Dynamic Resource Allocation and Energy Optimization in 5G O-RAN: Real-World Insights and Testbed Evaluations

Dynamic Resource Allocation and Energy Optimization in 5G O-RAN: Real-World Insights and Testbed Evaluations

1st Caterina Leonelli*[1,2], 2nd Dimitris Kefalas*[1,3], 3rd Serge Fdida[1], 4th Paolo Bellavista[2], and 5th Thanasis Korakis[3]
[1]Sorbonne Université, CNRS, LIP6, Paris, France
[2]Dept. of Computer Science and Engineering, University of Bologna, Bologna, Italy
[3]Dept. of Electrical and Computer Engineering, University of Thessaly, Volos, Greece
Emails: caterina.leonelli@sorbonne-universite.it, dkefalas@uth.gr, serge.fdida@sorbonne-universite.fr,
paolo.bellavista@unibo.it, korakis@uth.gr

***Abstract*—The energy efficiency of 5G Radio Access Networks (RANs) has become a critical area of research due to the growing energy demands of the 5G disaggregated architectures and the increasing environmental concerns surrounding mobile networks. In this paper, we address the fundamental challenge of optimizing energy consumption in Open RAN (O-RAN) 5G networks by dynamically scaling Central Unit (CU) components based on traffic demands. Leveraging a real-world testbed environment, we empirically analyze the relationship between data volume, architectural configurations, and energy usage. Our study identifies key tuning parameters for energy management and optimization, showcasing the impact of dynamic resource allocation on reducing energy consumption. Experimental results demonstrate that our implementation of a dynamic CU allocation policy can achieve energy savings of up to 60% compared to static configurations, without compromising quality of service (QoS).**

***Index Terms*—Energy Efficiency, CU Dynamic Scaling, Resource allocation, 5G O-RAN**

I. Introduction

Recent studies have shown that the sector of Digital Communication Technologies (DCT) is responsible for over 2% of the global greenhouse gas emissions [1]. A significant portion of these emissions comes from cellular networks, driven by the growing demand for connectivity and high traffic loads of mobile devices. As seen in [2], within the mobile network, the Radio Access Network (RAN) is responsible for 73% of the total energy consumption. This highlights the importance of ongoing research and development of 5G RAN architectures and technologies specifically aimed at minimizing energy consumption.

Considering the aforementioned, the energy efficiency of the 5G RAN is currently a major area of research due to the beyond 5G RAN architecture, where the Base Station (BS) is split into multiple components compared to previous monolithic approaches resulting in even greater energy demands than previous generations [3]. In fact, 5G RAN leverages a Network Function Virtualization (NFV) architecture, enabling greater flexibility and scalability for deploying both the 5G Core Network (5G CN) and 5G RAN, but may also generate higher energy usage compared to traditional RAN solutions. Despite that, this architecture presents opportunities to optimize energy consumption, as we demonstrate in our work.

In particular, 5G RAN is essentially a cloud-based version (known as C-RAN) that enables more flexibility in designing the architecture of the infrastructure since there is the possibility of choosing various functional splits, with 3GPP Option 2 being the most used one [4]. According to this split, the gNB has been decomposed to the Central Unit (CU), which hosts network layer functionalities such as the Packet Data Convergence Protocol (PDCP) and Radio Resource Control/Service Data Adaptation Protocol (RRC/SDAP), and the Distributed Unit (DU), which manages data link layer functions including the physical (PHY), medium access control (MAC), and radio link control (RLC) layers. Additionally, the Radio Unit (RU) remains at the edge of the network, directly interfacing with the antenna and focusing on radio frequency (RF) processing. The communication between the CU and DU is established over the F1 interface that is responsible for managing both control plane signaling and user plane data transmission. The F1 Application Protocol (F1AP) supports that a single CU can connect to multiple DUs, by improving the scalability of 5G C-RAN architectures. This can result in high traffic loads that the CU has to handle, resulting in increased processing demands and, consequently, power consumption.

