BeGREEN: Beyond 5G Energy Efficient Networking by Hardware Acceleration and AI-Driven Management of Network Functions

Mir Ghoraishi
Gigasys Solutions Ltd, UK
mir@gigasys.co.uk

Jose Oriol Sallent
Universitat Politècnica de Catalunya, Spain
jose.oriol.sallent@upc.edu

Miguel Catalan-Cid
i2CAT Foundation, Barcelona, Spain
miguel.catalan@i2cat.net

Guillermo Bielsa, Juan-Francisco Esteban-Rivas
RAN Innovation Department, Telefonica, Spain
guillermo.bielsa@telefonica.com
juanfrancisco.estebanrivas@telefonica.com

Vladica Sark, Jesus Gutierrez Teran
IHP GmbH – Innovations for High Performance Microelectronics, Germany
sark@ihp-microelectronics.com
teran@ihp-microelectronics.com

Simon Pryor
Accelleran, Belgium
simon.pryor@accelleran.com

***Abstract*—This paper presents a technical overview of BeGREEN project, a Horizon Europe, Smart Networks and Services Joint-Undertaking (SNS-JU) Phase 1 project kicked off on January 1, 2023 [1]. This paper is intended to describe BeGREEN's technical scope and objectives. These objectives aim at improving energy efficiency of the beyond 5G (B5G) networks. BeGREEN technical agenda includes analysis of the combined energy and spectrum efficiency of the B5G networks, based on massive multiple-input-multiple-output (mMIMO) scenarios. The project proposes a novel architecture that includes several innovative solutions. An offloading engine is used for hardware acceleration that is a solution for compute-heavy physical layer processing in 5G new radio (5G NR) mMIMO and beyond to improve the processing performance and energy efficiency. The architecture also includes joint communication and sensing (JCAS) for improving energy efficiency of the physical layer functions by, e.g., efficient beam-search and beam tracking, and uses reconfigurable intelligent surfaces (RIS) as an enabler for JCAS. BeGREEN proposes an artificial intelligence (AI)-assisted energy-aware "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. The project also proposes an AI Engine that is consist of an execution environment that can host AI models and will manage their lifecycle and access to data.**

Keywords—Energy Efficiency, Beyond 5G, Hardware Acceleration, Intelligent Plane, O-RAN Based Interface, AI Engine

I. INTRODUCTION

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

The energy usage of the telecoms is currently on the rise. The telecom operators account for 2-3% of the global energy consumption, making them one of the most energy-intensive industries. The impetus for saving network energy has grown remarkably as energy cost has become a significant part of network operational expenditures (OPEX). The fact that 5G New Radio (NR) is more efficient than previous generation radios, does not conceal the fact that a dense deployment of 5G networks (e.g., by incorporating massive-multiple-input-multiple-output, mMIMO, technologies) communicating on different frequency bands and with higher bandwidths, making the energy efficient concepts much more necessary for NR in upcoming 3GPP Rel-18 and -19 [3].

A recent report on 5G power consumption suggests that if operators act wisely and optimise the hardware used in their systems, this will bring a decline in power consumption of up to 70% [4]. It is calculated that a 5G base-station needs three times more energy to provide the same coverage as a 4G network, which, in turn, results in high energy costs and capital expenditure for operators. Around 70% of the consumed energy in the network is in the radio access network (RAN) [5]. A 5G RAN consumes up to 2.7 kilowatts of power with 64 transmit by 64 receive (64T64R) mMIMO configurations in a typical condition, whereas an LTE radio consumes about 0.8 kilowatts [6]. The dominant contributors to power consumption are power

amplifiers (PAs), baseband process modules, digital intermediate frequency (DIF) and transceivers. By using new generations of chipsets, and further by smart management of the network functions using Artificial Intelligent (AI) methods, it is estimated to achieve between 30-70% in energy saving. Combined with the rising costs of spectrum, capital investment and ongoing RAN maintenance/upgrades, energy-saving measures in network operations are necessary rather than nice to have.

