristics are required. Another issue examined in this chapter is the placement of the optimization agent, which in O-RAN architectures will generally be in the RIC. Finally, BeGREEN's proposed reference architecture is presented.

4.1 Optimization Strategies

4.1.1 Optimization at the level of network topology

This section intends to explain the evolution in radio architecture which wireless cellular networks have undergone since the first release of RAN. The origin of the term "cell" with the first BS deployment led to the appearance of smaller and heterogeneous cells, leading to the emergence of micro and femto-BSs. In the same way, the transmission of parallel data streams led to the appearance of MIMO, which later evolved into mMIMO and D-MIMO. The next step in RAN, which will mean erasing the cell boundaries, is the implementation of CF mMIMO communications, which will be studied in BeGREEN T2.2. Figure 4-1 shows the evolution of these architectures. This subsection will compare the different radio architectures including current state of the art analysis.

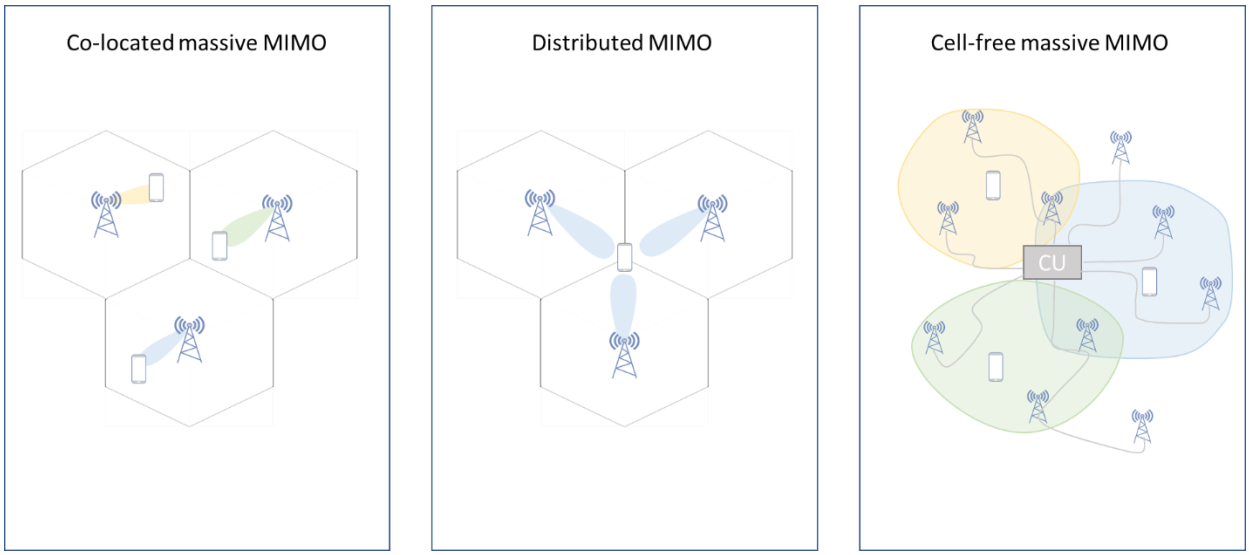


Figure 4-1. RAN Architectures

Traditional MIMO technologies started to be standardized in 3GPP Rel-8 [112] and are widely deployed in 4G and 5G. It has enabled spatial diversity, beamforming, and multiplexing capabilities, bringing important improvements in terms of reliability, SNR, and spectral efficiency. In recent years, the trend has been to constantly increase the number of transceivers in the BS, with the latest 5G deployments using 64T64R active antennas to take advantage of mMIMO, where the number of transmitting elements is much greater than the number of devices and layers.

The mMIMO architecture considers a significantly large number of antennas, in the order of tens or hundreds, providing a higher number of degrees of freedom that allows beamforming with very precise beams, improving signal quality, coverage and capacity extension in desired directions. Another advantage of mMIMO systems is interference management, which mitigates interferences through spatial filtering and cancellation techniques. Also, the increased number of antennas provides a high diversity gain to mitigate

fading and improve link reliability, while obtaining higher multiplexing gains due to the large degrees of freedom.

Such mMIMO systems tipically operate in TDD mode due to the complexity of accurate channel estimation. TDD operation exploits the DL-UL reciprocity, which makes it possible to estimate the channel in the UL direction, as the number of users is usually much smaller than the number of BS antenna transceivers. On the other hand, mMIMO systems require less complex signal processing due to the channel-hardening phenomenon that occurs due to the large number of antenna elements, where the channel between the BS and the users becomes less variable and more deterministic, thus simplifying the required adaptive algorithms for time-varying channel estimation and equalisation [113].

In order to find the best coverage and signal quality possible, different densification levels can be studied forming a heterogeneous network. This means that different RRH units will be used depending on the deployment area. For example, when not many users are served in a large and non-blocking area, a macro RRH can be used, while in crowded environment, or in a poor propagation channel environment as indoors, smaller RRH as femto-RRH would be suitable. As defined in Section 2.3.1, there is a wide range of RRH depending on the use case, some examples are to use macro-RRH with 64TRX or 32TRX mMIMO RRHs for extreme capacity, 8T8R macro RRHs for coverage solutions, 4T4R micro RRHs for street level and venue deployments and indoor pico RRHs for extensive in-building coverage.

The combination of network densification with MIMO technology provides great advancements in the communication data rate and user experience when dimensioned properly; however, cellular networks may lead to poor signal quality in the cell edges due to inter-cell interference (ICI) worsening the communication performance at the border between stations.

In order to avoid ICI in cell edges, Coordinated MultiPoint (CoMP) Transmission/Reception was introduced. This allowed the coordination of the transmission and reception of signals between different BS from a user device, improving the signal quality while reducing the interference. To implement CoMP, coordination and synchronization between the BS are needed, requiring some computational complexity [114].

As an evolution of CoMP, D-MIMO was introduced. In D-MIMO different RRH can transmit/receive different data streams to the same user. This way, the SINR is highly improved at the cell-edges, while MIMO capabilities are improved by means of the large separation between RRHs, implying a strong decorrelation between the channel components, thus providing a high channel capacity, spectral efficiency, and link reliability. The grouping of RRHs into clusters that connect to a user, can also lead into cluster interference, similarly to the ICI that was experienced from a regular cellular deployment. This cluster interference can be expected specially in high cell density environments.

CF-mMIMO, as an evolution of D-MIMO, is one of the promising solutions that is being developed during recent years. CF-mMIMO consists of spreading antennas over the deployment scenarios, erasing the cell boundaries obtaining several advantages compared to the traditional mMIMO, CoMP and D-MIMO deployments. Cell-edges disappear, as the network have the capability to dynamic reconfigure and select which RRHs (also named AP in the CF community) are serving a set of users considering the channel state information (namely AP-clustering). As there are no cell-boundaries and many APs are radiating the same signal, the SINR is stabilized over the whole network, having the different user-terminals always connected to the most-suitable APs, while the network decreases interference levels of the different users combining their signals appropriately. The CF solution needs to have more APs than users, therefore increasing network densification, making the APs closer to the users, shortening their wireless distances, and therefore consuming less radiated energy.

On the other hand, CF-mMIMO requires a higher complexity in terms of radio resource management, as it is required to dynamically compute and select the best combination of APs for different users or cluster of

users considering time-frequency resources and channel capacity that is achievable under different coherent transmissions techniques. Also, CF-deployments are expected to be more complex as the network elements are spread over the scenario, requiring backhaul, power supply, and installation permits for each of the antenna elements, making it a preferred option for indoor scenarios where it is easier to deal with such challenges. In the case of outdoor scenarios, it is reasonable to start thinking on a hybrid approach using the existing grid of macro BSs and strengthening the network with distributed deployments in the areas of expected higher requirements.

Even though in Section 3.1.1.3 it is said that current studies suggests that the most cost-effective way for deployment of the c-band frequencies is to utilize the existing infrastructure instead of installing new cellular towers, we need to study the previously defined network architectures to see how they behave in terms of energy efficiency. As the nature of the network may vary significantly with densification, the network energy efficiency concern becomes much more relevant. It will be necessary to consider not only the increment on the CAPEX side but also the OPEX savings due to the network efficiency using cell-free approach.

The energy efficiency of a small-cell network has been studied in [115]. The authors prove that increasing the BS density will improve the energy efficiency as long as the circuit power consumption remains below a threshold.

When increasing cell densification, inter-cell interference might be problematic. In order to deal with this, CoMP techniques can improve the signal quality and channel capacity. In [116] it is shown how DL is improved by joint processing scheme, where different BSs are used to transmit over the same Resource Block, or coordinated beamforming which reduces interferences and thus increases SINR. These techniques can provide about 50% of improvement in terms of throughput. On the UL side, different reception techniques can be used to exploit the spatial diversity of using multiple BSs, implementing minimum mean square error, zero forcing or maximum ratio combining joint combining techniques, achieving up to 54% of improvement in terms of capacity.

As previously described, D-MIMO was developed as an evolution of CoMP. This was studied in [117], where a comparative study of MIMO Cellular networks as co-located versus distributed is performed. Even though D-MIMO performs better in average, it is very sensitive to the user's position at the cell edge, needing a proper transmit power allocation mechanism to obtain a uniform SINR over the network. In this case, the deployment cost of distributed BS antennas also becomes significantly higher than in co-located MIMO, requiring more backhauling.

Comparing mMIMO with small-cell deployments, [118] and [119] show that small-cell systems are more energy-efficient than co-located mMIMO with the growth of antenna density. Reducing the cell size (or increasing the BS density) is a way to increase the energy efficiency. However, when the keep alive-circuit power of the BS dominates over the transmited power, this benefit ceases to be beneficial [120]. When antenna density is too high, the average energy efficiency of each active antenna in the small cell scenario is lower than in mMIMO, as the circuit idle power of each mMIMO antenna is usually lower than each of the antennas in small cell.

On the other hand, CF-mMIMO systems outperform small-cell systems when studied under the same conditions [121]. It is shown how CF-mMIMO systems obtain higher throughputs while these are more robust against shadowing effects, therefore making the network more energy efficient. On the negative side, CF-mMIMO requires more backhauling, making the network more complex.

CF-mMIMO is expected to outperform typical cellular mMIMO deployments in several ways [122]. First, achieving higher and more stable SINR within the coverage area: when the user is located close to the macro mMIMO antenna reaching good channel quality is affordable, but as the distance increases, channel quality drops. The CF-mMIMO deployment will provide better SINR values as the simultaneos connection to many

nodes providing supperior channel quality in every environment and setup. Then, by means of better interference management capabilities, mMIMO systems can easily manage interference separating spatially different users, however SINR will have a larger range due to the different distance locations among the coverage area. On the other hand, CF-mMIMO systems can mitigate the interferences providing a more homogeneous SINR within the coverage area. Finally, improving SNR using coherent transmissions, where all the antennas that are physically distributed over the coverage area outperforms macro mMIMO deployment in terms of SNR when both transmits using a coherent scheme.

Current O-RAN architecture supports the deployment of both co-located and distributed mMIMO solutions. However, it does not support the deployment of CF-mMIMO communications. As has been analysed in [17], CF-mMIMO could be implemented using O-RAN network elements with the interfacing of the different O-DUs with the O-CU, or implementing an inter-O-DU interface enabling a higher efficiency. The near-RT RIC should be in charge of defining the most suitable clusters of O-RUs serving each user.

Within the scope of BeGREEN, the project will examine the heterogeneous nature of the RAN based on the described reference scenarios. The main focus of BeGREEN will be to compare cellular environments with mMIMO technology against those utilizing CF-mMIMO deployments. Determining the optimal deployment strategy and configuration (for example number of radios, node separation, etc.) for each solution presents challenges. Addressing these challenges and finding the optimal balance will be a significant aspect of BeGREEN's objectives, and the outcomes will be documented in BeGREEN D2.2.

4.1.2 Coverage extension by means of relays

Communications using mmWave frequencies, where there is more bandwidth available, will allow to reach the high data-rates demanded in 5G-and-beyond systems. However, compared to the currently used sub-6 GHz frequencies, mmWave frequencies experience higher pathloss values, therefore smaller coverages are obtained and higher sensitivity to blocking by obstacles is expected.

Relays are being proposed as solutions to enhance the performance in mitigating obstructions by objects in outdoor mmWave deployments, augmenting capacity in high-density areas, providing coverage extension in outdoor and indoor areas, or improving resilience. The goal is to convert the communication link between BS and UEs into two (or more) links with better propagation conditions. This can transform a NLOS link into several LOS links to overcome an obstacle, or to extend the communication link to an UE which does not receive direct coverage from the BS.

As presented in Section 3.1.1.1, three types of relays have been identified: fixed relays, moving relays, and UE relays. Below we introduce the different technologies that exist for relaying, which are IAB, mobile IAB, and Proximity-based Services (ProSe).

Although 3GPP LTE in Rel-10 considers wireless backhaul with the so called “LTE relaying” [123], the proposal did not attract the interest of mobile operators. Mainly due to the limited available spectrum for 4G systems, and that LTE relaying only supports a single hop, with a static architecture from parent to child, and an inflexible partition of bandwidth between access and backhaul.

In the context of 5G systems, a wireless backhaul solution under the term IAB has been standardized. IAB is a technology which provides flexible wireless backhauling by means of relaying using 3GPP new radio (NR) technology. NR IAB has been proposed as technique for NG-RAN in 3GPP Rel-16 and continued in Rel-17 [124]. It consists of using NR radio-access technology for both the access link between BSs and UEs, and the backhauls links to transport information to the network. IAB is an interesting solution to enable the network densification required by 5G and beyond since it provides a fast and flexible deployment of new nodes [78][125][126].

The relaying node is referred to as IAB-node, and the terminating node of NR backhauling, that is connected to the rest of the network in a conventional way (for example fiber or microwave) is referred to as the IAB-

donor.

The IAB-donor serves relay nodes (IAB-nodes) and other UEs which are directly connected to it. Basically it is a BS with an additional adaptation layer on top of RLC, known as Backhaul Adaptation Protocol (BAP), which manages the routing via the IAB. Backhauling can occur via a single or via multiple hops as shown in the IAB architecture depicted in Figure 4-2.

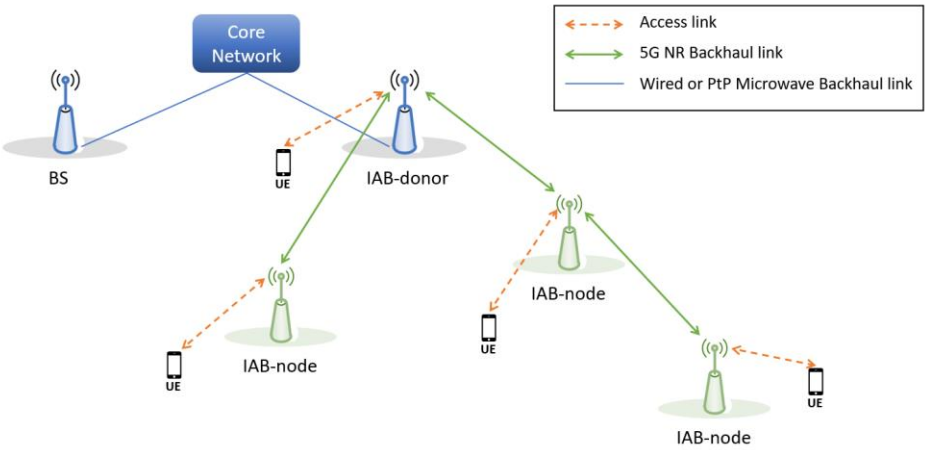


Figure 4-2. IAB deployed scenario

The IAB-donor is split in one CU and one or more DUs as a BS, while the IAB-nodes have a mobile termination (MT) part, referred to as IAB-MT, and a DU part, referred to as IAB-DU. The MT part is used to connect to a parent DU (which could be the donor DU or the DU part of another IAB-node), while the DU part of an IAB-node is used to serve UEs or the MT part of child IAB-nodes. Figure 4-3 illustrates the topology of a IAB network. DU functionality, as defined in TS 38.401 [127], terminates the 5G NR access interface (NR Uu) to UEs and next-hop IAB-nodes, and terminates the F1 interface with the BS-CU at the IAB-donor, referred as IAB-CU.

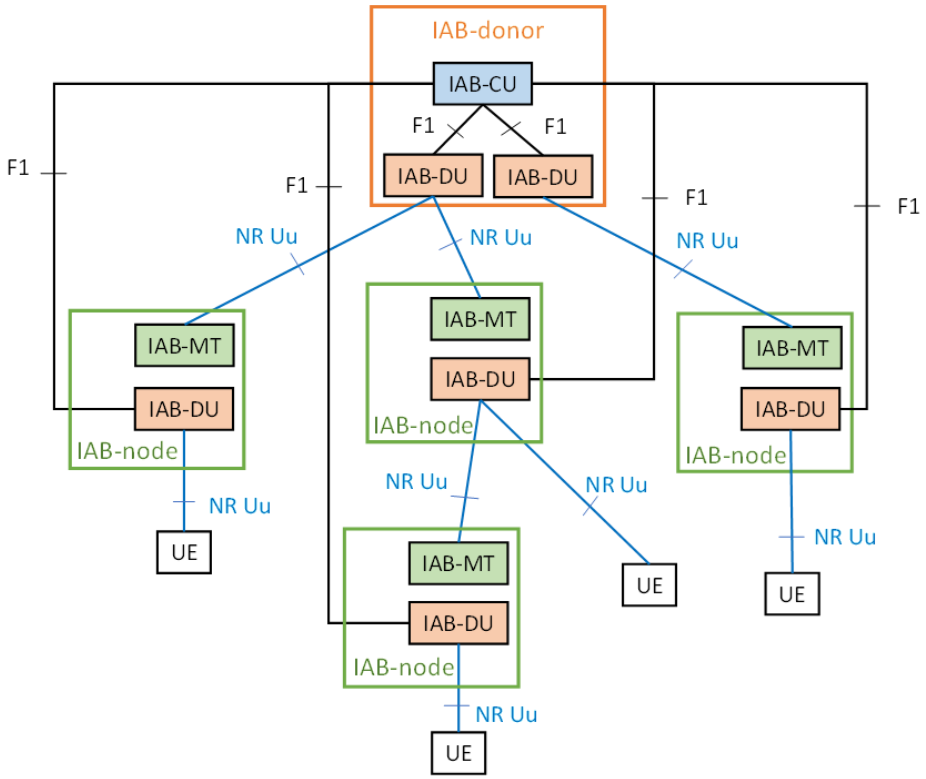


Figure 4-3. Topology of an IAB network

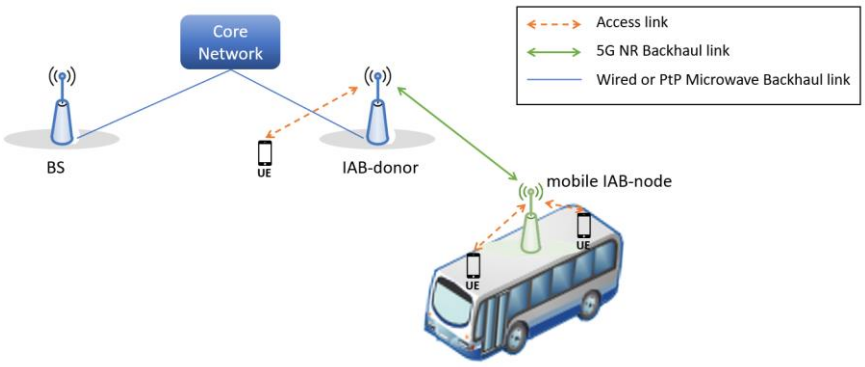


Figure 4-4. Mobile IAB deployed scenario

3GPP specifications for 5G NR IAB in Rel-16 and 17 only consider fixed IAB-nodes, while the interest in vehicle-mounted relays has become more relevant [31]. Several studies show that by deploying onboard mobile relays at the vehicles (for example trains or buses), in-vehicle access for the in-vehicle UEs through wireless backhaul links can be efficiently provided [58].