Pushed by the recognized advantages of open ecosystems also in the 5G sector, the Open RAN (O-RAN) architecture adds specifications to the RAN that promote the connection through open interfaces of the different network components, which are managed and optimized by Radio Intelligent Controllers (RIC) and the Session Management Orchestrator (SMO). Based on the response times requirements of the O-RAN system, two distinct types of RICs are introduced: the

This work is partly funded by the European Commission (Horizon Europe programme) under the Grant Agreement numbers 101079774 (SLICES-PP) and 101131207 (GreenDIGIT). The first two authors contributed equally to this work. Asterisk indicates corresponding authors

near Real-Time RIC (NearRT RIC) and the non-Near Real-Time RIC (non-NearRT RIC). In the context of aforementioned disaggregation of the architecture, O-RAN permits to develop different strategies in order to create different topologies of CUs and DUs placement. Moreover, the introduction of the RICs in O-RAN provides a powerful and flexible framework for improving resource management and in particular energy efficiency. This integration paves the way to develop RICs that implement energy management strategies such as dynamic scaling and resource management, the definition of energy metrics and support AI-driven optimization. As a result, RICs play a major role in creating energy-aware networks that can address the growing demands for sustainable and scalable connectivity.

Despite the potential of O-RAN, realistic experimental solutions for energy testing, metric evaluation, and assessment remain limited in the existing literature due to the complexity of such deployments. In this work, we use a real-world testbed environment to analyze the behavior of energy consumption in a 5G network infrastructure by considering different deployment scenarios with different data volume loads and different placements of disaggregated components. Based on those results and using a real-world dataset, we also deployed an execution environment where we prototyped and evaluated a scheduling policy that scales CU instances out or in to minimize the energy consumption of the 5G network. The main contributions of this work can be summarized as follows:

- Collecting energy measurements from real-world 5G experimental setups
- Analysis of the energy consumption behavior of 5G data plane entities
- Propose a policy for scheduling the number of CUs and implementing a proof of concept of this scheduler

The rest of the paper is organized as follows: Section II provides an overview of the related literature, focusing on existing approaches to minimize the energy consumption in the RAN along with scaling implementations in 5G networks. Section III provides our overall system model architecture and configuration, and Section III-A evaluates our implementation by detailing the experimental setup and presenting the results. Finally, in Section IV we conclude our work and discuss future directions.

II. Backgrounds and Related Works

The energy efficiency of the 5G RAN has gained significant research interest in recent years due to its critical importance. One of the first attempts to improve the energy efficiency of 5G RAN is the adoption of BS ON-OFF switching, originally created for IEEE 802.11 networks [5], as shown in [6] and [7]. In disaggregated 5G-RANs, BS ON-OFF switching becomes CU-DU-RU ON-OFF management with the most common approaches limited to DU-RU sleep mode schemes as seen in [8] since the CU is typically responsible for higher-layer control functions such as ensuring control-plane functionality and transmitting data plane packets. Nowadays, to minimize power consumption at the CU level, the most common approach involves optimizing the placement of CU functionalities across edge and regional clouds, as demonstrated in [9] and [10], leveraging techniques such as Integer Linear Programming (ILP), Mixed Integer Linear Programming (MILP), and heuristic algorithms to reduce active nodes while maintaining latency and QoS requirements.

Numerous strategies have been explored to optimize energy consumption in 5G networks, particularly in VNFs using scaling approaches. For example in [11], authors examined vertical scaling of the RAN and horizontal scaling of control and user plane functions (CN components). As demonstrated in the paper, the vertical scaling approach was beneficial for accommodating incoming demand, alleviating potential blockages caused by multiple UEs requesting network services concurrently. On the other hand, horizontal scaling of the Core Network components was found to be helpful for achieving significant energy gains. In addition, the authors used machine learning forecasting to develop a scheduler that proactively scales components based on traffic metric predictions. In stark contrast with them, our work focuses on horizontal scaling of RAN components rather than vertical scaling.