In this context, BeGREEN [2] 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. BeGREEN aims to establish a solid techno-economic basis for assessment of technology choices where energy efficiency is an explicit characteristic rather than an afterthought. The project will take an evidence-based assessment of current and emerging radio access technology choices to expose the practical energy cost of cutting-edge technologies to help the community achieve energy consumption targets. This paper is intended to provide a sketch of how BeGREEN project will tackle the energy efficiency problem in B5G context and an outline of the envisaged solution approaches. In this respect, Section II discusses an initial reference architecture and identify the potential areas of innovation. Then, Section III further elaborates the opportunities for energy efficiency improvements at the physical layer while Section IV discusses the relevant role that AI is expected to play in building an energy efficiency-centric intelligent plane together with NFV energy optimisation.

II. BeGREEN Architecture and Innovations

Figure II-1 shows the overall architecture of BeGREEN. The proposed architecture, based on mMIMO as well as lower order MIMO scenarios assisted with joint communications and sensing (JCAS) and reconfigurable intelligent surfaces (RIS), uses the advantage of an O-RAN Near-RT RIC and AI/ML xApps to provide extra features to B5G/6G RAN energy efficiency. Primarily, the project provides the analysis of mMIMO and cell-free distributed MIMO architectures, and their impact on energy utilisation in the network. With near-future deployments allowing a large number of antennas to achieve cell-free connectivity, the proper utilisation of spectrum resources, energy transmission and interference mitigation becomes essential. A small difference in the utilised joules (J) per bit per user would make a huge impact from the operator's perspective.

As such, BeGREEN's aim is to quantify the combined energy and spectral efficiency impact of practical constraints that conventional and emerging base-stations and mobile devices experience in providing data services to end users. Defining the energy efficiency of a radio network requires consideration of several factors.

If implemented at a base-station level, maximising the bit/J might involve a low level of network densification, avoiding the use of centralized mMIMO base-stations spanning a large surface of the coverage. Instead, increasing the number of base stations, or even granting cell-free deployments would imply emitting less power while mitigating interference more