So, as part of future Rel-18, 3GPP started to identify potential architecture and system level enhancements for the 5G system to support the operation of BS relays mounted on vehicles, using NR for wireless access toward the UE and for wireless access through an IAB-donor toward the 5GC [35]. This work item has been called as mobile IAB for NR [34].

In mobile IAB, the mobile IAB-node should have no descendent IAB-nodes, that is, it serves only UEs, as shown in Figure 4-4.

The relay UEs are based on the direct communication between terminals. In this context, ProSe are services that can be provided by the LTE and 5G 3GPP systems, based on UEs being in proximity to each other. They were initially introduced in LTE to address use cases related to public safety, and later on were extended to vehicle-to-vehicle communications.

In 3GPP Rel-17, the enhancement of ProSe for a wider range of applications with the support of UE-to-Network (U2N) relay was considered and improvements of this feature have been included in 3GPP Rel-18, as defined in the specifications of [129]. Specific procedures and identifiers were defined, respectively, for 5G ProSe Direct Discovery, 5G ProSe Direct Communication, 5G ProSe UE-to-Network Relay and 5G ProSe UE-to-UE Relay [129].

5G ProSe U2N Relay enables indirect communication between the 5G network and UEs (for example for UEs that are out of coverage of the network as shown in Figure 4-5). A new PC5 interface together with a radio link for direct transmissions between UEs, denoted as Sidelink, was defined in LTE and 5G NR. Only unicast traffic (UL and DL) between the Remote UE and the network shall be relayed by the ProSe U2N Relay.

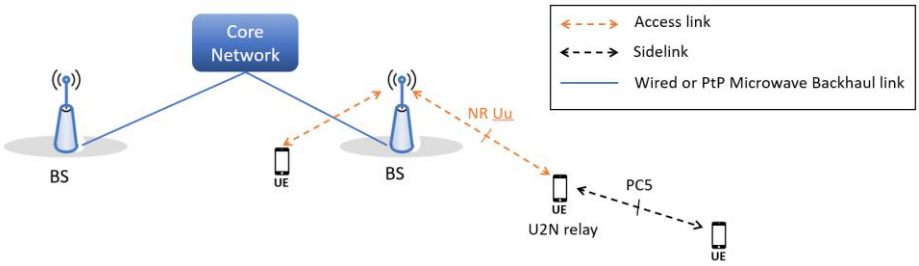


Figure 4-5. ProSe UE-to-Network Relay scenario

A RAN enhanced with relay nodes can reduce energy consumption while maintaining performance, and thus provide a more energy efficient network. More specifically, the deployment of fixed or moving relays in a RAN can alter the amount of energy consumed by the different entities involved. On the one hand, the BS consumes less power since the users connected through the relay experience better propagation conditions and thus they require less transmitted power from the BS than if they were directly connected to it. On the other hand, the relays themselves (fixed or mobile) consume power when they are active. Therefore, the overall operation of BSs and relays needs to be properly coordinated to achieve an overall energy consumption reduction. At the same time, users who connect to the relay need less transmitted power than if they were connected to the BS, and therefore their battery will last longer.

The assessment of the energy savings that can be achieved by means of relaying in a wireless system has been studied in previous works [27], [31], and [130]. In particular, the role played by scenario and channel conditions influencing the performance, such as the propagation and shadowing effects, the energy consumption model parameters, and the influence of the bit-rate and spectral efficiency, is highlighted in [31].

When looking to improve the energy efficiency in relay-enhanced RAN, there are several scenarios to consider. One example is to switch off a BS that has very little traffic, if the UEs that are within its coverage area can be served by a relay connected to a neighbouring BS, as shown in Figure 4-6. The relaying to UEs located at a large distance from the BS can also improve energy efficiency, as analysed in [31] and [131]. In this case, as illustrated in Figure 4-7, shorter UE-relay and relay-BS links mean less power requirement compared to the direct UE-BS link case.

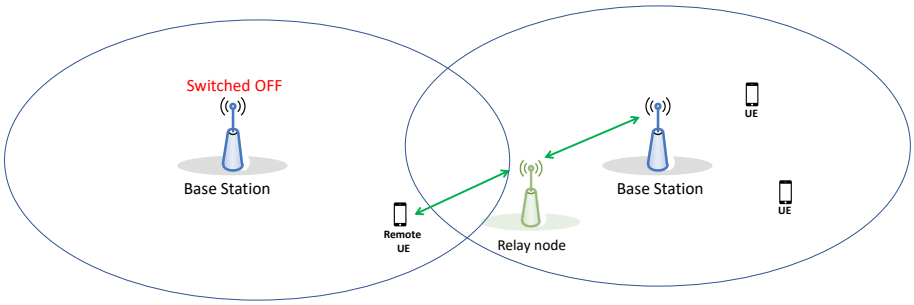


Figure 4-6. Relay serving a user from a diferent coverage area and allowing a BS with little traffic to be switched off

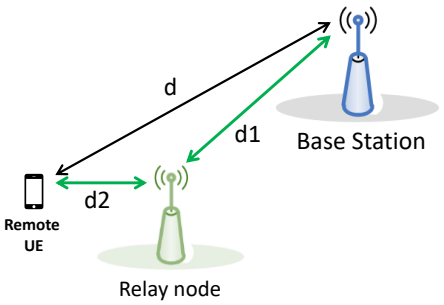


Figure 4-7. Relay serving UE located at large distance from the BS

Another situation where relays can improve energy efficiency is for groups of UEs moving together, for example in a train, bus, etc. In such a case a moving relay deployed in the vehicle that transports them allows the UEs in the vehicle to connect to the relay. Thus the BS has to serve only the relay instead of each of the UEs individually, so the better propagation conditions in the link BS-relay than in the BS-UE links will lead to less transmitted power at the BS. In [132] authors show that the efficient use of relay nodes to serve vehicular

users can greatly improve the energy efficiency of the network while maintaining the required quality-of-service (QoS).

In the context of a 5G RAN architecture using fixed relays, the authors in [133] proposed a coordinated mechanism to minimize network power consumption by utilizing the information of QoS in the whole network and determine which relay or UE should enter into sleep mode, using discontinuous transmission and reception when there is no data to be scheduled. By simulation using realistic network parameters, the presented mechanism is able to achieve up to 50% network power saving in comparison with LTE. The problem of finding optimal transmission powers for fixed relays using the same frequency channel, was addressed in [134]. In that work different communication schemes were considered: full-duplex, half-duplex, and hybrid relay schemes, and simulations indicate that a full-duplex access node is best suited for relatively small cells.

The power consumption of the RAN in different scenarios of device-to-device (D2D) relay systems, is analyzed in [135] where power consumption models for LTE BSs (macro, micro, pico, …) and for LTE user devices when transmitting and receiving data are proposed. Similarly, in [136] a comparison of the energy consumption when a UE communicates directly with the eNB in a LTE network, and when it communicates via an UE acting as a relay (Relay UE), shows that a reduction of up to 45% of the total energy consumption of the network and up to 40% of the user device can be achieved with a slightly increase in the relay device consumption.

In [137] a framework to evaluate power consumption and outage probability in UE-to-network relays is proposed, including a mode selection scheme which allows a remote UE to choose whether to connect directly to the BS, or through the relay UE. Similarly, in [138] they demonstrate that UE-to-Network relaying can be used to increase the energy efficiency of the network using a stable and proportional fair user association algorithm. Authors in [139] propose an algorithm for relay selection and power allocation, to minimize the total energy consumption of all remote and active Relay UEs, while guaranteeing the minimum data rate to each. Further details on specific algorithmic solutions related with the use of relays are presented in Section 6.1.4.

Aspects such as frequency planning, relay placement, power and resource allocation, and user association/relay selection, are key challenges for the optimization of energy efficiency in B5G networks that use relays. Fixed and moving relays need proper planning for indoor and outdoor deployment to reduce interference. Resource allocation between access and backhaul links will be dependent on whether the two links use the same frequency carrier (inband relay) or different frequencies/bands (outband relay), and this will have implications on achieved throughput and interference conditions.

4.1.3 Coverage extension by means of RIS

Another method to provide coverage extension is the use of RIS (Reflective Intelligent Surface). RIS is a hardware-efficient and highly scalable method to achieve dynamic changes in the wireless communication propagation environment in 5G and B5G systems [130]. RISs are man-made surfaces and consist of hundreds or thousands of nearly passive (reconfigurable) elements that can perform various functions, such as signal relaying/blocking, position estimation, obstacle detection, narrow beamforming, and multipath shaping. They enable to efficiently control radio wave propagation passively without needing active power amplifiers, focusing the impinging signals in specific areas to boost performance, mitigate obstructions, or extend radio coverage in dead zones [141]. There have been defined two approaches to control a RIS, in-band and out-of-band control.

- In-band control involves dynamically configuring the RIS using the wireless signals that are intended to affect, while ensuring that the control channel remains operational. This control can be done explicitly, where the RIS contains a full radio interface and can exchange control signals related to

the RIS, for example when multiple service providers in the same area need to make contractual requests to the RIS. On the other hand, in implicit in-band control, the RIS relies on sensing and interpreting received wireless signals.

- Out-of-band control involves controlling the RIS through a communication interface that is not affected by the RIS reconfiguration.

Various hybrid and composite control schemes can be designed based on these two main ways to control a RIS. The RIS resources that can be controlled via a control interface and allocated to stakeholders including space (physical subsections of the RIS), time (time intervals), frequency, or a combination of these.

One of the main challenges in controlling a deployed RIS serving different stakeholders, is its integration into an open and flexible network architecture. Recently, major players in the network business have made significant efforts towards the concept of O-RAN. Network operators may now acquire general-purpose hardware to customize their RAN following the O-RAN paradigm, as per the official O-RAN architecture [142]. The deployment of RIS may further increase interoperable smart radio environments, enabling drastic openness changes in the networking industry. In order to exert control on the RIS in an open, multi-vendor environment, network control components related with RIS may be integrated within a NFV/MANO system and plugged into an O-RAN platform. This is exactly what the RISE6G project (https://rise-6g.eu/) has been researching. Originated from [141] by authors in the RISE6G consortium, a new proposal about the network domain between the RAN elements and final UEs, where the RISs can be intelligently controlled by a RIS Controller to drive overall network efficiency at the edge, is presented. The RIS controller would interface with the orchestration layer (and the corresponding RIS plug-ins available in the SMO) through a new interface "RX" to instruct transmitter configurations for specific environments or use cases. The RIS Controller may also adjust the RIS parameters at short time scales based on feedback and desired KPIs by comunicating with the near-RT RIC. RISE6G proposes to use a new interface "R2" for such interaction, enabling near-RT RIC assisted joint RIS and BS beamforming, near-RT configuration of RIS parameters and RIS-based JCAS procedures.

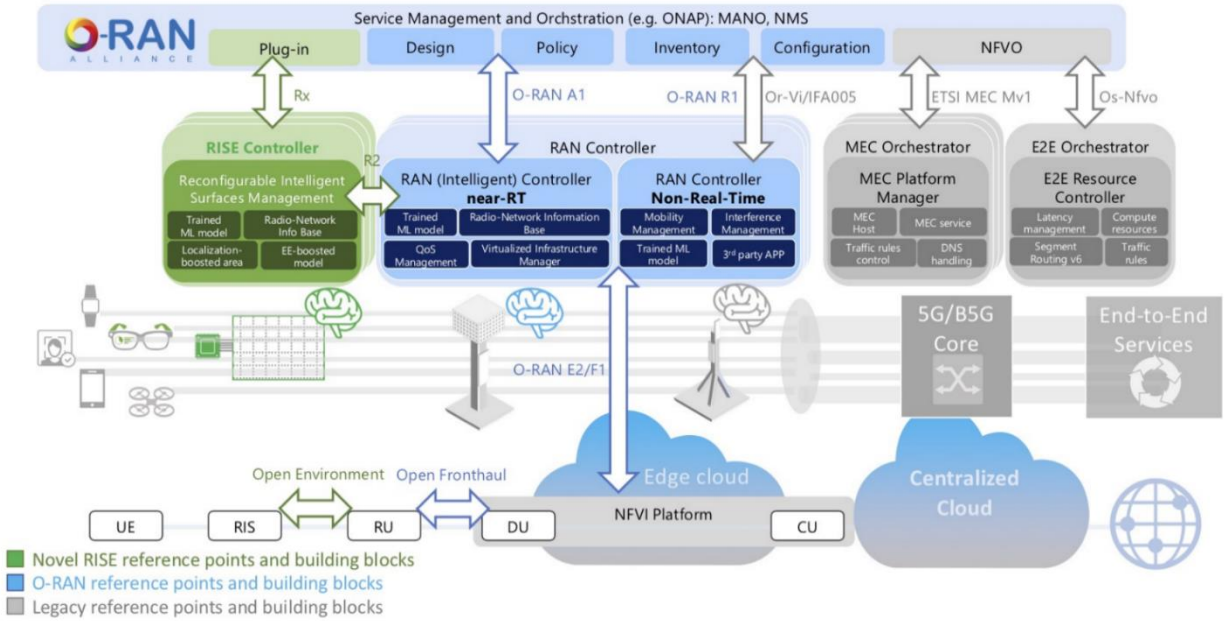


Figure 4-8 Proposed integration of a RIS-enabled RAN within the O-RAN architecture proposed by the RISE6G project

The use of a controller similar to real-time and non-real-time RICs that can operate with RIS allows AI/ML techniques to optimise the configuration parameters of the surfaces, similar to how RAN RICs optimise O-RAN BS components. There have already been various approaches for combining AI optimisation techniques

with RIS. However, they have yet to be integrated into O-RAN. Different deep neural network (DNN) architectures are utilized to design RIS beamformers, as described in references [143]-[145]. . [143] and [144] focus on maximizing the achievable rate of the system. In both papers, the authors use channel information to optimize the reflective properties of the RIS deployed. In [145], in order to maximize the received signal strength of a user, the authors develop a Deep Neural Network (DNN) to learn the mapping between the measured position information at the user's location and the optimal configuration of the RIS unit cells. Finally, in [145], authors propose adopting an unsupervised learning mechanism for passive beamforming design.

Besides the RIS control, an additional challenge is the design and deployment of a RIS-enabled 5G network. Adequate deployment strategies to extend the coverage and avoid dead zones existing in existing networks are crucial to improve efficiency and reduce costs. However, solving the RISs placement problem involves an increased overall deployment complexity, which calls for advanced optimization techniques to strike the optimal trade-off between RISs density and the corresponding spurious detrimental interference. Some works that have studied this trade-off are [222] for a mmWave access network and [111] in an empirical validation in a railway station. In general, studies of this kind of topology optimization have been done without regard to energy efficiency. BeGREEN will consider adding energy efficiency into the objective functions used in such optimizations.

4.1.4 Performance and spectrum efficiency using ISAC

ISAC will play a key role in delivering intelligent connectivity in future mobile communications. Unlike in the past [225], where sensing and communication operated independently, these functions will coexist in wireless communication systems. They will leverage and share the same set of resources including time, space and frequency. Additionally, enablers such as waveforms, signal processing capabilities, and hardware implementation targets [226][227][228] will facilitate this coexistence. This integration of sensing and communication gives birth to the concept of sensing-assisted communications [229] and communication-assisted sensing [230]-[238].

Within the ISAC framework, radio waves are used to sense the surroundings and to obtain information about the physical environment, which enable the provision of new services range. Sensing information opens the door to different applications such as improving the beamforming performance, assisting in the recovery from beam alignment failures and facilitating the tracking of the channel state information (CSI) between TX and RX. As a result, communication throughput and reliability are enhanced. The increased use of higher frequency bands, such as mmWave or THz, wider bandwidths, and denser distribution of massive antenna arrays will allow for additional opportunities to mutually enhance both functions [239]-[244].

From an architectural point of view, there exist four standard sensing modes:

- **Monostatic sensing**: A TX and RX are co-located at the same device and share a common clock and knowledge about the transmitted signal.
- **Bistatic sensing**: A TX and RX are located on separate devices and they may or may not share a common clock and full knowledge of the transmitted signal. Therefore, the positioning of a UE relies on bistatic sensing to and from multiple BSs.
- **Multistatic sensing**: A system comprising at least 2 transmitters (and 1 receiver) and/or at least 2 receivers (and 1 transmitter) separated in space, without a common clock.
- **Passive sensing**: The transmitted signal is provided by an external system (for example radio broadcast tower), while there is a sensing receiver, which has limited knowledge regarding the transmitted signal (for example only carrier frequency and bandwidth).

From a functional point of view, there exist two modes that may be combined with any of the above-mentioned architectural modes:

- **Radar-like sensing**, where the radio signal is processed to extract distances, angles, or Doppler shifts, to detect the presence and state (position, velocity) of objects/targets and track them over time. Radar-like sensing thus starts with detection/channel parameter estimation, followed by data association and by tracking. When objects are static, the process is called mapping, whereas when objects are moving, the process is called tracking.
- **Non-radar-like sensing**, as any other type of sensing, including pollution monitoring, weather monitoring, detection and tracking based directly on the received waveform or features extracted from the received waveform. These features can be applied to ML for classification or regression.

3GPP standardization has identified possible use cases for ISAC [110] that allow detecting objects that are not "connected" to the network. It is expected that the use cases will be motivated by the desired deployment scenarios and that imposes constraints in the type of technologies and their carrier frequency operation. There may be micro cellular deployments that require relying on FR2, non-3GPP carrier frequencies such as 60GHz, and sub-THz frequencies. These may be key to sense the environment in, for example industrial environments. Macro-cellular deployments would require Sub-6 bands only as they may serve the purpose of sensing in outdoor scenarios.

The ISAC systems are designed to fulfill four main types of use cases. Firstly, they enable high-accuracy localization and tracking, allowing precise identification and monitoring of objects or UEs. Secondly, they facilitate simultaneous imaging, mapping, and localization, enabling the creation of detailed maps while accurately determining the location of the system. Thirdly, they enhance human senses through augmented reality, providing users with additional information or sensory input. Lastly, the ISAC systems are capable of recognizing and interpreting gestures and activities, allowing for intuitive interaction and understanding of human actions.

4.1.5 Optimization at the level of the RAN intelligent controllers

As introduced in Section 2, the project assumes O-RAN standard as the baseline architecture since it provides inherent mechanisms to realise many of the infrastructure changes and techniques for energy optimization required for the Use Cases described in Chapter 3. In the O-RAN architecture, RAN intelligence is enabled by means of data-driven non-RT and Near-RT control-loops, which are managed by the non-RT RIC and Near-RT RIC components and implemented by the rApps and xApps, respectively [99]. The RIC and xApps/rApps form a vital framework in ORAN, driving the evolution of wireless networks. The RIC serves as a central control entity, while xApps/rApps utilize its capabilities to offer specialized functionalities. This symbiotic relationship allows for greater flexibility, scalability, and efficiency in managing radio resources. By deploying xApps/rApps on the RIC platform, operators can customize and enhance their networks with intelligent functionalities, leading to more agile and programmable radio networks. Using the current standard as starting point, BeGREEN will investigate and propose enhancements to existing O-RAN components, interfaces and functions related to RIC to incorporate energy-efficiency awareness in the RAN optimization decisions and to enable the implementation of AI/ML-based mechanisms.