Most of the literature has focused on orchestrating the BS function's resource allocation, validating their results only in simulated environments [12] [13]. For instance, Joda et al. [14] developed a strategy for the placement of CU-DU network functions in regional and O-Cloud nodes, while simultaneously addressing user association with RUs. Their simulations of various user mobility scenarios indicated that their strategy provides a good balance between cost minimization and performance. However, while these two works used "CPU cycle per second" or "GOPS" as *cost metric*, we empirically measure the energy consumption in Joules from a real-world test-bed by running actual workload experiments. We believe that our work extends the current state of the art by providing novel, measurable insights that can be verified whose reproducibility is guaranteed by the open-source nature of our framework (see Section III-A for more details). Concerning metric choice, we highlight that many works [15] [16] [9] focus on counting *power* (in Watts) rather than *energy* (in Joules) in their power-saving optimization strategies. Although these works primarily aim to reduce the overall system's power consumption, a recent study [17] highlights energy as a more meaningful metric. In fact, energy takes into account the power and the time span involved in managing dynamically changing BS functions. For those reason, in our paper we align with this assumption, though we also plan to explore the power-versus-energy debate in future work. This distinction between power and energy metrics underscores the need for precise evaluation methods, as optimization strategies must align with the dynamic and temporal nature of BS operations to achieve meaningful energy savings. There exist some novel work on methodologies for testing and measuring the essential parameters for energy saving in O-RAN [18]. Our work identifies key parameters offering tuning opportunities for energy management and optimization.

III. System Architecture & Experimental Evaluation

This section presents the architecture of our proposed 5G system, which includes a nonRT-RIC controller located at the SMO entity and is responsible for horizontally scaling the CU instances in a 5G O-RAN network in order to minimize energy consumption while dynamically adapting to network demands.

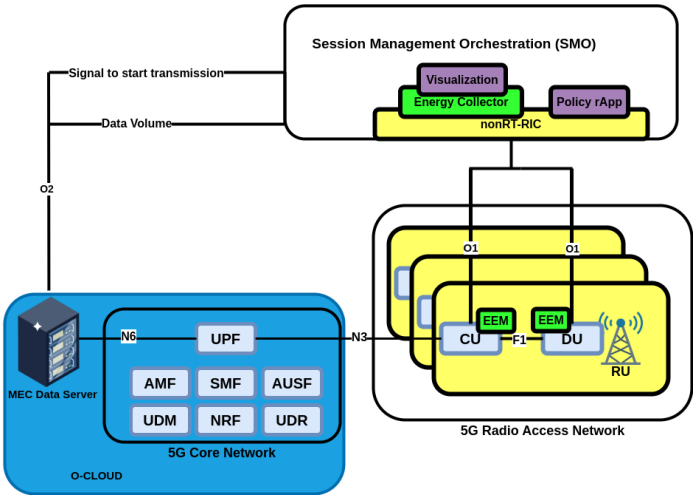

Fig. 1: Logical Architecture

The proposed 5G system architecture is based on a foundational 5G CN deployment, having only the necessary key components such as the Network Repository Function (NRF), User Data Repository (UDR), Unified Data Management (UDM), Authentication Server Function (AUSF), Access and Mobility Management Function (AMF), Session Management Function (SMF), and a User Plane Function (UPF). All the components of the 5GCN were placed at the O-CLOUD in order to be in a controlled environment from the SMO. We focused on the basic architecture of the 5GCN as it effectively meets the connectivity and service requirements of the UEs while aligning with our primary objectives. For the 5G-RAN, we follow the typical O-RAN architecture that supports only the existence of the F1 interface that implements the 3GPP Option 2 split, enabling the functional separation of the CU and DU components of a 5G network as mentioned in Section I. Our proposed 5G system is O-RAN compliant since we support the existence of the SMO platform responsible for the RAN and O-Cloud management. The SMO is equipped with an Energy Collector Module (ECM) executed as an rApp, application operating in nonRT-RIC, which aggregates energy data from different components of the disaggregated RAN. Each component of the disaggregated RAN (CUs, DUs) is equipped with Energy Exporter Modules (EEM) that monitor and report its energy consumption to the nonRT-RIC. According to O-RAN, O1 is the interface with which the nonRT-RIC communicates with all O-RAN Managed Elements (MEs). All the aforementioned components discussed are shown in Figure 1.

In our real-world testbed environment, we choose to implement a fully controlled scenario with respect to the data volume requests of the UEs to validate the proposed scaling policy. In this scenario, the rApp running on the nonRT-RIC informs the Multi-access Edge Computing (MEC) server about the demanded traffic volume for the UEs and signals it to start the transmission. The MEC server is located at the O-Cloud and the communication is happening via the O2 interface. Since the rApp already knows the data volumes that will be sent to the UEs, it can take the decision whether the CU component should be scaled out or in, in order to minimize the energy consumption of the O-RAN. These predefined conditions ensure the feasibility of the proposed scaling policy, as the focus of this work is on validating the policy itself, under the assumption that the traffic volume and duration is already known.