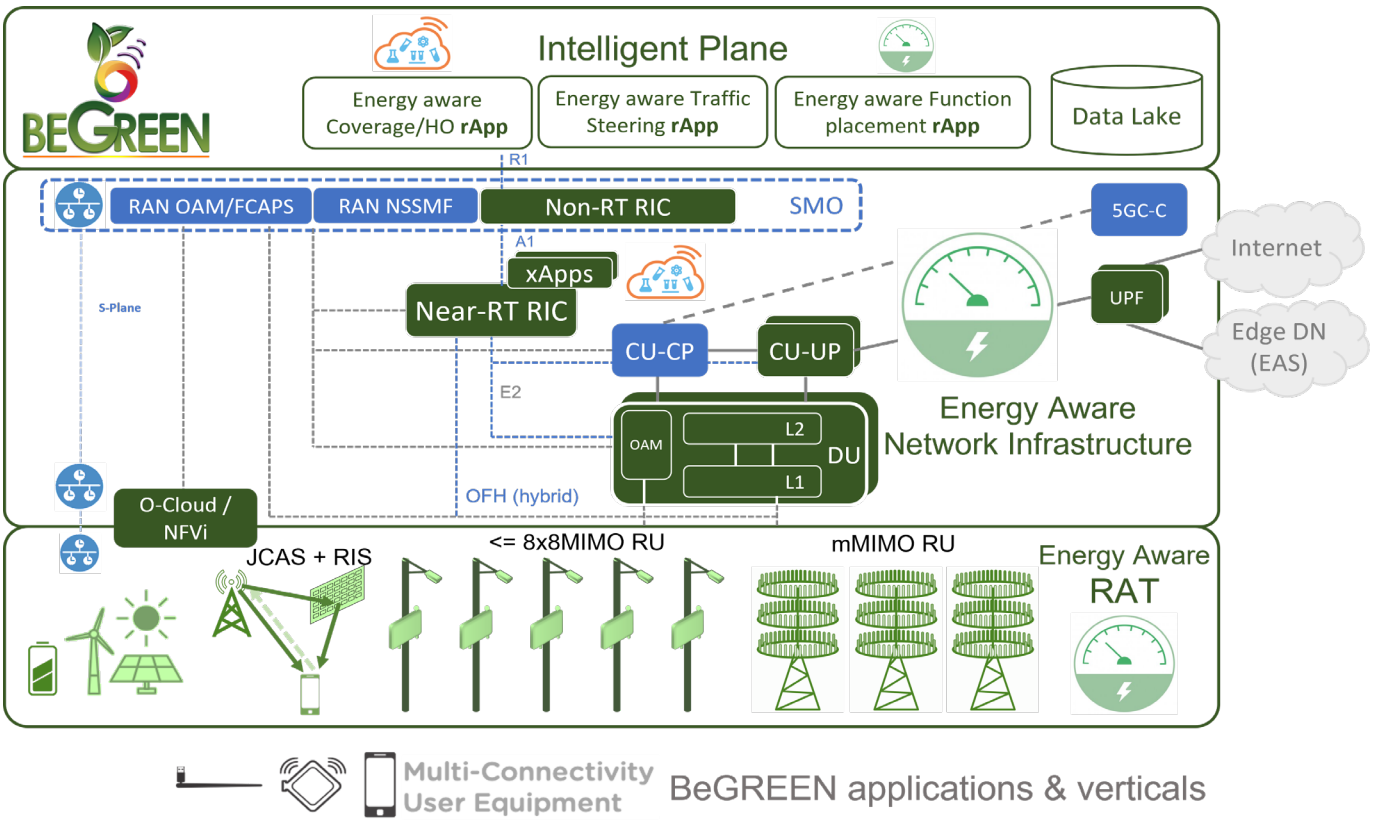

Figure II-1 BeGREEN proposed architecture

effectively, resulting in more energy-efficient communication links. This network densification will also be enhanced by the use of relays and RIS.

Based on the assumption that all offered traffic must be carried and no new spectrum is available, the alternative to provide additional capacity is therefore to build more base-stations, which will also increase energy consumption. BeGREEN considers combined energy and spectrum efficiency in terms of bit/s/Hz/J on an area basis and aim to use techniques to optimise against this metric for different architectures, allowing for future traffic growth. The areas over which the metric is applied are defined in terms of the traffic distribution, service type and temporal variability of load to identify which radio network solutions are most appropriate in the different circumstances. Furthermore, sensing-assisted link establishment approach, enabling optimised spatial resource-allocation, e.g., enhanced transmit power and beamwidth allocation, towards network energy usage optimization are included in the architecture analysis.

BeGREEN proposed architecture includes distributed unit (DU) hardware acceleration using a graphics processing unit (GPU) based offloading engine as a fundamental technology component to achieve energy efficiency for offloading heavy data processing to enable large scale mMIMO radio-access complex data processing, as well as centralized processing units for cell-free distributed MIMO control. In addition, direct interaction of Near-RT RIC with RU to improve energy efficiency by means of rApps/xApps controlling of radio unit (RU) functions, as well as artificial intelligence (AI)-assisted digital predistortion (DPD) and envelope tracking solutions are being considered. To enable longer RAN resource deactivation times (macro-sleeps), an appropriate management of the carriers and channels (antennas) of the BS according to traffic distributions and demands at a macro-time, e.g., minutes or even hours, can avoid the resource waste emanating from the over-dimensioning of the network to meet peak hour requirements. For the xApp control algorithms, depending on RU capabilities, the following functionalities are considered:

- Reduction of transmit power at idle times (reducing throughput and energy consumption)
- Smart handovers to move the UEs to adjacent base-stations, to create idle carriers which can be powered off
- Disabling transmit carriers of RUs, maybe with compensation of transmit power of adjacent base-stations
- Switching of 4x4 MIMO to 2x2 MIMO or single-input-single-output (SISO) to reduce throughput and energy consumption
- xApp driven eco-self-optimising-networks (eco-SON), extensions of typical SON but with energy reduction as the optimisation target instead of aggregate throughput, with policies (e.g., idle periods during the night) driven by rApps

BeGREEN proposes the design and development of an 'Intelligent Plane', incorporating O-RAN, along with user plane and data plane, for AI-assisted network function energy optimisation that allows the data, model and inference to seamlessly exchange over the network. The design of AI/ML algorithms that dynamically select central processing unit (CPU) power saving modes (e.g., C-states) or adapt the number of active virtual network function (VNF) instances to minimize energy consumption without affecting workload performance are included. Moreover, AI/ML algorithms that using *explainable* and interpretable AI algorithms (e.g., Shapley algorithms and partial dependency plots) that will accurately pinpoint energy influencing factors of the network functions beyond traffic are being analysed. Design and implementation of next generation Edge, aiming to minimize the overall energy cost by using, AI-assisted procedures to jointly control RAN resources and Edge service parameters, RAN user-plane NF acceleration plus EAS offload are also included in the architecture. These features are introduced in more details in Section IV.

III. Physical Layer Enhancements for Energy Efficiency Improvement

BeGREEN will investigate several physical layer enhancements in order to improve the energy efficiency of the RAN. These enhancements include hardware acceleration, physical layer processing, and joint communication and sensing.

As bandwidth increases and mMIMO RUs are deployed, conventional processor cores, e.g., x86, struggle to keep up and start driving impractical levels of power consumption. In this case, **hardware acceleration** is a solution for compute-heavy physical layer processing in 5G NR and beyond to improve the processing performance and energy efficiency. BeGREEN envisions to develop an innovative offloading engine, for energy efficient offloading of mMIMO related processing. It is desirable to have an implementation that can fit the scale needed from on-site DU to centralised cloud-based infrastructure using PCIe add-on or similar. One candidate is the FPGA path that has been very much explored with multiple options so far. Also, ASIC solutions are available, but these are not generic, and neither are future proof enough. BeGREEN's objective is to open a path with off-the-shelf GPU, by proposing a new architecture based on low-cost GPU embedded into the DU. Figure III-1 shows the envisaged offloading engine. Among the set of considerable tasks to be offloaded into the GPU there is the beamforming weights related calculation arithmetic. Such arithmetic includes heavy weight signal processing tasks that are required for the SRS channel estimation and the actual uplink (UL) and downlink (DL) beamforming weight calculations.

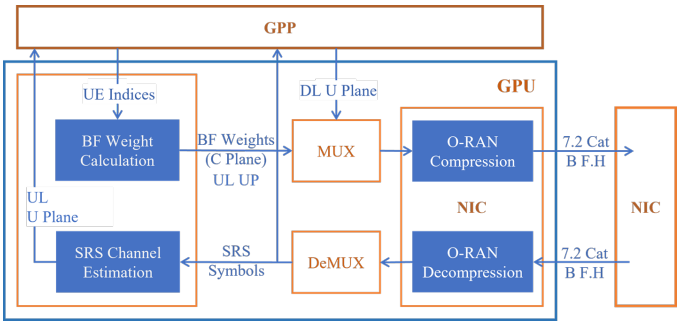

Figure III-1 GPU-based offloading engine

Those sets of calculation are heavily based upon FFT and matrix inversion arithmetic which makes them a perfect fit for highly parallelisable GPU platforms. Another well-suited task for GPU implementation is the LDPC FEC (Forward Error Correction) decoder. FEC adds redundancy to wirelessly transmitted bits so that the receiver can detect and correct errors.

Although LDPC codes have near capacity-achieving decoding performance, the decoding complexity is high and LDPC decoding creates a computational challenge for the 5G gNBs, which potentially need to decode many codewords from multiple users at high data rates.