4.1.5.1 O-RAN RIC architecture

Figure 4-9 depicts the non-RT RIC architecture as part of the SMO [12]:

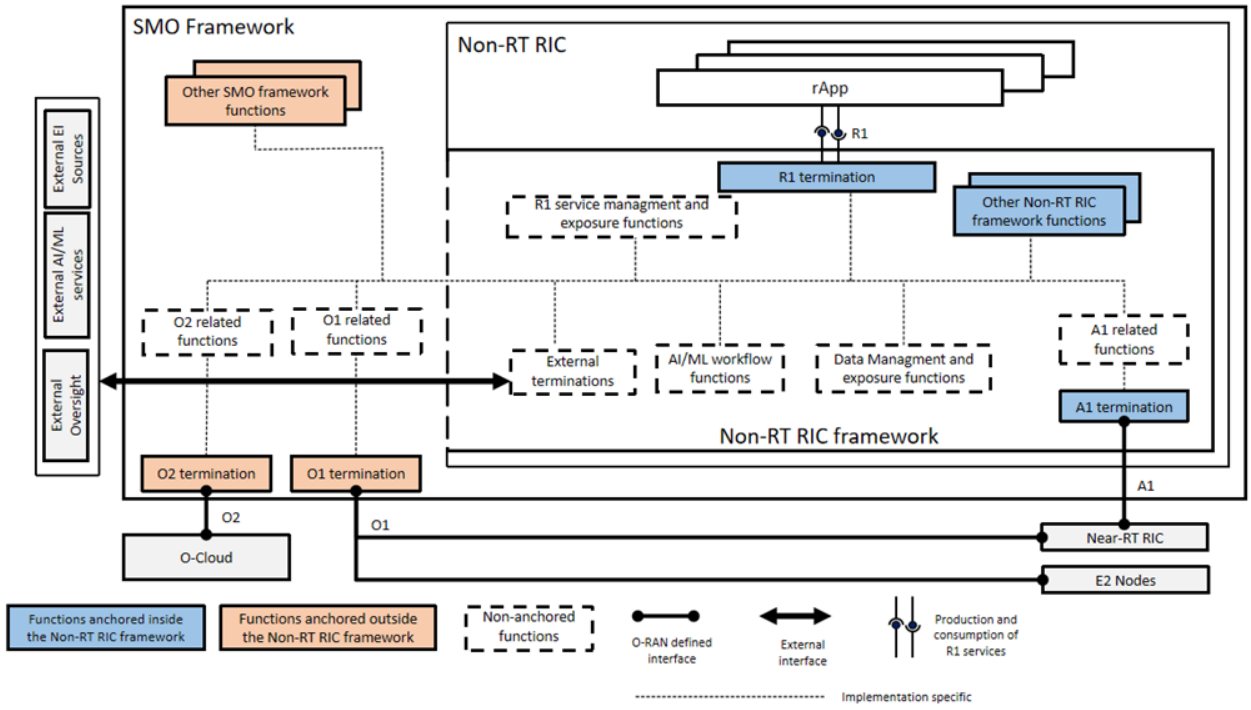


Figure 4-9 Non-RT RIC Reference Architecture [12]

Note that the main functionalities of the non-RT RIC are hosting rApps and implementing the termination of R1 and A1 interfaces. Similarly, the SMO terminates O1 and O2 interfaces. Other functions, such as those related with Data Management and Exposure (DME), AI/ML workflows or Service Management and Exposure (SME) can be located either in the non-RT RIC, in the SMO or even in external components. In this sense, the R1 interface, which is still under definition by the O-RAN alliance [104], plays a key role by exposing to the rApps aforementioned functions, which enable, among others, policy-driven guidance of Near-RT RIC/xApps through the A1 interface, monitoring and management of O-nodes and the Near-RT RIC through the O1 interface, or monitoring and management of the O-cloud infrastructure through the O2 interface. In addition to that, the R1 interface also facilitates the interconnection among rApps (rApp service and analytics. O-RAN is also considering R1 interface as a way to expose RAN analytics provided by rApps to external entities (for example to 3GPP NFs) and vice-versa (for example NWDAF or NEF functions and data torApps) [100].

Similarly, Near-RT RAN intelligent control is provided by the xApps hosted at the Near-RT RIC component, which architecture is depicted in Figure 4-10 [10].

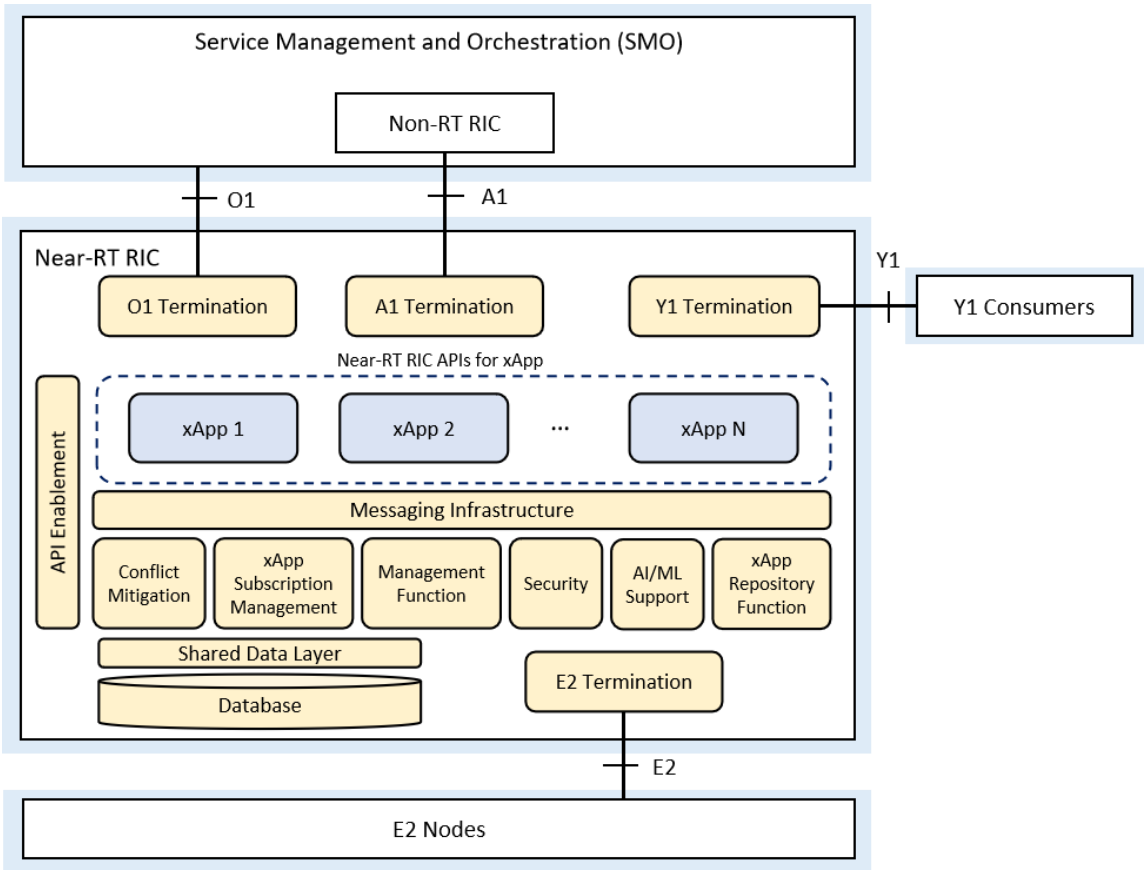


Figure 4-10: Near-RT RIC Internal Architecture [2]

In this case, xApps can make use of the services provided by the Near-RT RIC, mainly the E2 termination and the database, to implement Near-RT control-loop automations and analytics. The E2 interface allows monitoring and managing the E2 nodes according to the E2 Service Models [105], while the database might be used to store and to access to specific network data (for example UE or E2 nodes data). Also, the A1-EI interface enables the xApp subscribe to and receive Enrichment Information (EI) from the non-RT RIC, for example rApps' analytics, data from non-RAN components, and so on. In addition, in a similar way to the R1 in the non-RT RIC, the Y1 interface will allow consumers to subscribe to or request the RAN Analytics Information Exposure (RAIE) service provided by the Near-RT RIC [100], what can be useful for instance in MEC architectures with an RNIS [106] or to enhance application QoE [107].

4.1.5.2 O-RAN RIC AI/ML support

O-RAN has started the definition of the AI/ML support [103], which considers the workflows depicted Figure 4-11. Also, recent specifications of the Near-RT RIC [10] and the non-RT RIC [12] define several options for providing AI/ML workflow services, for example model management, model training, model inference, data preparation, etc., at the SMO, the non-RT RIC, the Near-RT RIC or through external components.

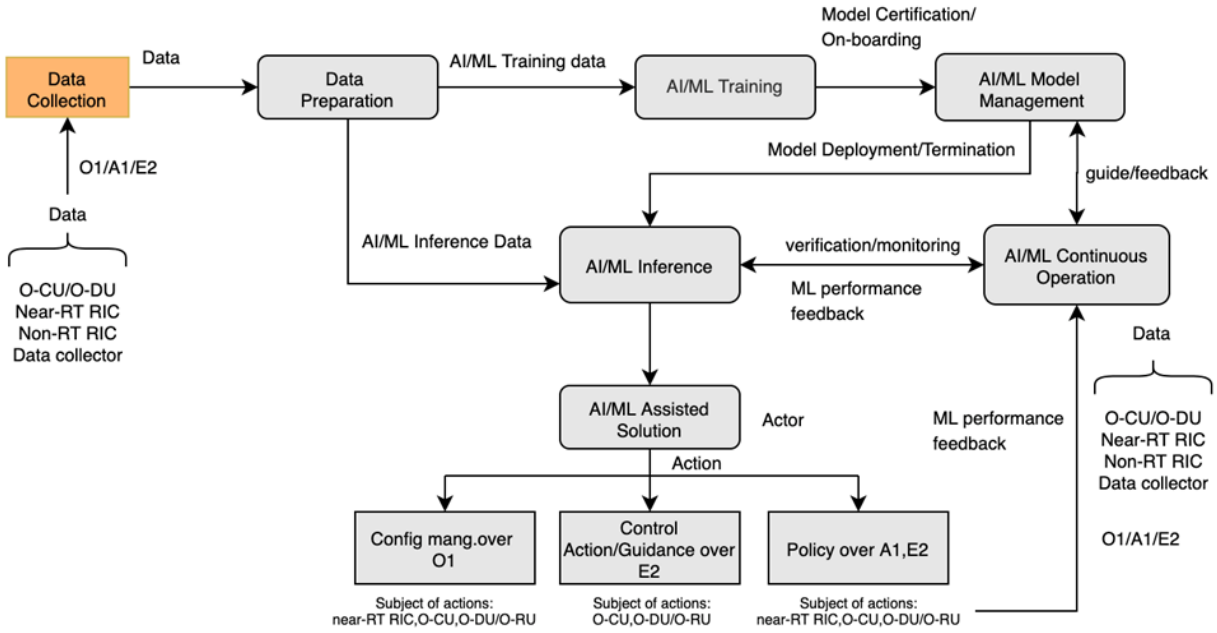


Figure 4-11: O-RAN AI/ML workflows [103]

Regarding AI/ML model deployment, two main options are considered and will be supported:

- Image-based deployment: The AI/ML model is deployed as or within a rApp/xApp instance which contains the ML training or irference runtime in the image. This simplifies the deployment and the update, since the AI/ML xApp is treated as a generic xApp/rApp.
- File-based deployment: AI/ML models are deployed based on AI/ML model files, which are decoupled from rApps/xApps. In such case, xApps need to have access to the AI/ML services, for example model interference or training, provided by the Near-RT RIC, the non-RT RIC, the SMO or other external components.

In any case, the support for multiple types of machine learning as supervised, unsupervised, reinforcement, and federated learning is assured and the architecture is flexible enough to enable diverse deployment scenarios, for example offline/online training at the SMO or the non-RT RIC (rApps), inference at the non-RT RIC rApp online training and inference at near-RT RIC (xApp), etc. The utilization of O-RAN interfaces for supporting AI/ML workflows between non-RT RIC and Near-RT RIC is also still under definition, but initial specifications consider the utilization of O1 to download/update ML models and of A1-ML to exchange model parameters, for example in federated learning approaches.

Within the scope of BeGREEN, the chaining of AI/ML models through rApps/xApps, together with the definition of energy efficiency specific policies and the utilization of enrichment information from different domains, for example RAN, core, application, UEs, will allow the creation of complex energy-aware AI-empowered closed-loop automations, which is one of the main research challenges for O-RAN [99]. As introduced in the previous paragraphs, this will require the definition and exposure of adequate metrics and KPIs through the different O-RAN interfaces (for example O1, O2, E2) and the definition and implementation of adequate policies, methods and services to manage O-Cloud (O2) and O-nodes (O1, A1, E2) operations in an energy efficient way. Regarding the implementation of the AI/ML architecture and workflows, some trends and challenges are described below:

- Federated Learning: Certain type of AI/ML algorithms such as deep learning require a huge amount of network events for the algorithm to outperform classical AI/ML algorithms at training stage, while also demanding higher compute and memory or need to parallelize model training for quicker

training. Performing such expensive operation at xApps level for several VNFs will be energy exhausting and usually not applicable at edge sites [103]. Therefore, federated learning approaches where the model is always trained, monitored and re-trained at SMO/non-RT RIC level, and the model is just deployed at xAppx level for quick inference using real-time RAN data available at this domain, will lead to higher energy efficiency.

- Explainable AI: Explainable AI can help in Energy Efficiency related optimizations in order to (i) identify data influencers which impact EE, (ii) define and calculate KPIs such as EE rating or score helping to understand the impact of an AI/ML algorithms, and (iii) identifying the areas, NFs or components which require AI-driven EE optimizations.
- AI/ML algorithm training and testing: In order to avoid impairing active services performance due to poor performing AI/ML algorithms, O-RAN specification mandates to perform offline training and evaluate model performance before deploying any AI/ML model. How to collect training and testing datasets that are heterogeneous and representative of large-scale deployments is a key challenge [99], especially for AI/ML-based solutions targeting energy efficiency. Envisioned solutions comprehend combining emulators/simulators with real scenarios connected to a common O-RAN RIC and AI/ML Engine.

4.1.5.3 O-RAN RIC application environment

Recent publications analyse and propose approaches to integrate AI/ML support within the O-RAN architecture. In [107] authors provide an exhaustive description of the AI/ML functionality in O-RAN architecture according to the last specifications, analysing in detail the utilization of O-RAN interfaces to support the design of AI/ML-enabled rApps/xApps and the implementation of Reinforcement Learning approaches. In [64] different AI/ML deployment scenarios and their associated workflows are studied (for example considering non-RT RIC or Near-RT RIC as training and/or inference hosts), analysing their application to two different use cases: traffic steering optimization and Cloud Virtual Reality QoE optimization. The authors in [65] describe the implementation of an AI/ML workflow leveraging O-RAN's Software Community non-RT and Near-RT RICs and exemplified with a traffic prediction AI/ML-driven xApp ML model is trained and packaged with Acumos AI platform and deployed as an xApp in Near-RT RIC. Acumos AI is an open-source platform which simplifies the creation, sharing, and deployment of AI applications, standardizing the infrastructure stack, enabling data scientists and model trainers to focus on their expertise and driving innovation. A similar approach called OpenRAN Gym is presented in [108]. OpenRAN Gym is a framework that enables the development, testing, and data collection for AI/ML algorithms applied to O-RAN networks as xApps. Through emulated wireless environment, called Colosseum, xApp developers prototype and evaluate their AI/ML solutions before deploying them in real scenarios.

Several studies leverage O-RAN architecture to design and deploy solutions capable of predicting, controlling, and automating the network behavior under dynamic conditions. Some examples include the use of Deep Learning and Deep Reinforcement Learning (DRL) to predict traffic load [171], beam alignment [172], radio resource allocation [173], etc.

The authors in [174] propose an online learning orchestration framework based on Bayesian online optimization. This framework operates in the non-RT RIC and performs radio resource allocation in intelligent virtualized RANs based on context information, for example channel quality and traffic load. In [173], radio and computational resources are jointly orchestrated from the non-RT RIC using DRL to comply with the service level agreements while the computational resources are scarce. The authors in [175] propose an energy-efficient orchestration framework for NextG that splits and allocates RAN components according to the current resource availability. In [176], a radio resource orchestration framework for 5G applications is presented. It re-assigns network slices dynamically to avoid inefficiencies and Service Level Agreements (SLA)

violations.

The authors in [177][178] present frameworks to orchestrate, place and disaggregate RAN components to minimize the computation and energy burden while accounting for diverse performance requirements. Concerning ML/AI models for orchestration in next generation systems, in [179][180] an architecture for the automated deployment of models in the 5Growth management and orchestration (MANO) platform [181] is presented. Finally, the authors in [182] propose an orchestrator to select and instantiate models at diverse locations of the network to obtain the target balance between latency and accuracy.

Recently, there are several companies which are working on early xApps/rApps to provide energy efficient solutions. Juniper Networks showcases an Energy Savings rApp/xApp running on an O-RAN compliant RAN Intelligent Controller (RIC) platform, which will enable service providers to achieve sustainability goals. This technology will allow service providers to implement features such as Booster Cell Switch off/on, Boost Throughput Power of Neighbouring Cells, and mMIMO RF channel switch off/on, among others [183].

Vodafone, Intel, Keysight, Radisys, and Wind River collaborated to demonstrate a new design and verification framework for dynamic closed-loop power management of O-Cloud resources in multi-vendor Open RAN systems in live networks. The demonstration showed how resource optimization, using Vodafone's network traffic profiles, achieved energy savings without affecting the customer experience. Dynamic modulation of cloud computing cores through power management capabilities, AI algorithms, and several KPIs were leveraged to optimize energy consumption for the O-RAN system. The use of Intel Xeon processors and power management features and accelerators helped achieve energy savings without affecting the subscriber experience or service KPIs. Results shows that energy savings can be higher than 11% [184].

Capgemini and Intel have implemented an "AI-Enabled Energy Savings" use case as an O-RAN rApp to reduce the energy consumption of the RAN by introducing intelligent energy-saving mechanisms in O-RU. The rApp uses real-time monitoring and advanced AI/ML prediction to forecast future energy consumption and carbon emissions, taking intelligent decisions on when to apply energy-saving measures like cell and carrier switch off/on. The rApp uses closed loop automation to learn and improve the energy-saving decisions over time, not only for a single RAN but also for neighboring RAN nodes [185].

In [186], Rimedo Labs, present a comprehensive whitepaper about the O-RAN energy efficiency and the role of the ML algorithms to develop energy efficiency solutions. In their solution, O-RAN can dynamically control the resources and hardware usage by adapting it to actual traffic demand through its Non-RT RIC and Near-RT RIC entities, which can monitor network data and control switching off hardware components. The implementation of ML-based algorithms in O-RAN can significantly improve energy efficiency, and the whitepaper provides specific solutions developed as ML-supported rApps for O-RAN use cases. The authors highlight the importance of considering energy efficiency in an end-to-end (E2E) fashion and coordinating with other features when switching off BSs or by switching RF Channels.

Another E2E 5G O-RAN intelligent control for energy savings including 5G O-RAN BS performance is developed by ITRI and PEGATRON. The solution involves the deployment and simulation of a total of eight BSs, with two BSs being executed on an x86-based server and six BSs being simulated on another x86-based server. It also includes the deployment of SMO and Near-RT RIC platforms. The Non-RT RIC platform is equipped with an ES rApp, which is responsible for monitoring the traffic load of the O-RAN BSs. When the ES rApp detects that the traffic load is below a certain threshold, it determines which BSs can be powered off to reduce power consumption. Prior to powering off the BSs, the Traffic Steering xApp in the Near-RT RIC platform will transfer the UEs they serve to other powered-on BSs. A web-based Graphical User Interface (GUI) is utilized to compare the resulting power consumption levels with and without ES. Additionally, the GUI provides information regarding the O-RAN BSs, Open fronthaul, load, stress, stability, and Security Assurance Specifications testing with M/C/U/S plans and peak throughput results [187].

Finally, the VMware Energy-saving rApp utilized the coverage/capacity umbrella approach [188]. The proposed solution features several components, including the Non-Real Time Radio Intelligent Controller (Non-RT RIC), as well as several open interfaces (O1, O2, E2, and A1), the RAN Element Management System, and an rApp. By leveraging VMware's RIC and ES rApp, the solution aims to optimize power consumption in Open RAN deployments, while maintaining network performance.

4.2 BeGREEN proposed architecture

The BeGREEN proposed reference architecture is shown in Figure 4-12. BeGREEN's Intelligent Plane will include the SMO plus the non-RT RIC, the Near RT-RIC and the developed rApps and xApps, which will be empowered by the AI Engine and the datalake. The proposed AI Engine will provide a serverless execution environment hosting the AI/ML models, offering inference and training services to the rApps/xApps. Also, it will manage the lifecycle of the AI/ML models. Together with the datalake, which will include metrics from different domains (such as RAN, core, ISAC and Edge applications (App)), the AI Engine will allow to design and implement innovative AI-driven rApps/xApps.