A. Experimental Architecture

In this section, we describe the use cases and our setup for the experiments that we performed. To conduct our experiments, we utilized the OneLab testbed located at the Sorbonne University, part of the French node of the SLICES-RI infrastructure [19]. SLICES-RI is a platform that enables researchers worldwide to deploy their experiments on a distributed testbed, with reproducibility being one of its key goals. To promote the reproducibility of our work, aligning with the SLICES-RI objective, the code for all the experiments conducted in this paper is available at the following URL: https://github.com/RootLeo00/Dynamic-Scaling-Policy-Energy-5G-ORAN.

To deploy an O-RAN 5G network we utilize the OpenAirInterface (OAI) [20] implementation of the 5GCN and 5G-RAN. OAI is one of the most popular, continuously integrated, and maintained open-source projects for 5G VNF implementation. These VNFs are containerized using Docker and orchestrated with Kubernetes. Specifically, four nodes equipped with Intel(R) Core(TM) i7-10700 CPU @ 2.90GHz CPUs, are used to deploy the different VNFs; in each node we deploy 1 DU-RU and 1 UE and we randomly place from 1 to 4 CU distributed across the nodes. The link among the RUs and UEs is emulated via the RF-simulator from the OAI implementation. We also utilize one extra node with the same CPU model to deploy our 5GCN.

To acquire the power consumption metrics of each VNF, we use Scaphandre [21], an open-source tool that has also been adopted in several recent research works, such as those in [22] [23]. Scaphandre leverages the RAPL (Running Average Power Limit) sensor and the powercap kernel module of the Linux OS to measure the instant power consumption of the node over a specific duration. It then maps this global energy measure to individual processes by scraping information from the /proc directory of the Linux OS. After obtaining the power consumption of a single process, we sum up those belonging to the same container, creating a coarser metric, the "power consumption of a container." These metrics are then available via the Prometheus exporter.

B. Experimental Evaluation

Our scenario involves multiple User Equipment (UE) devices downloading substantial data volumes, a common occurrence in

the AI era. Use cases include downloading large AI models or their outputs, such as images, documents, 3D objects, or videos. For example, a smartphone may download object detection results from an uploaded image, or an AR device may retrieve 3D objects for real-time interaction.

In our system, the traffic is generated at the MEC server and forwarded to the 5G network. Initially, it passes through the data plane component of the 5GCN, specifically the UPF, before being sent to the 5G RAN. Within the disaggregated RAN, the traffic flows from the CU to the DU and finally to the RU.

We conducted experiments scaling the CU from 1 to 4 nodes, while fixing the number of DUs and UEs to 4. Varying rates of traffic with a fixed packet size was generated using *iperf* between the MEC server and the UEs, to simulate diverse traffic loads. The traffic volume load ranged from 20 MB to 2 GB. To ensure simultaneous data transfer across all UEs, Python threads were used to enable parallel execution of the *iperf* sessions. While the *iperf* sessions were running, we collected power metrics (in Watts) from the Prometheus server. Upon completion of the tests, we calculated the energy consumption (in Joules) by integrating the power consumption over the duration of the experiment, using the SciPy library's *trapezoid* function.

The **Host Energy Consumption** (E_{host}), representing the baseline energy for keeping the physical machine operational, is calculated as:

$$E_{\text{host}} = P_{\text{avg}} \cdot T_{\text{service}}$$

where P_{avg} is the average host power consumption, calculated as the mean of the power values sampled over a 10-minute period, and T_{service} is the service duration. For deployemts with a single CU, T_{service} is the maximum recorded duration, whereas for multiple CUs, it sums the durations of all involved CUs.

The **Activation Energy Consumption** ($E_{\text{activation}}$) accounts for the energy required to deploy CUs on the node, computed as:

$$E_{\text{activation}} = P_{\text{activation}} \cdot T_{\text{deploy}} \cdot (N_{\text{CU}} - 1)$$

where $P_{\text{activation}}$ is the average activation power rate, T_{deploy} is the fixed deployment duration (averaged over 12 deployments of a single CU), and N_{CU} is the number of CUs being deployed. To guarantee UE connectivity, at least one CU must remain active in the RAN throughout its entire operation. Consequently, $N_{\text{CU}} - 1$ refers to the number of activations corresponding to the additional number of CUs that need to be deployed.