Joint Communication and Sensing (JCAS) is envisioned to become a part of future B5G and 6G networks. This means that the next generation mobile networks would have a so-called perceptive capability. JCAS offers a variety of functionalities which can be beneficial for improvement of the network energy efficiency. Notably, it can also be the "enabling effect" for improving the energy efficiency of other industry verticals.

JCAS will play an important role in optimizing some physical layer functions, such as beam-search/beam tracking. In conventional systems, the beam-search procedures are either based on extensive or hierarchical search. In both cases, the performed beam-search requires assignment of spectrum resources and demands energy and time, which is in practice a very inefficient procedure. BeGREEN will develop sensing facilitated beam-search algorithms which should prioritize the searching areas based on the sensing information. This should significantly improve the beam-search procedure, reduce the spectrum usage, and increase energy efficiency. Additionally, the same approach would be used to develop a predictive beam-tracking algorithm to efficiently track the user and reduce the number of additional beam-search procedures.

BeGREEN will also leverage the sensing functionality of (public mobile networks) PMNs for optimal resource allocation and planning. The JCAS would be used to estimate the number of users and their position in the network to optimally assess the network load. This information will be later used to assign optimal network resources and minimize energy usage, e.g., by turning off unnecessary equipment. Additionally, JCAS based algorithms for creating digital twin of the environment will be investigated in BeGREEN. The digital twin representation would help in the networks resource planning for further network optimization and efficient energy usage. Additionally, BeGREEN will use RIS infrastructure as one of the enablers for JCAS. In order to provide the PMN operator with RIS control mechanisms, we will investigate the requirements and deployment options for a O-RAN-compliant sub-millisecond control interface, paying special attention for the support of additional sensing channels that RIS may require. BeGREEN will also investigate how to design a self-configurable RIS, which does not use a control channel for those cases where using a control interface is not an option.

IV. AI-ASSISTED O-RAN-BASED EDGE AND NFV ENERGY OPTIMISATION

O-RAN alliance promotes a new open architectural framework featuring non-RT RIC and Near-RT RIC to host intelligent functions, i.e., rApps and xApps, that can react at different time scales to low level RAN telemetry in an