In addition to the envisioned Energy Efficiency rApps/xApps, which will be focused on the use cases introduced in Section 3.1 and the optimization strategies described in the following Chapters 5 and 6, a key innovation of BeGREEN will be providing energy efficiency ratings and energy scores through the Energy Efficiency Calculation rApps. These rApps will characterize the energy efficiency of the overall network, and with the help of eXplainable AI approaches, they will identify and quantify the contributions of the different RAN components to energy consumption and the contributions of the different enhancements proposed in BeGREEN to energy saving.

BeGREEN will study and propose enhancements to O-RAN interfaces to better support the implementation of energy efficiency approaches (marked with + in Figure 4-12). Also, it will consider interfaces to and from external components such as the core and the Edge applications (dotted blue line) to enable join optimisations. For instance, rApps may use data analytics provided by the 5GC Control Plane (CP) functions, for example the Network Data Analytics Function (NWDAF), modify UPF performance and energy consumption according to traffic predictions or policies, or modify AI service parameters according to RAN status to minimize the overall energy cost.

As shown in the proposed reference architecture, we will consider components and interfaces to support relay and RIS technologies and to integrate them into the O-RAN framework. Aligned with the discussions in Section 4.1.2, the relay represented in Figure 4-12 can be either an IAB-node or a relay UE. In the case of a relay UE, it is connected with the RU through the Uu interface. In the case of the relay being an IAB-node, this interface is also used to connect the IAB-MT with the RU and, in addition the interfaces F1-C/F1-U represented in dotted lines in Figure 4-12 are used to connect the IAB-DU of the relay with the CU-CP/CU-UP of the donor. The physical realisation of these interfaces is done on top of the Uu interface between the IAB-MT and the RU. AI/ML-based relay reconfiguration decisions, for example relay activation and deactivation to improve coverage and capacity and reduce energy consumption, are made by rApps in the non-RT RIC. Then, the enforcement of these relay reconfiguration decisions is done by the relay control function of the SMO through the O1+ interface. A similar approach will be followed to support RIS components through rApp, xApps or external components. RIS orchestration and control components will be implemented at the non-RT RIC and near-RT RIC levels respectively and will interact with other O-RAN components via O1+ and E2+ interfaces.

Furthermore, BeGREEN will explore the heterogeneity in the RAN depending on the reference scenarios described. BeGREEN will focus on comparing mMIMO cellular environments with CF-mMIMO deployments. In this comparison, it is expected that co-located solutions will be more energy efficient in rural environments, while CF-mMIMO will be more energy efficient in dense urban environments. However, it is difficult to find

the sweet-spot for deploying one or the other solution, as well as the sizing (for example number of radios, node separation, etc.). This is part of the work that will be pursued in BeGREEN and presented in D2.2.

Finally, as depicted in Figure 4-12, hardware acceleration and optimisation will be considered to reduce energy consumption in different components of the architecture, especially in the DU and the CU-UP.

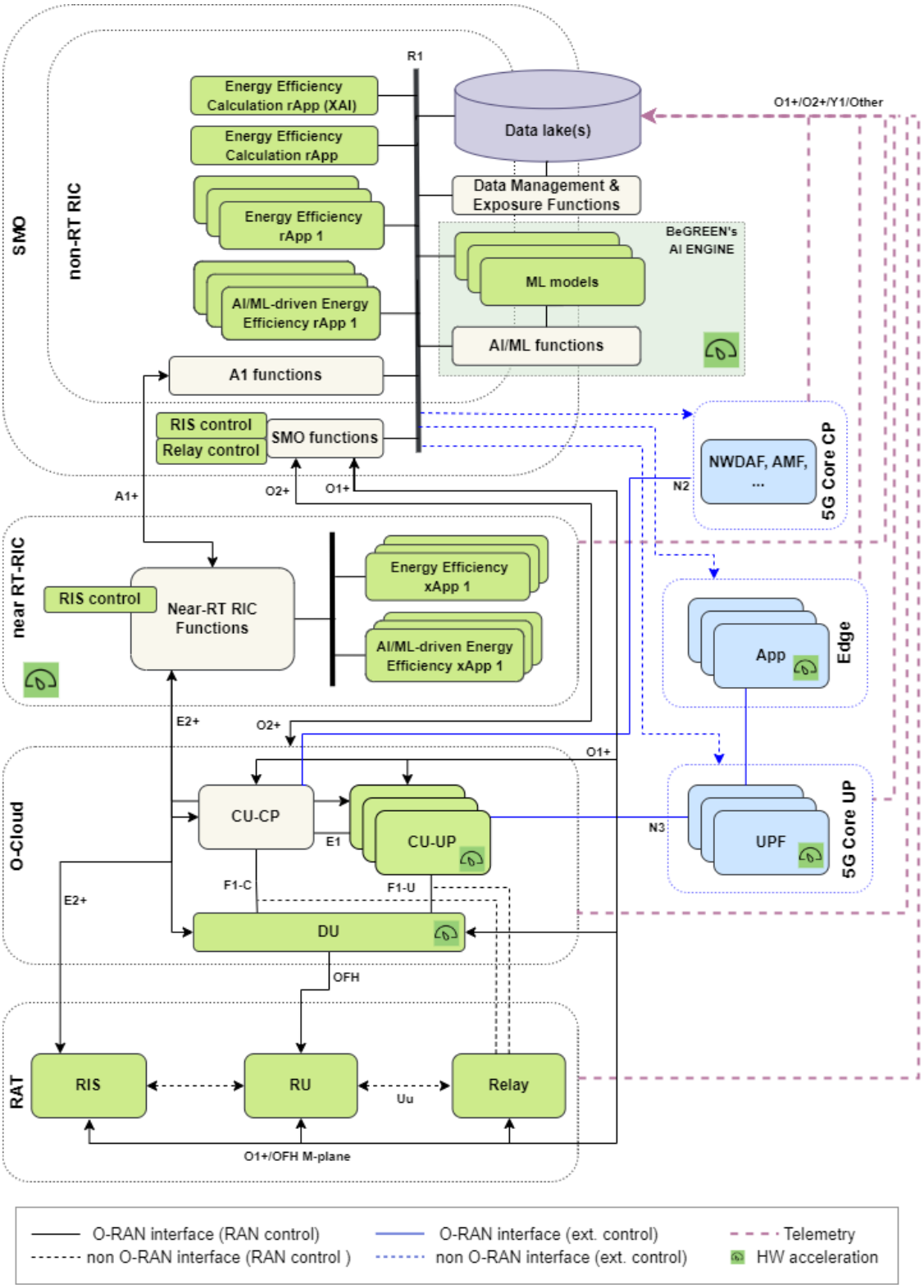


Figure 4-12: BeGREEN proposed architecture.

5 Energy-Aware Control of RAN

This chapter introduces various energy saving technologies that consider hardware and software, with a special look into the BeGREEN focus.

As depicted in Figure 5-1, the major energy consumer element in the mobile networks is the RAN with an energy consumption of 73% of the network's total energy consumption. The reason why RAN consumes more energy than other parts of the mobile networks is because the equipment in RAN, e.g. BSs, absolutely outnumber the equipment in other parts of the networks. Although the number of BSs increases in proportion to service coverage expansion, this is necessarily not the case with the equipment in the core network. In addition, it is shown that BS equipment accounts for 50% of the total energy consumption (followed by air conditioning equipment, which consumes 40%) [189].

In general, 5G equipment consumes more energy due to its increased number of transmitters in the case of mMIMO, which is 8-16 times more than LTE, and to its increased channel bandwidths, which are 5-10 times more than LTE [190]. It is inevitable that equipment's output power increases to improve its performance and, hence, simultaneous efforts to reduce energy consumption are required to reduce carbon emissions [191].

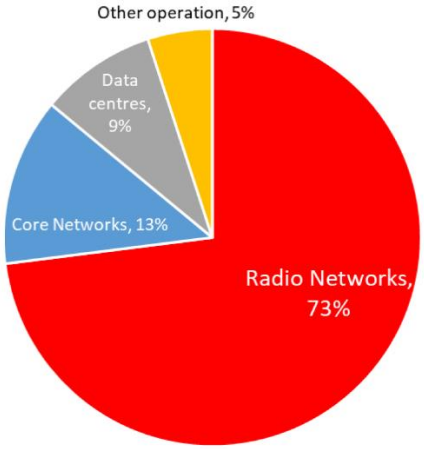


Figure 5-1 Energy consumption in mobile network

5.1 Radio unit (RU)

This section introduces different techniques investigated in BeGREEN that have the potential to substantially improve the RU energy efficiency. First techniques such as Envelope Tracking (ET) and Digital Pre-Distortion (DPD) are introduced in order to decrease the RU power amplifier consumption followed by methods to optimize the RU efficiency in low traffic conditions such as switching off channels (beams) and power amplifier data blanking.

RU itself is a major energy consuming part of the RAN. The RU RF assembly and, especially, its RF amplifiers, are major energy consuming elements. To achieve the network end-to-end QoS, RF signal distortion should be minimized. Specifically, linear RF amplifiers should be applied, coupled with adequate back-off, to avoid amplifier saturation which is required to keep the error vector magnitude low. However, such design considerations can significantly increase the energy consumption for 5G Orthogonal Division Multiple Access (OFDMA) signals and, particularly, at mMIMO enabled 5G systems.

5.1.1 RU energy consumption reduction

As an example, Table 5-1 shows the energy consumption ratio of RU to the energy consumption of RAN used

by SK Telecom (in the referenced study RU and baseband unit power consumptions are considered as the total) [192].

Table 5-1 Percentage of RU Energy Consumption in a RAN [REF]

	Energy Consumption per 5G cell @ maximum cell load	Energy Consumption in a 5G commercial network	Energy Consumption in a LTE commercial network
RU	78%	88%	82%
BBU	22%	12%	18%

Within the RU itself, the radio frequency (RF) components, that is, the power amplifier plus the transceivers and D/A converters, have been identified as the largest energy consumers, typically using about 65% of the total RU energy. The cooling system, the digital signal and base-band processing as well as the alternating current (AC)-direct current (DC) converters follow with an energy consumption of around 17.5%, 10%, and 7.5%, respectively.

In order to address the challenge of reducing the RU power consumption and improve the network energy efficiency the following methods can be applied:

- Improving the RF power amplifier efficiency by using Envelope Tracking (ET) method.
- Reducing the RF power amplifiers consumption by using Digital Pre-distortion (DPD) techniques.
- Dynamically turning the RUs on and off (Sleep and Idle modes) according to the actual data traffic distribution patterns.
- Using renewable energy sources for powering the RU sites, including solar power and smart lithium batteries.
- RU sharing between MNOs.
- Using more power efficient components in the RU as depicted in Figure 5-1.

	Item	Key Factor
1	Main chipset used	FPGA or SoC
2	Number of digital chipsets	How many digital chipsets are required?
3	Number of Tx/Rx	Tx/Rx antenna array per Digital Pre-Distortion (DPD)
4	Power amplifier element	LDMOS or GaN
5	Others	Circuit, fan, etc.

Figure 4- 1 Major design factors to minimize RU energy consumption [192]

5.1.2 RU power control

As above mentioned, RU can be turned off when traffic demand goes done and less network capacity is required, trying to maintain a zero Watt at zero traffic guide-line. mMIMO systems include a plurality of RF amplifiers that makes energy control essential.

Energy efficiency is essential in mMIMO solutions. Figure 5-2 depicts the BS RU energy efficiency adjustment along with the traffic load change during the the daytime [193]. Reducing the BS capacity by tuning off part of its beams can cause energy consumption reduction.

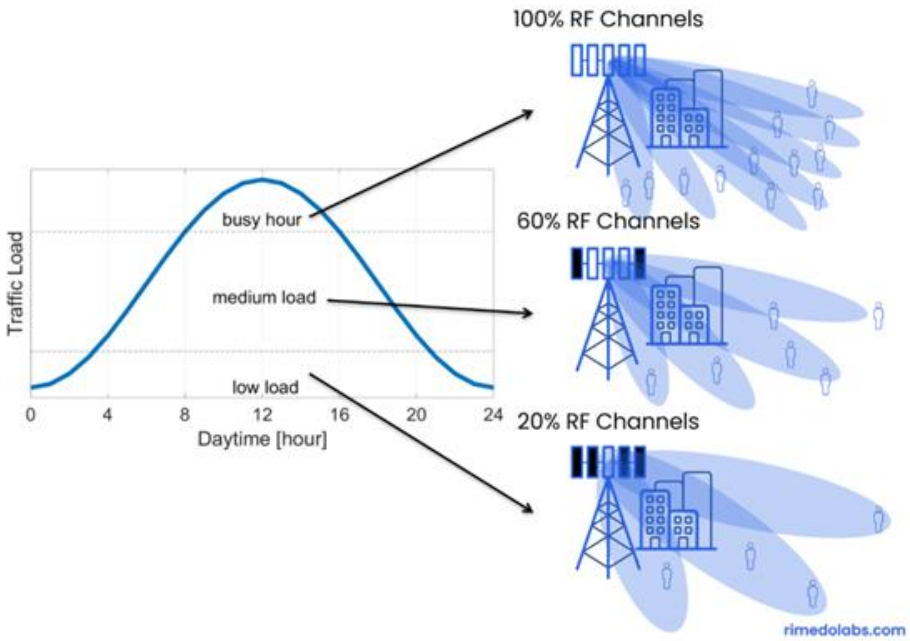


Figure 5-2 Traffic-load-sensitive illustration - switching off RF channels

Controlling the RU operation can be accomplished using the interface connecting the RU with the DU. O-RAN 7-2x interface can be used for the development of this functionality. O-RAN specifications are leading the RAN industry towards a multivendor interoperable open 5G RAN.

NETCONF is a client-server control and management standard for network devices which used in O-RAN specifications. NETCONF uses Extensible Markup Language-based data encoding for the network elements configuration data and also for its protocol messages.

O-RAN specifications use NETCONF standard protocol for RAN elements management including O-RU management and control. O-RAN management and control definitions are listed in O-RAN Working Group 4 (Open Fronthaul Interfaces WG) Control, User and Synchronization Plane Specification [194] and Management Plane Specification [195].

O-RAN specifications enable controlling RU beams status. NETCONF clients enable RU to put the tx-array-carrier / rx-array-carrier to “Sleep” by setting value of parameter "active" in the corresponding tx-array-carrier element / rx-array-carrier element to "SLEEP".

To achieve energy saving, O-RAN NETCONF client uses the existing parameter [tr]x-array-carrier::active to activate or de-activate [tr]x-array-carriers. NETCONF client configures the parameter to INACTIVE to deactivate a specific carrier. This causes the parameter [tr]x-array-carrier::state transition to DISABLED. The outcome is reduced power consumption of the RU.

Additionally, when [tr]x-array-carriers are inactive and energy-saving-enabled is set, the RU can turn off circuitry associated with carrier processing to further reduce the RU power consumption.

Data blanking: For cases when the Resource Element Block (REB) slot resources are not fully allocated, Data Blanking can be carried out. Data Blanking refers to the process of not sending physical resource blocks data bits on the fronthaul interface for some time and frequency resources in DL and/or UL directions. The objectives of data blanking technique are dynamic reduction of the fronthaul interface bitrate, power saving, and for providing an interval for antenna calibration. Data Blanking should be performed by the DU in DL. In the UL case, the RU shall follow the DU instructions received in the C-Plane message(s). DU advanced AI/ML-based scheduler should be an adequate enabler for such data blanking functionality.

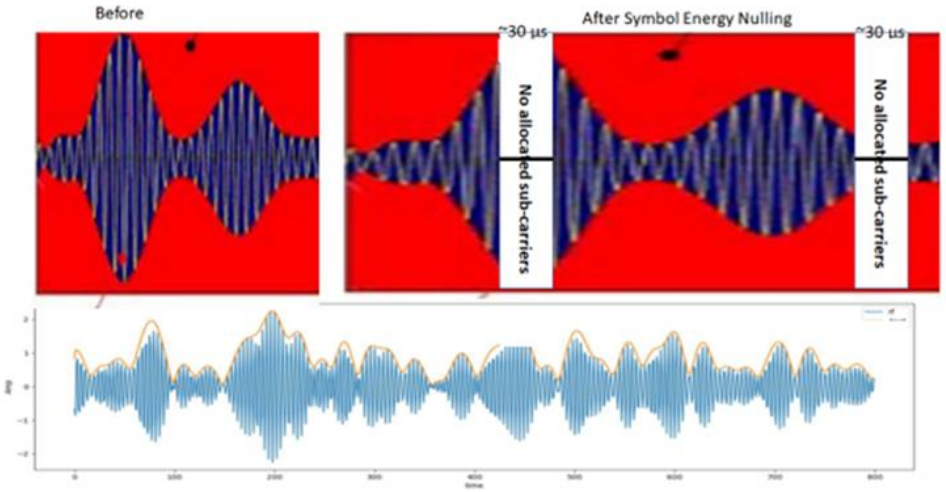


Figure 5-3 Dynamic energy management method using data blanking.

BeGREEN is considering the above techniques and dynamically selected data blanking applied to the RU. This method requires prompt response of the RU data processing following an analysis of the related slot control plane packets. More specifically, as part of the BeGREEN research the RU is upgraded to support data blanking independent of DU control plane messages. When RU identifies a non-allocated symbol at the incoming User Plane (UP) messages, RU will immediately blank this symbol power to decrease the RU energy consumption. This method is illustrated in Figure 5-3 where two consecutive time slots (in the lower part) with two symbols blanked by the RU data proccessing (at the right side of the figure) are shown.

Another embedded feature in the RU is lowering its power consumption in-line with DU allocation of resource elements. When DU reduces the allocated subcarriers in a symbol (due to a decrease in traffic load) the RU automatically reduces its instantaneous energy consumption. This is achieved by allocating less subcarriers, avoiding any power on subcarriers for which REBs were not allocated.

RU Power Measurements: Monitoring RU power is essential to **track the RU power consumption at all times as it is a major contributor to the network energy usage**. For RU measurements report *epe-stats* O-RAN message can be used to report RU power monitoring results (max, min, and average). The RU can monitor and report its power consumption using customized status reports.

5.2 Distributed unit (DU)

In addition to the high power consumption of the RU, depending on the configuration and load, the DU also serves as a significant factor in the overall BS power consumption. The modules which require the highest computational complexity in the DU are the L1 upper PHY algorithms, as described in the following.

Low Density Parity Check (LDPC) channel decoder, and sphere decoder – near maximum-likelihood receiver, where the latter enables SU-MIMO and MU-MIMO for cases where there is high correlation between the layers – due to correlation in the transmit antennas, the receive antennas or in the channel. The sphere decoder inputs are the received signal samples and the channel estimation matrix, whereas its outputs are the log-likelihood ratios (LLRs) per bit, which are fed to the LDPC decoder. As opposed to linear receivers, which operate in two stages, that is equalization and demapping, the sphere decoder does not pass through the equalization stage as LLR generation is embedded inside it. mMIMO related operations such as beamforming weights calculation based on SRS.

Many of the current DUs in the market are based on Intel x86 architecture, implemented on the CPU cores. In order to reduce power consumption, efficient and careful utilization of these cores need to be achieved. This can be done in a any of the following ways:

- Fully utilizing the vector processor unit
- Designing the optimal SW architecture
- Smart splitting of tasks into threads
- Parallelizing of the common functions
- Smart memory management strategies, for example how to transfer data from different memory locations, what data to put in cache, etc.

In addition to the above, the following processor archtiectures can be taken into account:

- Using ARM architecture which is inherently more power efficient than x86 architecture
- Using CPU architecture (either ARM of x86) and offloading some of the most computationally extensive modules to a GPU
- Using a GPU for performing the full DU, for example using the NVIDIA Aerial software development kit

In addition, ORAN suggests a number of new methods that are relevant for DU power efficiency [93] [216]:

- O-Cloud energy saving (discussed in Section 3.1.2.3) is expected to define ways to reduce cloud power consumption through optimization of resources like scaling of HW components and optimizing CPU power states. The same way that non-RT RIC optimizes radio resources, O-Cloud resources, and specifically DU processing, can also be optimized based on policy driven from non-RT RIC. Objective of O-Cloud energy saving can include parameters such as:
  - Configuration of logical nodes and resource pools
  - Scaling In & out of nodes
  - CPU/GPU power and memory usage, frequency, etc.
- Carrier and cell switch on/off, RF channel reconfiguration on/off and advanced sleep mode discussed in Section 3.1.2.4 will first and foremost reduce the RU power consumption, but will also help overall BS power consumption by re ducing DU processing. For instance:
  - If a carrier is switched off, the DU processing associated with it will also be shut down.
  - If an RF Channel reconfiguration policy decides to reduce the number of antennas, the DU processing that is associated with these antennas, such as beamforming weight calculation, will now reduce its complexity and hence consume less power.
  - Advanced sleep modes can also send DU to some level of energy conserving sleep mode, for example by reducing the CPU clock rate. Utilization of energy efficient scheduling with the corresponding sleep mode will reduce both RU and DU power consumption.