The **Service Energy Consumption** (E_{service}) reflects the energy used by RAN components especially the CU and DU during data processing and transmission, calculated as:

$$E_{\text{service}} = \int_0^{T_{\text{service}}} P_{\text{service}}(t)\, dt$$

where $P_{\text{service}}(t)$ represents the power usage of the service components sampled at one-second intervals.

Finally, the **Total Energy Consumption** (E_{total}) is given by:

$$E_{\text{total}} = E_{\text{host}} + E_{\text{activation}} + E_{\text{service}}$$

giving a comprehensive view of the energy requirements for the network operations.

C. In-the-field Experimental Results

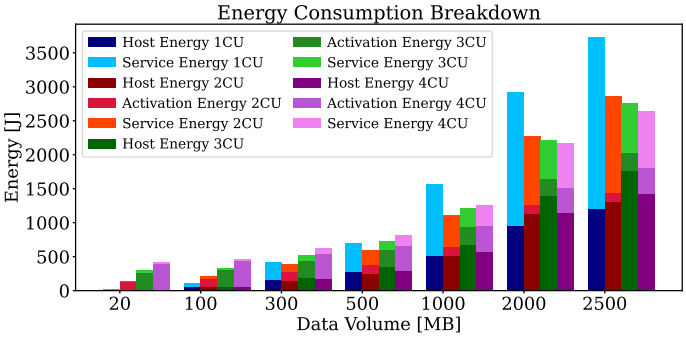

Fig. 2: Energy details

Figure 2 shows the energy consumption results for each experiment. Specifically, each bar represents the average value of the corresponding experiment with a specified number of MB requested. Each bar is divided into three sections, each corresponding to one of the three energy consumption metrics discussed in Section III-B. As expected, the energy consumption of each metric increases linearly with the amount of data requested, a trend confirmed by several other studies [24]. From this plot, we can intuitively observe that for certain ranges of requested data volume, there is an optimal number of CUs that consumes less energy compared to other configurations.

Building on this intuition, Figure 3 provides a clearer visual representation of the previous observation. As shown, each line follows a linear slope, indicating that energy consumption is directly proportional to the data volume. Additionally, we observe "interception points" where the energy efficiency of using a particular number of CUs surpasses that of others. These interception points indicate the data volume ranges where a specific configuration minimizes energy consumption. This insight is critical for developing adaptive policies that dynamically select the optimal number of CUs based on the identified data volume ranges. For example, if the requested data volume falls within a range where using 2 CUs is more energy-efficient than using 1 or 3 CUs, the system can adjust in real-time or proactively by predicting future data volume requests. Such policies enable energy savings by operating within the most efficient configuration for the given workload, reducing unnecessary energy consumption while maintaining performance. We will describe our proof-of-concept policy in Section III-D.

To understand the reasons behind these empirical results, Figure 4 provides insights into the time duration changes for each data volume and the number of CU instances scheduled. The time taken by the 5G Network to transmit the total requested data (in MB) increases as the data volume grows. Additionally, allocating more CUs allows the infrastructure to complete the service in less time, thereby improving performance in

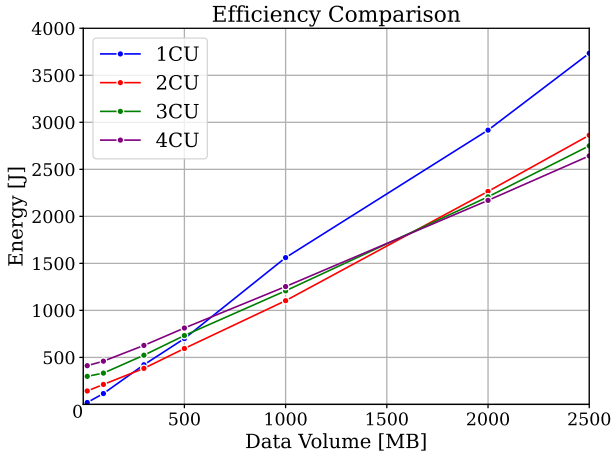

Fig. 3: Policy

any application case where latency is a critical QoS factor. While our work primarily focuses on optimizing the energy consumption of the overall system, this plot highlights the potential for achieving a trade-off between latency and energy consumption or, in other words, between QoS and resource allocation. As better described in Section III-D, we did not consider this trade-off in our policy manager for the moment, but we plan to explore it in future work.