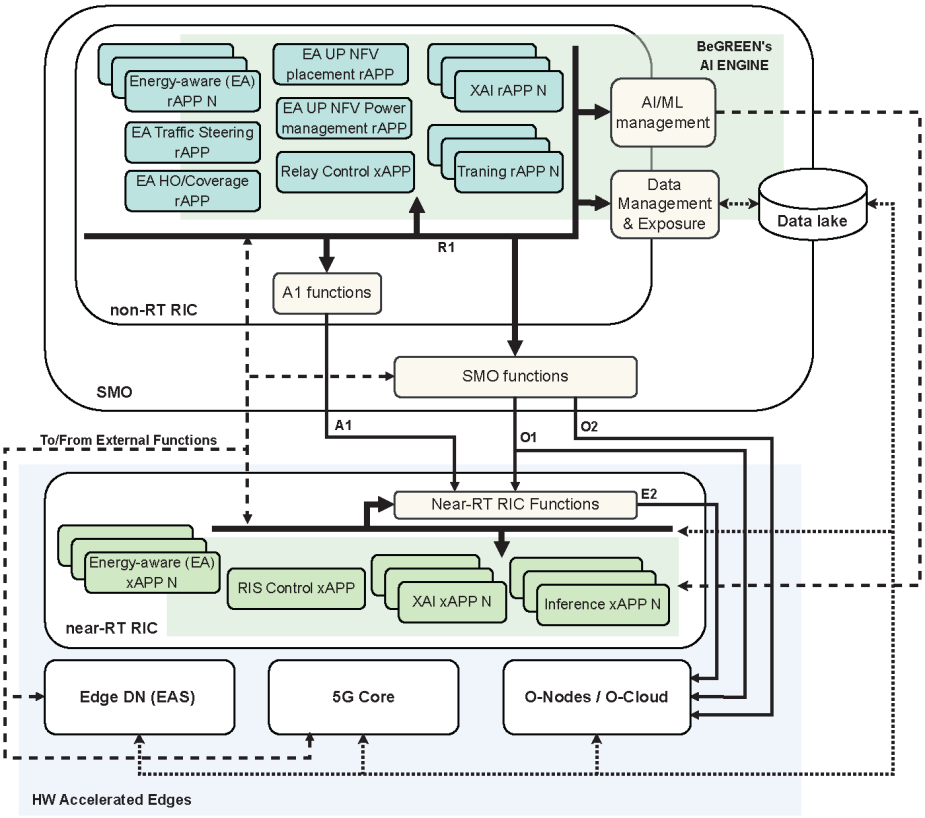

Figure IV-1 BeGREEN Intelligent Plane architecture

autonomous way. This enables the definition and implementation of data-driven optimizations, and closed-loop control and automation which can be based on AI/ML techniques [7]. While state-of-the-art (SotA) approaches are usually focused on performance-centric architectures and optimizations, BeGREEN will investigate how the O-RAN architecture can be extended to better address energy optimization.

BeGREEN proposes an AI-assisted energy-aware "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. This "Intelligent Plane" will evolve O-RAN components and interfaces to enhance power consumption monitoring and control in the RAN infrastructure. Furthermore, while O-RAN is currently focused solely on the RAN, BeGREEN will tackle beyond SotA evolution towards the infrastructure for the mobile core user-plane, for radio Edge located applications (EAS, with user plane function, UPF and CU-user plane , CU-UP, instances), including specific Edge AI services. Figure IV-1 depicts the envisioned architecture of BeGREEN's Intelligent Plane in order to address the main innovations and challenges listed below:

- **Energy-aware interfaces:** The designed architecture will extend O-RAN interfaces in order to enhance the exposure of energy related metrics and control services to the envisioned energy-aware rAPPs and xAPPs. The exposed data and services will leverage the findings of the project related to the definition of energy efficiency KPIs, and to the monitoring and optimization of CU/DU/RU components (E2 and O1 interfaces) and the O-Cloud (O2 interface). As reported in [8], these will be key features to enable O-RAN networks to gradually become more energy efficient, allowing rAPPs to implement and manage closed-loop automations through the R1 and A1 interfaces. Therefore, SMO functions like inventory or configuration will be also exposed to rAPPs. Additionally, as shown in Figure IV-1, the exposure of data and control services of external functions impacting energy consumption (e.g., 5G core network functions or application servers) through existent or new interfaces or components will be studied.
- **AI Engine:** As depicted in Figure VI-1, the AI Engine will consist of an execution environment that can host AI/ML models and will manage their lifecycle and access to data, where training and inference is envisioned to be performed by AI-driven rAPPs and xAPPs. Moreover, BeGREEN will study the impact of federated learning approaches on the energy efficiency of AI/ML algorithms and its lifecycle management. In such an approach, we envision that the model will be trained, monitored and re-trained at the SMO/rAPP level, and deployed at near-RT RIC/xAPP level for inference. This AI architecture, which is aligned with the initial specification of O-RAN regarding AI/ML support [9], will be further extended to support explainable and interpretable AI algorithms that will accurately pinpoint energy influencing factors of the network functions beyond traffic. Using the influencers, it will be possible to calculate energy efficiency rating and an associated energy score. As shown in Figure II-1 these scores will be exposed or applied by rAPPS/xAPPS, focusing energy optimizations into specific network areas needing additional orchestration. Additionally, we will study energy efficiency-aware training mechanisms for on-boarding the proposed AI/ML solutions.
- **AI-driven energy-efficient management of user plane network functions:** Software-based implementation of user plane functions has been widely adopted in 5G since they provide higher flexibility and adaptability to service requirements. This comes at a cost in terms of energy consumption, especially when considering user plane functions that are intensive in terms of processing [10]. Examples of these functions are the UPF, the CU-UP or the DU. BeGREEN will address this challenge by designing AI/ML algorithms that dynamically select CPU power saving modes (e.g., C-states) or orchestrate the number of VNF instances to minimize energy consumption according to the utilization patterns of the network and without affecting vRAN and UPF performance.
- **AI-driven energy optimization for edge computing applications:** BeGREEN will carry an in-depth evaluation of the impact of AI services' energy consumption on a mobile network. AI-based services are inherently energy-intensive, since they demand gathering, transmitting, and analyzing data flows to offer real-time services to end users. Therefore, BeGREEN will study and implement mechanisms to jointly control vRAN and Edge resources to minimize the overall energy cost while meeting AI service performance targets. Additionally HW acceleration in Near-RT RIC to offload AI/ML xApps will be advanced to quantify and manage the holistic Edge energy savings.
- **O-RAN-based control of innovative wireless technologies:** BeGREEN will study the benefits of innovative technologies like intelligent reconfigurable surfaces (RIS), cell-free distributed MIMO, and relay-enhanced RAN (e.g., Mobile Integrated Access and Backhaul) on energy efficiency, targeting its integration with O-RAN architectures. As depicted in Figure II-1, this will enable intelligent control by means of AI-driven rAPPs or xAPPs, which will also have an impact on the energy consumption of the Edge AI services.