5.3 Central unit (CU)

The power consumption of the CU in O-RAN can vary depending on several factors such as the hardware used, the specific implementation of the CU, and the workload being processed. Accordingly, the power consumption of the CU in O-RAN is expected to be relatively low compared to traditional centralized RAN (C-RAN) architectures, since the O-RAN is designed to distribute processing functions between multiple nodes, reducing the amount of centralized processing that needs to be performed.

The O-CU and O-DU shall provide Cloud-Native Network Function (CNF) level energy efficiency counters and KPIs (for example power consumption, traffic load, data volume, throughput), which shall be reported through O1 interface to the SMO or through North Bound Interface to external tooling to perform energy efficiency control. According to [223], O-CU/O-DU hardware (for example CPU, accelerators, Network

Interface Card (NIC), fan(s), etc.) shall have the capability to measure and report power consumption values to the O-Cloud; then the O-Cloud SW must collect these values and provide energy and environmental (PEE) parameters and report them through O2 interfaces.

5.3.1 CU energy and power consumption reduction

There are several strategies to reduce energy and power consumption on the O-CU in O-RAN. The O-RAN Alliance has identified new requirements for energy efficiency in various streams, including Cloud Infrastructure, O-CU/O-DU, and others [224], and can be grouped in the main following ones:

- Hardware optimization: Regarding hardware acceleration and optimization for the O-CU , CPUs need to have flexibility to select different ACPI power state profiles attending to traffic load that will help to reduce cores in use, frequency or any other relevant parameter affecting to overall O-CU/DU server power consumption. This includes the support of CPU C-state for power reduction and P-state to run processors at different voltage and/or frequency levels. ARM processors are a good choice for powering the O-CU due to their low power usage and high scalability. They are well-suited for the wireless environment of O-RAN networks and are commonly used in mobile devices. However, ARM processors tend to have lower per-core performance compared to x86 processors. x86-based servers are a typical for IT implementations since they have higher per-core performance compared to ARM processors, which can be important for workloads that require high processing power.
- Sleep Mode: This involves putting certain components of the system into a low-power state when they are not in use. For example, if a particular NIC card is not being used, it can be put into sleep mode to reduce its power consumption. Similarly, if a particular CPU core is not being used, it can be put into a low-power state to reduce its energy consumption. This can be achieved through various techniques, such as dynamic voltage and frequency scaling (DVFS) and CPU idle states. By using sleep mode following the concept of "Zero Watt at Zero Bits", the O-CU can reduce its overall energy consumption, which can lead to cost savings and improved energy efficiency.
- Virtualization and workload consolidation: The third strategy is to leverage virtualization and workload consolidation to optimize resource usage and reduce power consumption. By running multiple virtual instances of the O-CU on the same physical hardware, the overall power consumption can be reduced. Additionally, workload consolidation allows for more efficient use of resources by dynamically allocating them to the instances that need them most.
- Code Optimization: Optimizing the O-CU code can significantly impact its energy efficiency, with specific attention given to NIC card management and express data path (XDP) application performance. Optimizing algorithms and data structures can reduce the number of CPU operations needed to process data, minimizing power usage for data-intensive Open-RAN networks. For efficient NIC card management, offloading and packet steering techniques can reduce CPU workload and overall power consumption. Network packet processing can be optimised using XDP to bypass the generic packet handling in the kernel and apply specifically tailored processing for UP traffic. Careful consideration of workload and hardware components is necessary to optimize the O-CU code for energy efficiency while maintaining network performance and reliability.

5.3.2 Integrated sensing and communications

The objective of ISAC is to unify sensing and communication functionalities in a spectrum/energy/cost-efficient way with (ideally) a single transmission, a single device, and a single network infrastructure, allowing the delivery of many use cases that harness environment-aware scenarios. Both functionalities are in parallel benefiting from the use of higher frequency bands, larger antenna arrays, and miniaturization, thereby becoming increasingly similar in terms of hardware architectures, channel characteristics, and signal

processing. Nowadays, in a networked ISAC, sensing information can be shared cross-layer, cross-module and cross-node to achieve fully integrated communication and sensing, and the O-RAN ecosystem can help in this direction to optimize the energy efficiency of the RAN.

The successful transmission of the information comes at the cost of wireless resources, for example spectrum, spatial, and energy resources. Accordingly, efficiency is a metric that evaluates how much information is successfully delivered from the transmitter to the receiver, given (in many occassions) limited available resources. While the shared use of wireless resources between sensing and communications may potentially increase the efficiency and, therefore, allow achieving an integration gain, it is unclear how to measure this gain in a quantative manner. This accounts for the reason of why a meaningful metric is needed to measure the resource efficiency of sensing. Together with the spectral/energy efficiencies of communications, one may readily see how much efficiency improvement is achieved by using the ISAC transmission over that of individual sensing and communication designs.

Unlike energy efficiency (EE) maximization in traditional communication systems, whose performance is generally guaranteed by the SINR or transmission rate constraints, designs in ISAC system should take both communication performance and radar sensing performance into account.

5.4 BeGREEN proposed developments

Regarding the energy efficiency optimization enhancements conducted by BeGREEN the following components are being considered:

- Performing research and simulations in ML-based DPD and ET modules of the O-RU that has the potential to drastically reduce the overall network power consumption due to the fact that the O-RUs are currently the major power consumers in every cellular network.
- Providing an end-to-end link for BeGREEN lab test that includes an O-RU at 3.5 GHz band with interfaces to enable/disable the O-RU power amplifier as well as monitoring the O-RU power consumption in real time. This interface permits to develop and test applications related to the network power efficiency. RU blanking techniques will be integrated within the RU to enable energy savings inline with traffic load
- An accelerated energy efficient CU, supported by ARM or x86 servers able to support fast and reliable xApps/rApp focused on suporting E2E energy-efficient ML-based mechanisms.
- Define a telemetry framework that, articulated with the proposed architecture, permits supportting fast measurements and statistics from raw data obtained from southbound interafces and into a datalake to support fast xApps/rApps funtionabilities leveraged within the energy eficient goal of BeGREEN.
- For offloading L1 algorithms, there are two configurations to be evaluated:
  - Running the whole L1 on ARM architecture, which is more power efficient than x86 architecture
  - Running computationally extensive algorithms such as LDPC decoder, sphere decoder, and MIMO beamforming algorithms, on a GPU, while running the rest of the L1 algorithms on the ARM cores
- Analysing O-RAN power optimization methodology, for example O-Cloud energy saving, carrier and cell switch off/on, RF channel reconfiguration off/on and advanced sleep modes; and examining their utilisation inBeGREEN architecture.
- In relation to achieving an energy efficient RAN with relay nodes BeGREEN will analyse the different interference scenarios that can arise in a relay-enhanced B5G network. Based on this, suitable frequency assignment strategies and resource allocation mechanisms will be devised.

- Regarding the use of ISAC, BeGREEN will feature the RAN with sensing capabilities to allow for better use of the communication capabilities towards the UE.

6 AI/ML Enhanced RAN

This chapter sets out some AI/ML-based optimization strategies that address the BeGREEN objectives to introduce AI/ML services for enhancing energy efficiency of the BeGREEN RAN infrastructure. These strategies are based on AI/ML algorithms that optimise energy efficiency of software-based network user plane processing functions and develop service-aware AI/ML algorithms that optimise overall system energy efficiency of RAN and edge infrastructure.

6.1 Optimization strategies

The following sections describe the optimization strategies that the BeGREEN project will use to achieve an AI/ML enhanced RAN for energy saving purposes. This includes energy usage measurement, dynamic adaptation of user-plane functions, energy aware coverage and capacity optimization with relay nodes, energy-efficient resource orchestration in virtualized RANs with shared computing infrastructure and AI edge services and service-aware energy-efficient RU control.

6.1.1 Energy Usage Measurement

The basis for a holistic optimization strategy that can help to reach energy efficiency targets is measuring the energy usage of entities in the network prior to the application of energy saving enhancements so that the required energy reduction in those entities can be verified.

In [248] the authors study current and future wireless networks from the viewpoint of energy efficiency and sustainability to meet the planned network and service evolution toward, along, and beyond 5G. Quantifying the energy consumed by each BS is the starting point for the research done in that study.

A benchmark on the measured energy usage of network entities under various network conditions can be created and the contribution of various factors to energy reduction can be analyzed. This can be used as a basis for comparison of energy usage before and after energy saving enhancements are applied to the network. The calculations for energy usage benchmarks can be carried out in the AI Engine in the BeGREEN proposed architecture (Section 3.2) using the RAN metrics and compute metrics which are stored in the datalake.

Explainable AI has experienced a significant growth over the last few years. This is due to the widespread application of machine learning, particularly deep learning, that has led to the development of highly accurate models that lack explainability and interpretability [247]. Explainable AI in which humans can understand the reasoning behind decisions or predictions made by the AI contrasts with the "black box" concept in machine learning, where even the AI's designers cannot explain why it arrived at a specific decision.

With the advancement of Explainable AI [246], it is possible to identify influencers of energy consumption beyond traffic. Using VNF telemetry data, influencers can be identified, for example interference or mobility in case of RAN. Using the knowledge of such influencers, it is possible to calculate energy efficiency ratings and to calculate energy score for zooming into areas that need additional orchestration for achieving maximum energy efficiency.

Explainable AI and ML can be used to generate values for the marginal contribution of various variable network factors, such as load and interference, to the energy usage of each network element where the power used is measurable. Using this data, it will be possible to quantify the contribution to energy savings from the energy efficiency enhancements implemented in the BeGREEN project.

6.1.2 Dynamic adaptation of the energy consumption of software-based user-plane network functions

The energy consumption of software-based implementations of user-plane functions such as the UPF or the CU-UP is highly dependent on network traffic workload. Virtualized or containerized architectures, as is the case of O-RAN's O-Cloud, open the door to power consumption-aware optimizations in order to enhance the energy efficiency of IT servers. In addition, the dynamic tunning of NFVs can be empowered with AI/ML-based predictions (for example expected load or energy consumption) and/or with AI/ML-based algorithms inferring the best configurations.

As reported in [92][93] and [95], increasing the overall energy efficiency in O-RAN involves considering three main pillars alongside the hardware and software characteristics of the servers hosting the VNFs: (i) EE related KPIs definition and reporting, (ii) mechanisms to apply EE optimizations and (iii) EE-aware intelligent control of mechanisms and components. Accordingly, at the user-plane side the following requirements and features should be considered:

1. Measure, provide and correlate performance and EE related KPIs and Power, Energy and Environmental (PEE) parameters, including the hardware component level (for example CPU, NIC, power supply…), the workload level (for example pod, node, host) and the RAN level (for example energy consumption, traffic load, data volume). [96] and [97] provide relevant KPI definitions within the scope of this use case.
2. Ability to apply dynamic horizontal (for example by increasing or reducing the number of VNF instances, or migrating active VNF instances to free resources) and vertical scaling (as by increasing or reducing the number of available resources of active VNFs). An application of the former comprehends switching off CU-UP and/or UPF [94]instances during off-peak periods, while applications of the latter are usually focused on managing the pool of CPU resources by controlling CPU sleep modes/C-states/P-states[92][94].
3. Implementation of intelligent control of VNF instances and the resource pool according to the measured and objective KPIs and the available dynamic management mechanisms. In the case of O-RAN architectures, this involves the implementation of rApps which, according to RAN and O-cloud measurements and the defined KPIs, assist the SMO for optimizing the energy efficiency of the O-cloud [93]. A similar approach can be followed at the 5G Core, where Application Functions (AF) with access to Network Data Analytics Function (NWDAF), can apply control-loop automations to guide orchestrator decisions or VNF operations exposed by the NEF[98]. The implementation of AI-driven algorithms is also a key enabler to enhance decision-making, for instance by using traffic load predictors or by developing Reinforcement Learning (RL) solutions.

Recent O-RAN specifications have started the definition of Network Energy Saving use cases, including O-Cloud Resource Energy Saving Model use case [93], which has the aim to reduce power consumption of O-Cloud components without impairing the network performance. Although still under definition, envisioned optimizations will join compute (O-Cloud, via O2) and RAN (O-Nodes, via O1) data to feed AI/ML-based rApps capable of guiding O-Cloud configurations and actions through the Federated O-Cloud Orchestration and Management and the Network Function Orchestration (NFO) components in the SMO, such as shutting down a specific node or managing its CPU frequencies. A similar approach is proposed in [147], which presents a project from the Open Networking Foundation targeting to build a RAN platform and PoC to demonstrate different energy efficiency techniques. The initial concept considers the optimization of CPU utilization through the RAN and Core infrastructure, by tuning CPU P/C states or scaling in/out the NFV instances, and according to AI/ML-based traffic predictive models fed with real world datasets.

A ML-based framework is proposed and validated in [148] which uses Bayesian methods to optimize the dynamic tuning of the UPF aiming to minimize power consumption and packet drops. First, through an

offline phase, the model is trained and evaluated with different load levels in order to obtain the best configurations. Then, in the online phase, a classifier selects the best available configuration according to the actual load and resource status. Finally, the dynamic tunning of the CPU is performed using Linux *cpufreq* subsystem tools, which allow to define governors defining the policies to be used for frequency control. A different CPU management mechanism for Intel-based processors is presented in [149], which is called Speed Select Technology Core Power (SST-CP) and enables a user-defined and flexible allocation of CPU core prioritization. This allows pinning high-priority workloads (for example high loaded UPFs or UPFs serving high priority traffic) to specific cores, which might use higher frequencies, while low-priority workloads (for example low loaded UPFs or UPFs serving low priority traffic) are assigned to cores using lower frequencies, thus reducing the overall energy consumption of multi-tenant or multi-slice servers.

6.1.3 Energy-efficient resource orchestration in virtualized RANs with shared computing infrastructure and AI edge services

There have been quite a number of pioneering works on the vRAN orchestration that embraces and builds upon the Open RAN paradigm to provide intelligent solutions on resource allocation for the deployment of vBSs over a shared cloud computing platform (as [196]) and provide energy-aware solutions (as [197]) to optimize the energy consumption of underlying computing resources.

In the spectrum of computing resource allocation, the work of [198] introduced a Bayesian learning model to optimize radio policies subject to hard power consumption constraints. Concordia [199] addressed sharing computing resources with latency-elastic applications. For a given network optimization/automation objective, it is also important to know how to select which data-driven models should be deployed and where, which parameters to control, and how to feed them appropriate inputs. OrchestRAN [200] proposed a novel orchestration framework to execute in the non-Real-time RIC to automatically compute the optimal set of data-driven algorithms and their execution location (for example in the cloud, or at the edge) to achieve high-level control/inference objectives specified by the Network Operators while meeting the desired timing requirements and avoiding conflicts among them. O-RAN [201] introduced a large-scale testing framework with software-defined radios-in-the-loop to evaluate O-RAN near or non-real-time RIC algorithms. Making a decisive step forward towards cost-effective implementation of virtual and Open RAN, vrAIn [196] was the first work to jointly optimize the CPU allocation and radio policies for a given number vBSs deployment, and [197] extended the idea to share hardware accelerators to further bring down energy consumption and costs.

Other works study the deployment of AI services at the network edge, specifically, MVA. Some of them focus on performance optimization by using different strategies such as variable encoding, caching, visual tracking, or adaptive compression [202][203][204][205]; or study the trade-off between accuracy and latency [206][207][208].

Other studies, in turn, address the resource orchestration in MVA systems. For example, [209] and [210] search greedily in real-time for the most resource-prudent configuration; [211] and [212] allocate computing resources and decide the image compression (or, video quality) and neural network model. Although some works like [213][214][215] consider the energy in edge AI services, none of them consider the energy consumption at the network infrastructure (BS and edge server), only at the user device. The joint orchestration of the network infrastructure plus edge AI services is still an open issue in the literature.

6.1.4 Relay node assisted energy-aware coverage and capacity optimization

One of the challenges faced by MNOs is to improve RAN coverage and capacity while minimising energy consumption. Traditionally, network planning and optimisation tools have been used for this purpose, for example by optimising the placement of BSs, adjusting power levels, etc. In this context, SON plays a critical role in effectively improving cell coverage and capacity and reducing energy consumption. SON is a network

operation and maintenance technology that uses advanced algorithms and machine learning techniques to automatically deal with these challenges [150]. In this context, Coverage Capacity and Optimisation (CCO) is one of the key SON functions. The CCO function [150] continuously monitors the network performance and adjusts the power levels and antenna parameters to maximize network throughput and capacity and guarantee certain quality requirements. On the other hand, Energy Saving [151] is another relevant SON function that aims to decrease the operator's energy consumption by adjusting BS power levels (or even switching off BSs) in time periods with low traffic while satisfying the coverage, capacity and quality requirements. Other SON functions such as Inter-cell Interference Coordination or Mobility Load Balancing play a relevant role for the reduction of energy consumption.

The deployment of relays in wireless networks can also contribute to obtaining coverage and capacity improvements and reduce energy consumption. In fact, relays have been traditionally used for coverage extension and increasing the RAN capacity [33]. In recent years, the use of relays has gained interest and it is expected to take a relevant role in B5G RAN. As mentioned in Section 3.1.1.1, the 3GPP has been working on the introduction of new relaying technologies, namely the IAB as an alternative to fibre backhaul, the use of relay BSs placed in moving vehicles, and the UE-to-network relaying feature in which a RUE (that is a UE with relaying capabilities) can relay the traffic of another UE to and from the network in a two-hop link by using the D2D sidelink.

The inclusion of relays can be useful for different purposes such as mitigating signal blocking in millimetre wave deployments or increasing coverageand capacity in high-density areas leading to a reduction in the number of BSs to deploy. The deployment of relays can also lead to energy savings through the reduction of transmit power in mobile networks thanks to better propagation conditions in the involved links. The assessment of the energy savings that can be achieved by means of relaying in a wireless system has been studied in previous works [27][31] and [130] as mentioned in Section 4.1.2.

Several challenges arise for exploiting the use of relays for the optimisation of both energy efficiency and spectral efficiency in B5G networks, such as determining the adequate place to locate fixed, the power and resource allocation at the BSs and the relays, user to cell association and relay selection, relay activation/deactivation, etc., Table 6-1 summarises the state of the art in this field presenting relevant works for different considered problems and indicating the methodology used in each case for solving the problem.

Table 6-1 State of the Art in Relay Deployment and Optimisation

Applicability	Reference	Considered Methodology
Assessment of energy savings achieved with relays	[27][31] [130]	Theoretical/simulations
Relay placement for coverage extension	[152]	Greedy Solution
	[153]	Particle Swarm Optimisation
	[154]	Genetic Algorithm
	[155]	Neural Networks
	[156]	Heuristic
Resource allocation and relay selection for improving energy and spectral efficiency	[157][158]	Non-Linear Programming
	[159][160]	Game Theory
	[161][162]	Particle Swarm/ Genetic Algorithm
	[163][163]	Neural Networks
	[164][165]	Q-learning
	[166] -[168]	Deep Q Network (DQN)
Relay UE (RUE) activation/deactivation	[169][170]	Heuristic

Concerning the problem of relay placement, several solutions have been proposed in the last few years. Specifically, [152] deals with coverage extension by means of D2D relays and proposes a greedy solution to determine the optimal relay locations. Similarly, [153] makes use of different approaches based on Particle Swarm Optimisation for determining the adequate placement of the relays in order to extend coverage at the cell edge in 4G/5G cellular networks. Other works proposed methodologies to obtain the appropriate placement of IAB nodes. In particular, [154] makes use of Genetic-Algorithm-based optimisation. Neural Networks methodologies have also been proposed, for example in [155], in order to optimise the relative position of the relay with respect to a specific BS and a specific user assuming that the path between the BS and the UE is correlated with the path between the BS and the relay. Most of the previously mentioned works deal with the problem of relay placement by using simulation and mathematical models. In turn, in [156] a design of a Digital Twin platform has been developed in order to obtain an accurate modelling of a real propagation that is used for a more adequate relay deployment.