V. SUMMARY

This paper introduced SNS BeGREEN project's scope and architecture. The project objective is improved energy efficiency for the beyond 5G (B5G) networks. This is proposed at multiple levels, i.e., introducing a reference architecture for combined energy and spectrum efficiency in mMIMO scenarios, of the B5G networks, based on massive multiple-input-multiple-output (mMIMO) and cell-free scenarios, an offloading engine for hardware acceleration of the compute-heavy physical layer processing, as well as an AI-assisted energy-aware "Intelligent Plane" as an additional plane along with user plane and data plane to allow the data, model, and inference to be seamlessly exchanged between network functions using an AI Engine that is consisted of an execution environment that can host AI models and will manage their lifecycle and access to data.

ACKNOWLEDGMENT

This work is supported by the European Commission's Horizon Europe, Smart Networks and Services Joint Undertaking, research and innovation program under grant agreement #101097083, BeGREEN project.

REFERENCES

[1] 6GSNS Overview of Phase 1 Projects. Available: https://smart-networks.europa.eu/sns-phase-1/

[2] SNS BeGREEN Project: www.sns-begreen.com

[3] 3GPP TR38.864, "Study on network energy savings for NR (Release 18)", 3rd Generation Partnership Project (3GPP), Technical Report (TR) 38.864, Dec. 2022

[4] 5G Energy Consumption and Operator Sustainability Initiatives, ABI Research, March 2022. Available: https://www.abiresearch.com/market-research/product/7779678-5g-energy-consumption-and-operator-sustain/

[5] David Lopez-Perez, et. Al., "A Survey on 5G Radio Access Network Energy Efficiency: Massive MIMO, Lean Carrier Design, Sleep Modes, and Machine Learning," IEEE Comm. Surveys and Turotials, January 2022.

[6] "Green 5G: Building a Sustainable World," Huawei, August 2020.

[7] 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.

[8] Telecom Infra Project, Open RAN MoU Group, "Open RAN Technical Priority "Release 2" Document: Focus on Energy Efficiency", March 2022.

[9] O-RAN Alliance, "O-RAN AI/ML Workflow Description and Requirements 1.03", October 2021.

[10] Y. Anser, J. -L. Grimault, S. Bouzefrane and C. Gaber, "Energy-Aware Service Level Agreements in 5G NFV architecture," 2021 8th International Conference on Future Internet of Things and Cloud (FiCloud), Rome, Italy, 2021, pp. 377-382, doi: 10.1109/FiCloud49777.2021.00061.