Other works, such as [157] -[162], focused on resource allocation and relay selection to optimize the energy efficiency or the spectral efficiency using different search methods. As an example, [157][158] propose the use of non-linear programming techniques for determining the transmitted power and spectrum allocation to maximise the energy efficiency. Concerning the relay selection problem, [160] presents a methodology based on Game Theory to enhance both energy and spectral efficiency in a multi-hop scenario. In turn, [161] proposes a multi-objective relay selection methodology with the aim to maximise throughput while minimizing the delay and battery power consumption. The proposed methodology in [161] is based on meta-heuristic algorithms such as Particle Swarm Optimisation and Genetic Algorithms. Other works, such as [162], proposed a joint power allocation and relay selection methodology. However, these approaches may lead to a high computational cost. In the last years, recently developed machine learning approaches outperform these previously mentioned optimisation methods in terms of reduced complexity and increased accuracy. As an example, [163] proposed a supervised machine learning technique based on Artificial Neural Networks (ANNs) for relay selection and resource allocation with the aim to optimise the energy and spectral efficiency. As shown in [163], the proposed ANN technique provides better performance than a Particle Swarm Optimisation approach.

Other approaches make use of traditional RL techniques for relay selection. RL paradigm defines an action-state Q-table that represents the expected accumulated reward when taking an action in a certain state. By iteratively interacting with the environment, the RL agent learns which actions lead to better rewards. As an example, [164] makes use of a Q-learning methodology for power allocation and relay selection. Similarly, in [165], Q-learning is used for user relay selection with the objective of minimising the total transmitted power. As shown in [165], Q-learning provides similar results to a Genetic Algorithm while exhibiting much lower computational complexity. The main drawback of traditional RL techniques is that, in general, the number of action-states can be very large, making it computationally intractable to build an action-state Q-table. To address this limitation, a Q-function is usually defined, which achieves the same purpose of mapping an action-state to a Q-value in a large action-state space. Since Neural Networks can provide an accurate modelling of complex non-linear functions, several proposals make use of a Neural Network to estimate this Q-function. This leads to combined approaches of RL with Neural Networks usually called DRL. As an example, several works have proposed the use of a DQN technique for relay selection and power allocation [166] -[168].

While several approaches have been proposed in the context of power and resource allocation and relay selection, very few works can be found in the literature regarding the activation/deactivation of the different relays. Adequate decisions of relay switching on/off by means of AI/ML techniques can contribute to obtain large energy savings in the network while maintaining the service requirements. In this context, a critical functionality is the “relay activation” function. This functionality is in charge of deciding under what type of

conditions a fixed or moving relay or a RUE can be switched off to obtain energy savings. The decisions of activating or deactivating each relay will change the radio environment conditions. For this reason, a RAN reconfiguration functionality will be necessary to perform the reconfigurations needed in the network in order to both obtain energy savings and satisfy the user service requirements.

In the context of UE-to-Network relaying, different RUE activation strategies have been studied in [169] based on different criteria and context information. Results of [169] revealed that the most efficient strategies from the perspective of outage probability reduction are those that account for the number of UEs that would be served by a RUE based on the experienced spectral efficiency. Leveraging the outcomes of this previous work, a functional framework for supporting the RUE activation was presented in [170] based on the characterization of each potential RUE through a utility metric that measures the coverage enhancements brought to the network when the RUE is activated.

6.1.5 Service-aware energy-efficient RU control

The intelligent and dynamic energy-efficient control of RUs is one of the main challenges and opportunities of the O-RAN architecture due to its impact on the global energy consumption of the network [218]. As introduced in Section 2.3.4, different approaches such as switching on/off the cells or enabling advanced sleep modes can be considered to orchestrate RU status according to the network status and the targeted energy efficiency, while AI/ML can be incorporated to enhance decision making through predictions (for example load, mobility, interference...) or inference.

The O-RAN alliance has recently started defining energy savings use cases [93], where intelligent RU management is one of the main targeted optimisations. The report considers carrier and cell on/off and RF channel reconfiguration strategies, describing in both cases a similar approach which mainly requires: collecting data from the O-RU through O1 or OFH-MP (directly or via DU), defining energy-saving targets in the Non-RT RIC (rApp), collecting data from the O-RU through O1 or OFH-MP (directly or via DU), training AI/ML models with the collected data in the non-RT RIC or in the near-RT RIC, deploy and activate AI/ML models in the non-RT RIC or in the near-RT RIC and perform required actions according to the models and the data through A1, O1, E2 or OFH-MP interfaces. The output of the AI/ML models is not specified and may include prediction of future traffic, user mobility, and resource usage, or expected energy efficiency enhancements, resource usage, and network performance for different ES optimization states [93]. The report also describes the data that might be considered for training and inference, and the possible gains of applying these optimisations. The authors in [102] describe a similar approach for controlling cell switch off/on and RF channel switching in order to increase energy efficiency. In the first case, the rApp hosts a RL agent which uses user positions to decide whether to switch on/off actions, using the obtained energy efficiency (throughput vs energy consumption) as a reward. In the case of the RF Channel Switching rAppr RL agent considers parameters such as the number of users per beam, the user throughput or the power consumption, to select one of the possible O-RU's array configurations. The reward is based on power consumption and QoS metrics.

The authors in [219] implement a Q-learning approach to infer the best combination and duration of Advanced Sleep Modes (ASM) according to traffic load and latency requirements. The work considers three ASM levels with different deactivation, duration and activation times, which also lead to different energy consumptions. Depending on energy and latency requirements, which constitute the reward function, the algorithm prioritizes shorter or longer sleep states. In [219], a Q-learning approach is also presented, in this case to join decisions on radio resource adaptation (bandwidth, number of users, and active array size) and advanced sleep modes according to traffic demands. The work in [221] proposes a Deep Neural Network solution to enhance energy efficiency by joining decisions on sleep modes and beamforming weights, which are contradictory: that is, a scenario with less active RUs may need a higher transmission power to serve users, while scenarios with a lower transmission power may need of a higher number of actives RUs. The

objective of the DNN is to derive the optimal set of active RUs and the beamforming weights to minimize energy consumption according to QoS demands.

6.2 BeGREEN proposed developments

This chapter contains optimization strategies that BeGREEN project will use to achieve an AI/ML enhanced RAN for energy-saving purposes. The Energy Measurement is a prerequisite step to evaluate the energy savings achieved.

BeGREEN optimization strategies will be aligned with O-RAN Network Energy Saving use cases, in particular the Carrier and Cell Switch on/off and the O-Cloud Resource Energy Saving Mode scenarios, which identification and definition have started as described in [93]. BeGREEN will implement innovative AI/ML-based strategies through novel rApps empowered by the Intelligent Plane in order to further optimize the energy effienciency of the RAN, including the 5G Core.

In the context of relays, BeGREEN will propose a relay activation function to smartly decide the activation and deactivation of relays by means of AI/ML techniques with the aim to improve the spectral efficiency and reduce the energy consumption.

The RIC framework provided in BeGREEN will be able to support AI/ML xApps to enhance the performance of the RAN by means of the hardware acceleration, while implementing techniques to reduce the energy consumption of this framework.

The BeGREEN consortium common effort towards energy efficient networks include the development of AI-based ET and DPD modules for the O-RU that have the potential to substantially decrease the O-RU power consuption especially in high and medium power amplifiers. Taking into account that the O-RUs are the major power consumers in the cellular networks, the ET and DPD modules will be an essential components in Green Cellular Networks.

7 Summary and Conclusions

BeGREEN's objective is to offer pioneering solutions for enhancing energy efficiency in the RAN. This deliverable provides an overview of the motivation on energy-efficient design choices and general optimization strategies. Additionally, it introduces the main components of the O-RAN framework, which will be used as basis for BeGREEN's proposed optimizations and architecture.

Chapter 2 presented the energy efficient design choices and models and a high-level analysis of the RAN architecture.

Chapter 3 presented the reference use case criteria where the reference scenarios and reference use cases , followed by the definition of the preliminary KPIs.

Heterogeneity in the RAN depending on the described reference scenarios is studied in chapter 3, focusing on comparing macro MIMO environments with CF-mMIMO deployments as well as coverage extension aided by relays and RIS. Finding the sweet-spot for deploying one or the other solution will be part of the work that will be pursued in BeGREEN and presented in deliverable D2.2. In this chapter, relay and RIS technologies to provide coverage extension are described. Finally, BeGREEN's architecture system requirements are mapped onto O-RAN standards, identifying suitable locations in an O-RAN based architecture for the proposed Energy Efficiency and AI/ML components. In there, the central role of the RICs is emphasized. The integration of RIS and relays in BeGREEN's architecture has been considered in detail. The proposed architecture also aims to integrate external components such as the 5G Core and the Edge to allow a joint orchestration of allocated resources.

In chapter 4, various energy saving technologies are introduced considering hardware and software, focusing on BeGREEN's considered technologies and objectives. It is understood that RAN is the major energy consumer in a mobile network, and RAN equipment play a major role in this where RU can be notified as the most significant energy consumer. RU on/off scheme, beam-switching and data-blanking are studied as well as AI-based DPD and ET modules of RU. Hardware acceleration, for DU or CU, is another power-saving strategy that can contribute significantly, especially when computation heavy processing, for example in mMIMO systems, is used. An analysis of the most suitable processes and platforms, for example GPUs, for DU hardware acceleration are introduced, and several strategies to reduce energy consumption in O-CU are discussed. Finally, link quality enhancements by incorporating ISAC concept and relays in the RAN are introduced and the potential to improve energy efficiency in the RAN are discussed. Chapter 5 is finalised by introducing BeGREEN's proposed development on each of the discussed items.

Chapter 5, aims at providing innovative solutions based on the AI/ML-enhanced RAN. AI and ML algorithms will play a key role in optimizing the future systems. This chapter outlines a proposal for the formulation, and assessment of AI/ML-based algorithmic solutions for RAN control at the non-RT RIC and Near-RT RIC of the O-RAN architecture targeting enhanced energy efficiency. This includes carrier and cell off/on switching, relay activation/deactivation, energy-efficient resource orchestration in virtualized RANs with shared computing infrastructure and AI edge services and service-aware energy-efficient RU control.

Chapter 6 sets out some AI/ML-based optimization strategies that address the BeGREEN objectives to introduce AI/ML services for enhancing energy efficiency of the BeGREEN RAN infrastructure.

To sum up, this deliverable will work as input for the rest of the tasks in WP2, as well as for the rest of the WPs. BeGREEN architecture, scenarios and KPIs will be considered along for the evolution of the whole project.

8 Bibliography

[1] European 6G Smart Networks and Services. Available at: https://smart-networks.europa.eu/

[2] Ericsson, "Breaking the Energy Curve," 2020. [Online]. Available: https://www.ericsson.com/4aa14d/assets/local/about-ericsson/sustainability-and-corporate-responsibility/documents/2022/breaking-the-energy-curve-report.pdf

[3] GSM Association, Going green: benchmarking the energy efficiency of mobile, June 2021.

[4] Ericcson, "Radio Systems Solutions - Energy Efficiency," 2022. [Online]. Available: https://www.ericsson.com/en/portfolio/networks/ericsson-radio-system/radio-system-solutions/energy-efficiency

[5] N. Piovesan, D. Lopez-Pérez, A. D. Domenico, X. Geng, H. Bao, and M. Debbah, "Machine Learning and Analytical Power Consumption Models for 5G Base Stations," IEEE Communications Magazine, vol. 60, no. 10, October 2022.

[6] AIMM project https://aimm.celticnext.eu/

[7] AIMM simulator https://github.com/keithbriggs/AIMM-simulator

[8] ACPI Specification 6.5 documentation, https://uefi.org/specs/ACPI/6.5/index.html

[9] O-RAN Alliance "O-RAN Working Group 1 (Use Cases and Overall Architecture); O-RAN Architecture Description".

[10] O-RAN Alliance WG3, "Near-RT RIC Architecture v4", March 2023.

[11] O-RAN-WG2.A1GAP-v02.03, "O-RAN Working Group 2; A1 interface: General Aspects and Principles v02.03".

[12] O-RAN Alliance WG2 "O-RAN Working Group 2; Non-RT RIC Architecture Technical Specification v02.01", Non-RT-RIC-ARCH-TS-v01.00, October 2022.

[13] O-RAN-WG4.MP.0-v07.00, "O-RAN Alliance Working Group 4; Management Plane Specification v07.00".

[14] O-RAN-WG6.CAD-v02.01, "Cloud Architecture and Deployment Scenarios for O-RAN Virtualized RAN v02.01".

[15] 3GPP TR 36.401, "Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Architecture Description".

[16] 3GPP TR 38.300, "NR; NR and NG-RAN Overall Description; Stage 2".

[17] Ranjbar, Vida; Girycki,, Adam; Rahman, Md Arifur; Pollin, Sofie; moonen, marc; Vinogradov, Evgenii (2021): Cell-free mMIMO Support in the O-RAN Architecture: A PHY Layer Perspective for 5G and Beyond Networks. TechRxiv. Preprint. https://doi.org/10.36227/techrxiv.15144516.v2

[18] O-RAN next Generation Research Group (nGRG), https://public.o-ran.org/display/NGRG/Introduction

[19] G. Auer et al., How much energy is required to run a wireless network? IEEE Wireless Communications, October 2011, 40-49.

[20] A. Israr et al., Power consumption analysis of access network in 5G mobile communication infrastructures - an analytical quantification model. Pervasive and Mobile Computing 80 (2022), 1-17.

[21] Mavenir Whitepaper, A Holistic Study of Power Consumption and Energy Savings Strategies for Open vRAN Systems. February 2023.

[22] FierceWireless webinar, Building Sustainable Networks with Energy Efficient Open RAN. Feb 22, 2023.

[23] NGMN Whitepaper, Green Future Networks: sustainability challenges and initiatives in mobile networks v1.0. July 2021.

[24] ATIS' Next G Alliance white paper "Next G Alliance Green G: The Path Toward Sustainable 6G", 2022, https://www.nextgalliance.org/white_papers/green-g-the-path-towards-sustainable-6g/

[25] S. Zhou, A. J. Goldsmith, Z. Niu, "On Optimal Relay Placement and Sleep Control to Improve Energy Efficiency in Cellular Networks", IEEE International Conference on Communications, 2011.

[26] 5G Americas White paper, "The Evolution of Open RAN", February 2023.

[27] R. Fantini, D. Sabella, M. Caretti, "Energy efficiency in LTE-Advanced networks with relay nodes", IEEE VTC in Spring, 2011.

[28] R. Fantini, D. Sabella, M. Caretti, "An E3F based assessment of energy efficiency of Relay Nodes in LTE-Advanced networks", IEEE PIMRC, 2011.

[29] X. Li, H. Wang, X. You, "Dynamic User Association for Energy Minimization in Macro-Relay Network", International Conference on Wireless Communications & Signal Processing (WCSP), 2012.

[30] J. Li , N. Liu, "User association for minimum energy consumption with macro-relay interference", IEEE PIMRC, 2015.

[31] J. Pérez-Romero, O. Sallent, R. Agustí, "Energy Saving Potentials in the Radio Access through Relaying in Future Networks", EUCNC 2014.

[32] 3GPP TR 22.866 v17.1.0, "Enhanced Relays for Energy Efficiency and Extensive Coverage; Stage 1 (Release 17)", December, 2019.

[33] J. Sydir, R. Taori, "An Evolved Cellular System Architecture Incorporating Relay Stations", IEEE Communications Magazine, June 2009.

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

[35] J. Pérez-Romero, O. Sallent, "Leveraging User Equipment for Radio Access Network Augmentation", IEEE Conf. on Standards for Communications and Networking (CSCN'21), December 2021.

[36] V. F. Monteiro et al., "Paving the Way Toward Mobile IAB: Problems, Solutions and Challenges", in IEEE Open Journal of the Communications Society, 2022.

[37] RP-222671 "Mobile IAB (Integrated Access and Backhaul) for NR", 3GPP TSG RAN Meeting #97, Sept. 2022.

[38] 3GPP TR 23.700-05 v18.0.0, "Study on architecture enhancements for vehicle-mounted relays (Release 18)", December, 2022.

[39] Infoantennas Government of Spain https://geoportal.minetur.gob.es/VCTEL/vcne.do

[40] 3GPP, "NR; Multi-connectivity; Overall description; Stage-2," 3GPP Technical Specification (TS) 37.340, v17.3.0, Jan. 2023.

[41] Hexa-X, "D1.2 - Expanded 6G vision, use cases and societal values", Dec. 2021. Accessed: April 6, 2023. [Online] Available: https://ec.europa.eu/research/participants/documents/downloadPublic?documentIds=080166e5dc8b611b&appId=PPGMS.

[42] S. Chen, J. Zhang, J. Zhang, E. Bjornson, B. Ai, "A survey on user-centric cell-free massive MIMO

systems," Digital Communications and Networks, 2022.

[43] G. Interdonato, E. Björnson, H. Q. Ngo, P. Frenger, and E. G. Larsson, "Ubiquitous cell-free massive MIMO Communications," In EURASIP J. Wireless Commun. and Networking, vol. 2019, no. 1, pp. 197-209, 2019a.

[44] J. Zhang, S. Chen, Y. Lin, J. Zheng, B. Ai, and L. Hanzo, "Cell-free massive MIMO: A new next-generation paradigm," In IEEE Access, vol. 7, pp. 99878-99888, 2019.

[45] V. Jungnickel, K. Manolakis, W. Zirwas, B. Panzner, V. Braun, M. Lossow, M. Sternad, R. Apelfröjd, and T. Svensson, "The role of small cells, coordinated multipoint, and massive MIMO in 5G," In IEEE Commun. Mag., vol. 52, no. 5, pp. 44-51, 2014.

[46] E. Björnson, J. Hoydis, and L. Sanguinetti, "Massive MIMO has unlimited capacity," In IEEE Trans. Wireless Commun., vol. 17, no. 1, pp. 574-590, 2018.

[47] S. Buzzi, C. D'Andrea, A. Zappone, and C. D'Elia, "User-centric 5G cellular networks: Resource allocation and comparison with the cell-free massive MIMO approach," In *IEEE Trans. Wireless Commun.*, vol. 19, no. 2, pp 1250–1264, 2019.

[48] E. Nayebi, A. Ashikhmin, T. L. Marzetta, H. Yang, and B. D. Rao, "Precoding and power optimization in cell-free massive MIMO systems," In *IEEE Trans. Wireless Commun.*, vol. 16, no. 7, pp. 4445–4459, 2017.

[49] G. J. Foschini, K. Karakayali, and R. A. Valenzuela, "Coordinating multiple antenna cellular networks to achieve enormous spectral efficiency," In IEEE Proceedings Communications, vol. 153, no. 4, pp. 548-555, 2006.

[50] D. Gesbert, S. Hanly, H. Huang, S. S. Shamai, O. Simeone, W. and Yu, "Multi-cell MIMO cooperative networks: A new look at interference," In IEEE J. Sel. Areas Commun., vol. 28, no. 9, pp.1380–1408, 2010.

[51] X. Hong, Y. Jie, C.-X. Wang, J. Shi, and X. Ge, "Energy-spectral efficiency trade-off in virtual MIMO cellular systems," In IEEE J. Sel. Areas Commun., vol. 31, no. 10, pp. 2128-2140, Oct. 2013.

[52] J. Zhang, E. Björnson, M. Matthaiou, D. W. K. Ng, H. Yang, and D. J. Love, "Prospective Multiple Antenna Technologies for Beyond 5G," In IEEE Journal on Selected Areas in Communications, vol. 38, no. 8, pp. 1637-1660, Aug. 2020.

[53] Samsung C-Band Network Solutions Portfolio

[54] Pan, Cunhua, et al. "Reconfigurable intelligent surfaces for 6G systems: Principles, applications, and research directions." IEEE Communications Magazine 59.6 (2021): 14-20.

[55] Zhaohui Yang et al. 2022. Energy-Efficient Wireless Communications With Distributed Reconfigurable Intelligent Surfaces. IEEE Transactions on Wireless Communications 21, 1 (2022), 665–679.

[56] Li Yezhen et al. 2020. A Novel 28 GHz Phased Array Antenna for 5G Mobile Communications. ZTE Communications 18, 3 (2020), 20–25.

[57] Georgios Trichopoulos et al. 2021. Design and Evaluation of Reconfigurable Intelligent Surfaces in Real-World Environment. arXiv preprint arXiv:2109.07763 (2021).

[58] Linglong Dai et al. 2020. Reconfigurable intelligent surface-based wireless communications: Antenna design, prototyping, and experimental results. IEEE Access 8 (2020), 45913–45923.

[59] Jingzhi Hu et al. 2020. Reconfigurable Intelligent Surface Based RF Sensing: Design, Optimization, and Implementation. IEEE Journal on Selected Areas in Communications 38, 11 (2020), 2700–2716.

[60] Romain Fara et al. 2021. A Prototype of Reconfigurable Intelligent Surface with Continuous Control of the Reflection Phase. arXiv preprint arXiv:2105.11862 (2021).

[61] Xin Tan, Zhi Sun, Dimitrios Koutsonikolas, and Josep M Jornet. 2018. Enabling indoor mobile millimeter-wave networks based on smart reflect-arrays. In IEEE INFOCOM 2018-IEEE Conference on Computer Communications. IEEE, 270–278.

[62] Venkat Arun and Hari Balakrishnan. 2020. RFocus: Beamforming using thousands of passive antennas. In USENIX NSDI.

[63] Manideep Dunna et al. 2020. ScatterMIMO: Enabling virtual MIMO with smart surfaces. In ACM MobiCom.

[64] Q. Sun, N. Li, C. -L. I, J. Huang, X. Xu and Y. Xie, "Intelligent RAN Automation for 5G and Beyond," in IEEE Wireless Communications, doi: 10.1109/MWC.014.2200271.

[65] H. Lee, J. Cha, D. Kwon, M. Jeong and I. Park, "Hosting AI/ML Workflows on O-RAN RIC Platform," 2020 IEEE Globecom Workshops (GC Wkshps, Taipei, Taiwan, 2020, pp. 1-6, doi: 10.1109/GCWkshps50303.2020.9367572.

[66] O-RAN- 3GPP TR 38.913 version 17.0.0 Release 17O-RAN-WG6.CAD-v02.01, “Cloud Architecture and Deployment Scenarios for O-RAN Virtualized RAN v02.01”.

[67] “SNS-JU BeGREEN Project,” 2016. [Online]. Available: ww.sns-begreen.com

[68] Samsung, “Virtualized Radio Access Network: Architecture, Key technologies and Benefits.” Technical Report, 2019, Link.

[69] A. Garcia-Saavedra, X. Costa-Perez, “O-RAN: Disrupting the virtualized ran ecosystem,” IEEE Communications Standards Magazine, 2021.

[70] Cisco, Rakuten, Altiostar, “Reimagining the End-to-End Mobile Network in the 5G Era,” White Paper, 2019.

[71] G. Garcia-Aviles et al., “Nuberu: Reliable RAN virtualization in shared platforms,” in Proceedings of the 27th MobiCom, 2021, pp. 749–761.

[72] Heavy Reading, “5G Transport: A 2021 Heavy Reading Survey,” White Paper, Feb. 2022.

[73] NTT DOCOMO. (2013) Docomo to develop next-generation BSs utilizing advanced c-ran architecture for lte-advanced. Link.

[74] Ericsson. (2021) Exploring new centralized ran and fronthaul opportunities. Link.

[75] AT&T. (2022, Feb.) Cloudifying 5G with an Elastic RAN. [Online]. Available : https://about.att.com/innovationblog/2022/cloudifying-5g-with-elastic-ran.html

[76] O-RAN Alliance, “Cloud Architecture and Deployment Scenarios for O-RAN Virtualized RAN (O-RAN.WG6.CADS-v04.00) ,” Technical Report, Oct. 2022.

[77] A. Tootoonchian et al., “ResQ: Enabling SLOs in Network Function Virtualization,” in Proceedings of the 15th USENIX NSDI, 2018, pp. 283–297.

[78] A. Manousis et al., “Contention-aware performance prediction for virtualized network functions,” in Proceedings of the ACM SIGCOMM,2020, pp. 270–282.

[79] C. Sun et al., “NFP: Enabling network function parallelism in NFV,” in Proceedings of the ACM SIGCOMM, 2017, pp. 43–56.

[80] P. Kumar et al., “PicNIC: predictable virtualized NIC,” in Proceedings of the ACM SIGCOMM, 2019, pp.

351–366.

[81] J. Gong et al., “Microscope: Queue-based performance diagnosis for network functions,” in Proceedings of the ACM SIGCOMM, 2020, pp. 390–403.

[82] Microsoft, “Seeing AI,” [Online]. Available: https://www.microsoft.com/en-us/ai/seeing-ai.

[83] Chatzopoulos, Dimitris, et al. “Mobile augmented reality survey: From where we are to where we go.” Ieee Access 5 (2017): 6917-6950.

[84] Taigman, Yaniv, et al. "Deepface: Closing the gap to human-level performance in face verification." Proceedings of the IEEE conference on computer vision and pattern recognition. 2014.

[85] Garcia-Saavedra, Andres, et al. "Joint optimization of edge computing architectures and radio access networks." IEEE Journal on Selected Areas in Communications 36.11 (2018): 2433-2443.

[86] Lu, Donna. "Creating an AI can be five times worse for the planet than a car." New Scientist 6 (2019). [Online]. Available: https://www.newscientist.com/article/2205779-creating-an-ai-can-be-five-times-worse-for-the-planet-than-a-car/

[87] GSMA - Future Networks, “Energy efficiency: An overview,” 2019. [Online]. Available: https://www.gsma.com/futurenetworks/wiki/energy-efficiency-2

[88] 3GPP TR 38.913 version 17.0.0 Release 17 “Study on scenarios and requirements for next-generation access technologies”

[89] D. Lake, N. Wang, R. Tafazolli and L. Samuel, “Softwarization of 5G Networks–Implications to Open Platforms and Standardizations,” in IEEE Access, vol. 9, pp. 88902-88930, 2021, doi: 10.1109/ACCESS.2021.3071649.

[90] 5G IA, “European Vision for the 6G Network Ecosystem”, White Paper, version 1.0, June 2021

[91] Erik Ekudden, “Energy-efficient packet processing in 5G mobile systems”, Ericsson blog post. Accessed in March 2023. Available: https://www.ericsson.com/en/reports-and-papers/ericsson-technology-review/articles/energy-efficient-packet-processing-in-5g-mobile-systems

[92] Telecom Infra Project, Open RAN MoU Group, “Open RAN Technical Priority “Release 2” Document: Focus on Energy Efficiency”, March 2022

[93] O-RAN Alliance WG1, “O-RAN Network Energy Savings Use Cases Technical Report 1.0”, March 2023

[94] L. M. P. Larsen, H. L. Christiansen, S. Ruepp and M. S. Berger, "Toward Greener 5G and Beyond Radio Access Networks—A Survey," in IEEE Open Journal of the Communications Society, vol. 4, pp. 768-797, 2023, doi: 10.1109/OJCOMS.2023.3257889.

[95] 3GPP TS 28.310 v18.0.0, “Energy efficiency of 5G (Release 18)”, December 2022

[96] Nielsen, Lars, Gavras, Anastasius, Dieudonne, Michael, Mesogiti, Ioanna, Roosipuu, Priit, Houatra, Drissa, & Kosmatos, Evangelos, “Beyond 5G/6G KPIs and Target Values”, Version 1.0, 2022, Zenodo, https://doi.org/10.5281/zenodo.6577506

[97] 3GPP TS 28.554 v18.0.0, “5G end to end Key Performance Indicators (KPI)”, December 2022

[98] Capgemini engineering and Intel, “Project Bose: A smart way to enable sustainable 5G networks”, White paper, July 2022.

[99] 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, doi: 10.1109/COMST.2023.3239220.

[100] O-RAN Alliance WG1, “Uses Cases Analysis report” v10, March 2023

[101] D. Mavrakis, M. Saadi, “Is Open vRAN Power Efficient?”, Abi Research, February 2023

[102] Rimedo Labs, “The O-ran Whitepaper 2023: Energy Efficiency in O-RAN”, Rimedo, February 2023

[103] O-RAN Alliance WG2, “O-RAN AI/ML workflow description and requirements 1.03”, October 2021

[104] O-RAN Alliance WG2, “O-RAN R1 interface: General Aspects and Principles 3.0”, October 2022

[105] O-RAN Alliance WG3, “E2 Service Model (E2SM) v3”, March 2023

[106] ETSI, “ETSI GS MEC 012 V2.2.1 (2022-02): Multi-access Edge Computing (MEC); Radio Network Information API”, February 2022

[107] O-RAN Alliance, “Near-Real-time RAN Intelligent Controller,Use Cases and Requirements v3”, March 2023

[108] Leonardo Bonati and Michele Polese and Salvatore D’Oro and Stefano Basagni and Tommaso Melodia, “OpenRAN Gym: AI/ML development, data collection, and testing for O-RAN on PAWR platforms”, Computer Networks, Volume 220, https://doi.org/10.1016/j.comnet.2022.109502

[109] Strinati, Emilio Calvanese, et al. “Reconfigurable, intelligent, and sustainable wireless environments for 6G smart connectivity.” *IEEE Communications Magazine* 59.10 (2021): 99-105.

[110] 3GPP, “Feasibility Study on Integrated Sensing and Communication,” TSG SA, Technical Report 22.837, 2023, Version 1.0.0 (Rel. 19).

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

[112] Bendlin, Ralf, et al. "From homogeneous to heterogeneous networks: A 3GPP long term evolution rel. 8/9 case study." *2011 45th Annual Conference on Information Sciences and Systems*. IEEE, 2011.

[113] Torres, Rafael P., Jesús R. Pérez, and Luis Valle. "Channel Hardening: A Comparison between Concentrated and Distributed Massive MIMO." *IEEE Antennas and Wireless Propagation Letters* (2023).

[114] Jang, Uk, et al. "CoMP-CSB for ICI nulling with user selection." *IEEE Transactions on Wireless Communications* 10.9 (2011): 2982-2993.

[115] Li, Chang, Jun Zhang, and Khaled B. Letaief. "Throughput and energy efficiency analysis of small cell networks with multi-antenna BSs ." *IEEE Transactions on Wireless Communications* 13.5 (2014): 2505-2517.

[116] Sawahashi, Mamoru, et al. "Coordinated multipoint transmission/reception techniques for LTE-advanced [Coordinated and Distributed MIMO]." *IEEE Wireless Communications* 17.3 (2010): 26-34.

[117] Liu, Zhiyang, and Lin Dai. "A comparative study of downlink MIMO cellular networks with co-located and distributed BS antennas." *IEEE Transactions on Wireless Communications* 13.11 (2014): 6259-6274.

[118] Liu, Wenjia, Shengqian Han, and Chenyang Yang. "Energy efficiency comparison of massive MIMO and small cell network." *2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE, 2014.

[119] Björnson, Emil, Luca Sanguinetti, and Marios Kountouris. "Deploying dense networks for maximal energy efficiency: Small cells meet massive MIMO." *IEEE Journal on Selected Areas in Communications* 34.4 (2016): 832-847.

[120] Wesemann, S., Du, J., & Viswanathan, H. (2023). Energy Efficient Design of Extreme Massive MIMO. *arXiv preprint arXiv:2301.01119.*

[121] Ngo, H. Q., Ashikhmin, A., Yang, H., Larsson, E. G., & Marzetta, T. L. (2017). Cell-free massive MIMO

versus small cells. *IEEE Transactions on Wireless Communications*, *16*(3), 1834-1850.

[122] Demir, Özlem Tugfe, Emil Björnson, and Luca Sanguinetti. "Foundations of user-centric cell-free massive MIMO." *Foundations and Trends® in Signal Processing* 14.3-4 (2021): 162-472.

[123] 3GPP TS 36.216 V10.3.1, “Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer for relaying operation (Release 10)”, September, 2011.

[124] 3GPP TS 38.300 V17.3.0, “NR and NG-RAN Overall Description; Stage 2 (Release 17)”, December, 2022.

[125] M. Polese et al., "Integrated Access and Backhaul in 5G mmWave Networks: Potential and Challenges", IEEE Communications Magazine, March 2020.

[126] C. Madapatha et al., "On Integrated Access and Backhaul Networks: Current Status and Potentials", IEEE Open Journal of the Communications Society, 2020.

[127] 3GPP TS 38.401 V17.3.0, “NG-RAN; Architecture description (Release 17)”, December, 2022.

[128] 3GPP TR 22.839 v18.1.0, “Study on Vehicle-Mounted Relays; Stage 1 (Release 18)”, December, 2021.

[129] 3GPP TS 23.304 v18.0.0, "Proximity based Services (ProSe) in the 5G System (5GS) (Release 18)", December, 2022.

[130] B. Saleh, Ö. Bulakci, S. Redana, B. Raaf, J. Hämäläinen, “Evaluating the Energy Efficiency of LTE-Advanced Relay and Picocell Deployments”, IEEE WCNC Conference, 2012.

[131] C. V. Anamuro, N. Varsier, J. Schwoerer and X. Lagrange, "Simple modeling of energy consumption for D2D relay mechanism," 2018 IEEE Wireless Communications and Networking Conference Workshops (WCNCW), Barcelona, Spain, 2018, pp. 231-236

[132] Y. Sui, A. Papadogiannis, W. Yang and T. Svensson, "The Energy Efficiency Potential of Moving and Fixed Relays for Vehicular Users", 2013 IEEE 78th Vehicular Technology Conference (VTC Fall), Las Vegas, NV, USA, 2013, pp. 1-7.

[133] A. Prasad, M. A. Uusitalo and A. Maeder, "Energy Efficient Coordinated Self-Backhauling for Ultra-Dense 5G Networks," 2017 IEEE 85th Vehicular Technology Conference (VTC Spring), Sydney, NSW, Australia, 2017.

[134] D. Korpi, T. Riihonen, A. Sabharwal and M. Valkama, "Transmit Power Optimization and Feasibility Analysis of Self-Backhauling Full-Duplex Radio Access Systems," in IEEE Transactions on Wireless Communications, vol. 17, no. 6, pp. 4219-4236, June 2018.

[135] Höyhtyä, M.; Apilo, O.; Lasanen, M. “Review of Latest Advances in 3GPP Standardization: D2D Communication in 5G Systems and Its Energy Consumption Models”. Future Internet 2018, 10, 3.

[136] A. F. E. Mohammed, M. H. Altayeb and N. I. Osman, "Evaluating the Energy Efficiency of UE-to-Network Relay Assisted Device-to-Device Communication," 2020 International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE), Khartoum, Sudan, 2021, pp. 1-5.

[137] 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.

[138] W. Lv, Y. Zeng, T. Song, T. Xu and H. Hu, "Stable and Proportional Fair User Pairing Algorithm for D2D-Relay Systems," 2018 IEEE Global Communications Conference (GLOBECOM), Abu Dhabi, United Arab Emirates, 2018, pp. 1-6.

[139] S. Zhang, X. Xu, M. Sun, X. Tang and X. Tao, "Energy efficient uplink transmission for UE-to network

relay in heterogeneous networks," 2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), Montreal, QC, Canada, 2017.

[140] C. Huang et al., "Reconfigurable intelligent surfaces for energy efficiency in wireless communication," IEEE Trans. Wireless Commun., vol. 18, no. 8, pp. 4157–4170, Aug. 2019.

[141] Strinati, E. C., Alexandropoulos, G. C., Wymeersch, H., Denis, B., Sciancalepore, V., D'Errico, R., ... & Popovski, P. (2021). Reconfigurable, intelligent, and sustainable wireless environments for 6G smart connectivity. *IEEE Communications Magazine*, *59*(10), 99-105.

[142] O-RAN Alliance. "O-RAN Architecture Description 8.0" Technical Report. March 2023 O-RAN.WG1.OAD-R003-v08.00

[143] Taha, Abdelrahman, Muhammad Alrabeiah, and Ahmed Alkhateeb. "Enabling large intelligent surfaces with compressive sensing and deep learning." *IEEE access* 9 (2021): 44304-44321.

[144] Aygül, Mehmet Ali, Mahmoud Nazzal, and Hüseyin Arslan. "Deep learning-based optimal RIS interaction exploiting previously sampled channel correlations." *2021 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2021.

[145] Huang, Chongwen, et al. "Indoor signal focusing with deep learning designed reconfigurable intelligent surfaces." *2019 IEEE 20th international workshop on signal processing advances in wireless communications (SPAWC)*. IEEE, 2019.

[146] Gao, Jiabao, et al. "Unsupervised learning for passive beamforming." *IEEE communications letters* 24.5 (2020): 1052-1056.

[147] T. Sloane, et al., "SMaRT 5G Conceptual Overview: Sustainable Mobile and RAN Transformation (SMaRT) project", ONF, October 2022

[148] J. Sydir et al., "DPM-NFV: Dynamic Power Management Framework for 5G User Plane Function using Bayesian Optimization," GLOBECOM 2022 - 2022 IEEE Global Communications Conference, Rio de Janeiro, Brazil, 2022, pp. 4099-4105, doi: 10.1109/GLOBECOM48099.2022.10001394.

[149] P. Veitch, C. Macnamara and J. J. Browne, "Balancing NFV Performance and Energy Efficiency," 2022 25th Conference on Innovation in Clouds, Internet and Networks (ICIN), Paris, France, 2022, pp. 71-75, doi: 10.1109/ICIN53892.2022.9758133.

[150] 3GPP TS 28313, v17.8.0, "Management and orchestration, Self-Organizing Networks (SON) for 5G networks", (Release 17), March 2023.

[151] 3GPP TS 28310, v18.1.0, "Management and orchestration, Energy Efficiency of 5G", (Release 18), March 2023.

[152] W. Qu, G. Li and Y. Zhao, "On the Coverage Problem in Device-to-Device Relay Networks," in *IEEE Communications Letters*, vol. 23, no. 11, pp. 2139-2143, Nov. 2019, doi: 10.1109/LCOMM.2019.2931543.

[153] R. M. Mokhtar, H. M. Abdel-Atty and K. R. Mahmoud, "Optimization of the Deployment of Relay Nodes in Cellular Networks," in *IEEE Access*, vol. 8, pp. 136605-136616, 2020, doi: 10.1109/ACCESS.2020.3011472.

[154] C. Madapatha, B. Makki, A. Muhammad, E. Dahlman, M. -S. Alouini and T. Svensson, "On Topology Optimization and Routing in Integrated Access and Backhaul Networks: A Genetic Algorithm-Based Approach," in *IEEE Open Journal of the Communications Society*, vol. 2, pp. 2273-2291, 2021, doi: 10.1109/OJCOMS.2021.3114669.

[155] H. Amiriara, M. R. Zahabi and V. Meghdadi, "Power-Location Optimization for Cooperative Nomadic Relay Systems Using Machine Learning Approach," in *IEEE Access*, vol. 9, pp. 74246-74257, 2021, doi: 10.1109/ACCESS.2021.3079171.

[156] K. Tokugawa *et al.*, "Design of mmW Digital Twin platform toward B5G/6G – High-Precision

Measurement System and Relay Station Deployment–," *2022 IEEE 33rd Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, Kyoto, Japan, 2022, pp. 813-818, doi: 10.1109/PIMRC54779.2022.9977450.

[157] D. Bilibashi, E. M. Vitucci, V. Degli-Esposti, and A. Giorgetti, "An energy-efficient unselfish spectrum leasing scheme for cognitive radio networks" Sensors, vol. 21, 2021, Art. no. 6161.

[158] Q. Wang and F. Zhao, "Joint Spectrum and Power Allocation for NOMA Enhanced Relaying Networks" in *IEEE Access*, vol. 7, pp. 27008-27016, 2019, doi: 10.1109/ACCESS.2019.2900225.

[159] A. Khodmi, S. B. Rejeb, N. Agoulmine and Z. Choukair, "Joint User-Channel Assignment and Power Allocation For Non-Orthogonal Multiple Access In A 5G Heterogeneous Ultra-Dense Networks" *2020 International Wireless Communications and Mobile Computing (IWCMC)*, Limassol, Cyprus, 2020, pp. 1879-1884, doi: 10.1109/IWCMC48107.2020.9148116.

[160] S. Sawsan and R. Bouallègue, "Energy and spectral efficient relay selection and resource allocation in mobile multi-hop device to device communications," IET Commun., vol. 15, no. 14, pp. 1791–1807, 2021.

[161] S. A. Alvi et al., "QoS-oriented optimal relay selection in cognitive radio networks," Wireless. Commun. Mobile Comput., vol. 2021, 2021, Art. no. 5580963

[162] C. E. Garcia, M. R. Camana, and I. Koo, "Relay selection and power allocation for secrecy sum rate maximization in underlying cognitive radio with cooperative relaying NOMA," Neurocomputing, vol. 452, no. 10, pp. 756–767, Sep. 2021.

[163] B. S. Khan, S. Jangsher, N. Hussain and M. A. Arafah, "Artificial Neural Network-Based Joint Mobile Relay Selection and Resource Allocation for Cooperative Communication in Heterogeneous Network," in *IEEE Systems Journal*, vol. 16, no. 4, pp. 5809-5820, Dec. 2022, doi: 10.1109/JSYST.2022.3179351.

[164] X. Wang, T. Jin, L. Hu and Z. Qian, "Energy-Efficient Power Allocation and Q-Learning-Based Relay Selection for Relay-Aided D2D Communication," in *IEEE Transactions on Vehicular Technology*, vol. 69, no. 6, pp. 6452-6462, June 2020, doi: 10.1109/TVT.2020.2985873.

[165] J. Pérez-Romero, J. Sánchez-González, R. Agustí, B. Lorenzo and S. Glisic, "Power-Efficient Resource Allocation in a Heterogeneous Network With Cellular and D2D Capabilities," in *IEEE Transactions on Vehicular Technology*, vol. 65, no. 11, pp. 9272-9286, Nov. 2016, doi: 10.1109/TVT.2016.2517700.

[166] H. Zhang, S. Chong, X. Zhang and N. Lin, "A Deep Reinforcement Learning Based D2D Relay Selection and Power Level Allocation in mmWave Vehicular Networks," in *IEEE Wireless Communications Letters*, vol. 9, no. 3, pp. 416-419, March 2020, doi: 10.1109/LWC.2019.2958814.

[167] A. Abdelreheem, O. A. Omer, H. Esmaiel and U. S. Mohamed, "Deep Learning-Based Relay Selection In D2D Millimeter Wave Communications," *2019 International Conference on Computer and Information Sciences (ICCIS)*, 2019, pp. 1-5, doi: 10.1109/ICCISci.2019.8716458.

[168] H. Kim, T. Fujii and K. Umebayashi, "Relay Nodes Selection Using Reinforcement Learning," *2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, 2021, pp. 329-334, doi: 10.1109/ICAIIC51459.2021.9415208.

[169] *J. Pérez-Romero, O. Sallent, "*On the Value of Context Awareness for Relay Activation in Beyond 5G Radio Access Networks" IEEE 95th Vehicular Technology Conference (VTC2022-Spring), Helsinki (Finland), June, 2022.

[170] *J. Pérez-Romero, O. Sallent, O. Ruiz, "*On Relay User Equipment Activation in Beyond 5G Radio Access Networks" 96th Vehicular Technology Conference (VTC2022 Fall), London(UK) / Beijing (China), September, 2022

[171] Bega, Dario, et al. "DeepCog: Cognitive network management in sliced 5G networks with deep learning." *IEEE INFOCOM 2019-IEEE conference on computer communications*. IEEE, 2019.

[172] Polese, Michele, Francesco Restuccia, and Tommaso Melodia. "DeepBeam: Deep waveform learning for coordination-free beam management in mmWave networks." Proceedings of the Twenty-second

International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing. 2021.

[173] Ayala-Romero, Jose A., et al. "vrAIn: Deep learning based orchestration for computing and radio resources in vRANs." IEEE Transactions on Mobile Computing 21.7 (2020): 2652-2670.

[174] Ayala-Romero, Jose A., et al. "Orchestrating energy-efficient vrans: Bayesian learning and experimental results." *IEEE Transactions on Mobile Computing* (2021).

[175] Singh, Rajkarn, et al. "Energy-efficient orchestration of metro-scale 5g radio access networks." IEEE INFOCOM 2021-IEEE Conference on Computer Communications. IEEE, 2021.

[176] Chatterjee, Shubhajeet, Mohammad J. Abdel-Rahman, and Allen B. MacKenzie. "On Optimal Orchestration of Virtualized Cellular Networks With Statistical Multiplexing." IEEE Transactions on Wireless Communications 21.1 (2021): 310-325.

[177] Morais, Fernando Zanferrari, et al. "PlaceRAN: Optimal placement of virtualized network functions in the next-generation radio access networks." arXiv preprint arXiv:2102.13192 (2021).

[178] Matoussi, Salma, et al. "5G RAN: Functional split orchestration optimization." IEEE Journal on Selected Areas in Communications 38.7 (2020): 1448-1463.

[179] Baranda, Jorge, et al. "On the Integration of AI/ML-based scaling operations in the 5Growth platform." 2020 IEEE Conference on Network Function Virtualization and Software Defined Networks (NFV-SDN). IEEE, 2020.

[180] Baranda, Jorge, et al. "AIML-as-a-Service for SLA management of a Digital Twin Virtual Network Service." IEEE INFOCOM 2021-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS). IEEE, 2021.

[181] Li, Xi, et al. "5Growth: An end-to-end service platform for automated deployment and management of vertical services over 5G networks." IEEE Communications Magazine 59.3 (2021): 84-90.

[182] Salem, Tareq Si, et al. "Towards inference delivery networks: Distributing machine learning with optimality guarantees." 2021 19th Mediterranean Communication and Computer Networking Conference (MedComNet). IEEE, 2021.Ericsson, "Breaking the Energy Curve," 2020. [Online]. Available: https://www.ericsson.com/4aa14d/assets/local/about-ericsson/sustainability-and-corporate-responsibility/documents/2022/breaking-the-energy-curve-report.pdf.

[183] O-RAN Alliance, *"Juniper RIC Energy Savings Use Cases"*, Feb 2023, https://www.virtualexhibition.o-ran.org/classic/generation/2023/category/intelligent-ran-control-demonstrations/sub/intelligent-control/284.

[184] O-RAN Alliance, *"Dynamic power management to achieve energy savings in multi-vendor Open RAN systems",* Feb 2023, https://www.virtualexhibition.o-ran.org/classic/generation/2023/category/open-ran-demonstrations/sub/open-interface/280

[185] O-RAN Alliance, *"AI Enabled Energy Savings in Open RAN",* Feb 2023, https://www.virtualexhibition.o-ran.org/classic/generation/2023/category/intelligent-ran-control-demonstrations/sub/intelligent-control/256

[186] M. Hoffmann, M. Dryjanski, "*The O-RAN Whitepaper 2023 – Energy Efficiency in O-RAN*", Whitepaper, Rimedo Labs, Feb 2023

[187] O-RAN Alliance, *"5G O-RAN intelligent control for energy savings and 5G O-RAN gNB performance",* Feb 2022, https://www.virtualexhibition.o-ran.org/classic/generation/2022/category/intelligent-ran-control-demonstrations/sub/intelligent-control/167

[188] VMWare, "*Reduce power consumption with VMware Energy Savings rApp*", Sep 2022, https://vmwaretv.vmware.com/media/t/1_rcq6r4qg

[189] NGMN, Green Future Networks: Network Energy Efficiency, December 2021.

[190] Green Mobile Network: Energy Saving Efforts by SK Telecom and NTT DOCOMO, February 2023.

[191] ITU-R, "Minimum requirements related to technical performance for IMT2020 radio interface(s)," Int. Telecommun. Unit, Tech. Rep. M.2410-0, Nov. 2017.

[192] Green Mobile Network: Energy Saving Efforts by SK Telecom and NTT DOCOMO, February 2023.

[193] Remedo Labs White Paper, O-RAN Network Energy Saving: RF Channel Switching, Marcin Hoffmann Feb. 2023.

[194] O-RAN Specifications, User and Synchronization Plane Specification, O-RAN.WG4.CUS.0-R003-v11.00.

[195] O-RAN Specification, O-RAN Working Group 4 Management Plane Specification,O-RAN.WG4.MP.0-R003-v11.00.

[196] Ayala-Romero, Jose A., et al. "vrAIn: A deep learning approach tailoring computing and radio resources in virtualized RANs." *The 25th Annual International Conference on Mobile Computing and Networking*. 2019.

[197] Garcia-Aviles, Gines, et al. "Nuberu: Reliable RAN virtualization in shared platforms." *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*. 2021.

[198] Ayala, Jose A., et al. "Bayesian Online Learning for Energy-Aware Resource Orchestration in Virtualized RANs." *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*. IEEE, 2021.

[199] Foukas, Xenofon, and Bozidar Radunovic. "Concordia: Teaching the 5G vRAN to share compute." *Proceedings of the 2021 ACM SIGCOMM 2021 Conference*. 2021.

[200] D'Oro, Salvatore, et al. "Orchestran: Network automation through orchestrated intelligence in the open RAN." *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*. IEEE, 2022.

[201] Polese, Michele, et al. "ColO-RAN: Developing machine learning-based xApps for open RAN closed-loop control on programmable experimental platforms." *IEEE Transactions on Mobile Computing* (2022).

[202] Zhang, Wenxiao, Bo Han, and Pan Hui. "Jaguar: Low latency mobile augmented reality with flexible tracking." Proceedings of the 26th ACM international conference on Multimedia. 2018.

[203] Jain, Puneet, Justin Manweiler, and Romit Roy Choudhury. "Overlay: Practical mobile augmented reality." Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services. 2015.

[204] He, Zhaoliang, et al. "Adaptive compression for online computer vision: an edge reinforcement learning approach." ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 17.4 (2021): 1-23.

[205] Lu, Guo, et al. "Deep learning for visual data compression." Proceedings of the 29th ACM International Conference on Multimedia. 2021.

[206] Galanopoulos, Apostolos, et al. "Measurement-driven analysis of an edge-assisted object recognition system." ICC 2020-2020 IEEE International Conference on Communications (ICC). IEEE, 2020.

[207] Li, Hongshan, et al. "Jalad: Joint accuracy-and latency-aware deep structure decoupling for edge-cloud execution." 2018 IEEE 24th international conference on parallel and distributed systems (ICPADS). IEEE, 2018.

[208] Ran, Xukan, et al. "Deepdecision: A mobile deep learning framework for edge video analytics." IEEE INFOCOM 2018-IEEE conference on computer communications. IEEE, 2018.

[209] Jiang, Junchen, et al. "Chameleon: scalable adaptation of video analytics." Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication. 2018.

[210] Hung, Chien-Chun, et al. "Videoedge: Processing camera streams using hierarchical clusters." 2018 IEEE/ACM Symposium on Edge Computing (SEC). IEEE, 2018.

[211] Liu, Qiang, and Tao Han. "Dare: Dynamic adaptive mobile augmented reality with edge computing." 2018 IEEE 26th International Conference on Network Protocols (ICNP). IEEE, 2018.

[212] Yang, Peng, et al. "Edge coordinated query configuration for low-latency and accurate video analytics." IEEE Transactions on Industrial Informatics 16.7 (2019): 4855-4864.

[213] Li, Yongbo, et al. "Mobiqor: Pushing the envelope of mobile edge computing via quality-of-result optimization." 2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS). IEEE, 2017.

[214] Alyamkin, Sergei, et al. "Low-power computer vision: Status, challenges, and opportunities." IEEE Journal on Emerging and Selected Topics in Circuits and Systems 9.2 (2019): 411-421.

[215] Goel, Abhinav, et al. "A survey of methods for low-power deep learning and computer vision." 2020 IEEE 6th World Forum on Internet of Things (WF-IoT). IEEE, 2020.

[216] "O-RAN Work Group 1 (Use Cases and Overall Architecture) Network Energy Saving Use Cases Technical Report", March 2023

[217] https://developer.nvidia.com/aerial-sdk

[218] L. M. P. Larsen, H. L. Christiansen, S. Ruepp and M. S. Berger, "Toward Greener 5G and Beyond Radio Access Networks—A Survey," in IEEE Open Journal of the Communications Society, vol. 4, pp. 768-797, 2023, doi: 10.1109/OJCOMS.2023.3257889.

[219] F. E. Salem, T. Chahed, Z. Altman and A. Gati, "Traffic-aware Advanced Sleep Modes management in 5G networks," 2019 IEEE Wireless Communications and Networking Conference (WCNC), Marrakesh, Morocco, 2019, pp. 1-6, doi: 10.1109/WCNC.2019.8886051.

[220] S. K. G. Peesapati, M. Olsson, M. Masoudi, S. Andersson and C. Cavdar, "Q-learning based Radio Resource Adaptation for Improved Energy Performance of 5G Base Stations," 2021 IEEE 32nd Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), Helsinki, Finland, 2021, pp. 979-984, doi: 10.1109/PIMRC50174.2021.9569420.

[221] G. Du, L. Wang, Q. Liao and H. Hu, "Deep Neural Network Based Cell Sleeping Control and Beamforming Optimization in Cloud-RAN," 2019 IEEE 90th Vehicular Technology Conference (VTC2019-Fall), Honolulu, HI, USA, 2019, pp. 1-5, doi: 10.1109/VTCFall.2019.8891410.

[222] Moro, Eugenio, et al. "Planning Mm-Wave access networks with reconfigurable intelligent surfaces." *2021 IEEE 32nd Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*. IEEE, 2021.

[223] *OPEN RAN TECHNICAL PRIORITY DOCUMENT – ENERGY EFFICIENCY by Deutsche Telekom, Orange, Telefónica, TIM and Vodafone, 2023*

[224] Polese, Michele, et al. "Understanding O-RAN: Architecture, interfaces, algorithms, security, and research challenges." IEEE Communications Surveys & Tutorials (2023).

[225] F. Liu, L. Zheng, Y. Cui, C. Masouros, A. P. Petropulu, H. Griffiths, and Y. C. Eldar, "Seventy years of radar and communications: The road from separation to integration," arXiv:2210.00446, 2022.

[226] Y. Cui, F. Liu, X. Jing, and J. Mu, "Integrating sensing and communications for ubiquitous IoT:

Applications, trends, and challenges," IEEE Network, vol. 35, no. 5, pp. 158–167, 2021.

[227] A. Liu, Z. Huang, M. Li, Y. Wan, W. Li, T. X. Han, C. Liu, R. Du, D. K. P. Tan, J. Lu et al., "A survey on fundamental limits of integrated sensing and communication," IEEE Commun. Surveys Tuts., 2022.

[228] F. Liu, Y. Cui, C. Masouros, J. Xu, T. X. Han, Y. C. Eldar, and S. Buzzi, "Integrated sensing and communications: Towards dualfunctional wireless networks for 6G and beyond," IEEE J. Sel. Areas Commun., 2022.

[229] F. Liu, W. Yuan, C. Masouros, and J. Yuan, "Radar-assisted predictive beamforming for vehicular links: Communication served by sensing," IEEE Trans. Wireless Commun., vol. 19, no. 11, pp. 7704–7719, 2020.

[230] R. Liu, M. Li, Q. Liu, and A. L. Swindlehurst, "Dual-functional radarcommunication waveform design: A symbol-level precoding approach," IEEE J. Sel. Topics Signal Process., vol. 15, no. 6, pp. 1316–1331, 2021.

[231] F. Liu, L. Zhou, C. Masouros, A. Li, W. Luo, and A. Petropulu, "Toward dual-functional radar-communication systems: Optimal waveform design," IEEE Trans. Signal Process., vol. 66, no. 16, pp. 4264–4279, 2018.

[232] L. Chen, F. Liu, W. Wang, and C. Masouros, "Joint radarcommunication transmission: A generalized Pareto optimization framework," IEEE Trans. Signal Process., vol. 69, pp. 2752–2765, 2021.

[233] X. Liu, T. Huang, N. Shlezinger, Y. Liu, J. Zhou, and Y. C. Eldar, "Joint transmit beamforming for multiuser MIMO communications and MIMO radar," IEEE Trans. Signal Process., vol. 68, pp. 3929–3944, 2020.

[234] T. Huang, N. Shlezinger, X. Xu, Y. Liu, and Y. C. Eldar, "MAJoRCom: A dual-function radar communication system using index modulation," IEEE Trans. Signal Process., vol. 68, pp. 3423–3438, 2020.

[235] X. Yu, X. Yao, J. Yang, L. Zhang, L. Kong, and G. Cui, "Integrated waveform design for MIMO radar and communication via spatiospectral modulation," IEEE Trans. Signal Process., vol. 70, pp. 2293– 2305, 2022.

[236] Z. Xiao and Y. Zeng, "Waveform design and performance analysis for full-duplex integrated sensing and communication," IEEE J. Sel. Areas Commun., vol. 40, no. 6, pp. 1823–1837, 2022.

[237] C. Xu, B. Clerckx, S. Chen, Y. Mao, and J. Zhang, "Rate-splitting multiple access for multi-antenna joint radar and communications," IEEE J. Sel. Topics Signal Process., vol. 15, no. 6, pp. 1332–1347, 2021.

[238] J. Qian, P. Xu, Z. Liu, N. Fu, and S. Wang, "Enhancement for spectrum compatibility by mutual interference management," IEEE Wireless Commun. Lett., 2023.

[239] Z. Du, F. Liu, W. Yuan, C. Masouros, Z. Zhang, S. Xia, and G. Caire, "Integrated sensing and communications for V2I networks: Dynamic predictive beamforming for extended vehicle targets," IEEE Trans. Wireless Commun. doi: 10.1109/TWC.2022.3219890, 2022.

[240] W. Chen, L. Li, Z. Chen, T. Quek, and S. Li, "Enhancing THz/mmWave network beam alignment with integrated sensing and communication," IEEE Commun. Lett., 2022.

[241] J. Li, Y. Sun, T. Zhang, and R. Wang, "An indoor environment sensing and localization system via mmWave phased array," arXiv:2206.02996, 2022.

[242] S. Huang, M. Zhang, Y. Gao, and Z. Feng, "MIMO radar aided mmwave time-varying channel estimation in MU-MIMO V2X communications," IEEE Trans. Wireless Commun., vol. 20, no. 11, pp. 7581–7594, 2021.

[243] H. Sarieddeen, N. Saeed, T. Y. Al-Naffouri, and M.-S. Alouini, "Next generation TeraHertz

communications: A rendezvous of sensing, imaging, and localization," IEEE Commun. Mag., vol. 58, no. 5, pp. 69–75, 2020.

[244] Z. Chen, C. Han, Y. Wu, L. Li, C. Huang, Z. Zhang, G. Wang, and W. Tong, "Terahertz wireless communications for 2030 and beyond: A cutting-edge frontier," IEEE Commun. Mag., vol. 59, no. 11, pp. 66–72, 2021.

[245] 3GPP TS 28.552 version 17.10.0 Release 17 - 5G; Management and orchestration; 5G performance measurements

[246] A. -D. Marcu, S. K. Gowtam Peesapati, J. Moysen Cortes, S. Imtiaz and J. Gross, "Explainable Artificial Intelligence for Energy-Efficient Radio Resource Management," *2023 IEEE Wireless Communications and Networking Conference (WCNC)*, Glasgow, United Kingdom, 2023, pp. 1-6, doi: 10.1109/WCNC55385.2023.10119130.

[247] Notions of explainability and evaluation approaches for explainable artificial intelligence, https://www.sciencedirect.com/science/article/pii/S1566253521001093

[248] M. Masoudi et al., "Green mobile networks for 5G and beyond," in IEEE Access, vol. 7, pp. 107270-107299, 2019, doi: 10.1109/ACCESS.2019.2932777.