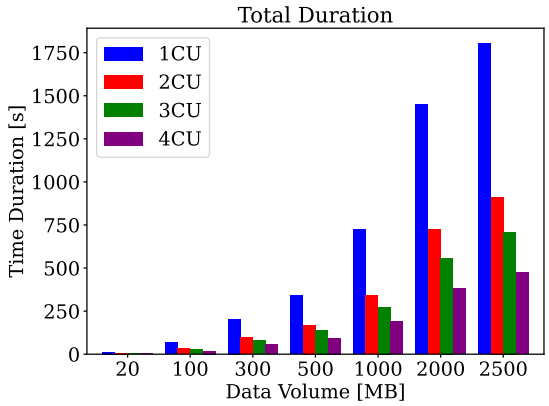

Fig. 4: Durations detail of the experiments

D. Policy Manager

Building on the insights derived from the results, we decided to implement a proof-of-concept scheduler based on our policy. Specifically, the policy dictates that the scheduler must allocate the optimal number of CUs, which is determined between consecutive intercept points based on the requested real-time data volume. This policy manager is executed on the rApp as illustrated in 1. The rApp dynamically adjusts the number of CU instances based on Data Volume advertisements received from the MEC server, as outlined in Algorithm 1. Upon completion of the configuration process, the rApp signals the MEC server to continue data transmission. For instance, if the requested data volume falls within the range of 0 to 300 MB, the scheduler will allocate 1 CU; if the range is from 300 to 576 MB, it will allocate 2 CUs, and so on.

Algorithm 1 Dynamic CU Allocation Policy

Require: Real-time data volume V (in MB)
CU allocation thresholds $\{T_0, T_1, \ldots, T_n\}$,
corresponding CU allocations $\{CU_1, CU_2, \ldots, CU_{n+1}\}$
Ensure: Allocated number of CUs $CU_{allocated}$
1: Find k such that $T_k \leq V < T_{k+1}$
2: $CU_{allocated} \leftarrow CU_{k+1}$
3: **return** $CU_{allocated}$

E. Benchmarks

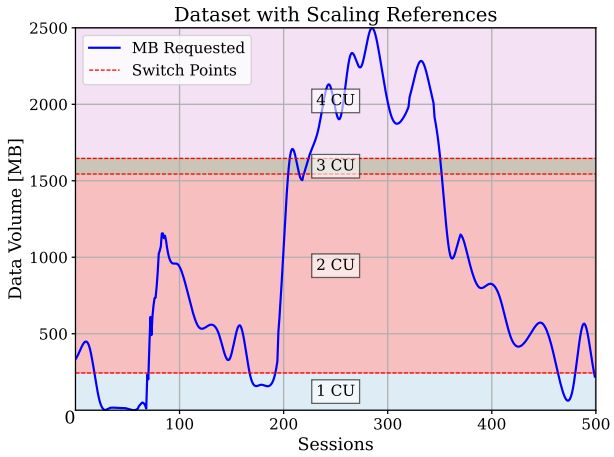

Fig. 5: Dataset

To test the scheduler, we selected a real dataset [25], which is a 5G trace of data collected from a major Irish mobile operator. The dataset includes two mobility patterns (static and car) and two application patterns (video streaming and file download). For our experiment, we choose the file download trace in the static scenario. In Figure 5, the data volume over the entire duration of the dataset is shown. Instead of considering each MB requested at every time unit, which would analyze the latency implications of our scheduler–a concern not intended for this work–we map the data volume to a single "session", representing the entire duration of the corresponding data transfer. To make the test feasible, we normalize the MB values between 0 and the maximum data volume that we encounter in our experiments; however, this does not affect the overall behavior of the curve. We conducted the test using the policy described in 1 and calculated the total energy consumed to complete all sessions by adding the average energy consumed by the entire infrastructure during each session, based on the experiments outlined in Section III-A, for each corresponding number of CUs allocated.

In Figure 6, we present a comparison between our dynamic scaling policy and the static one, where a fixed number of CUs is deployed across all sessions. As shown, our scheduler reduces

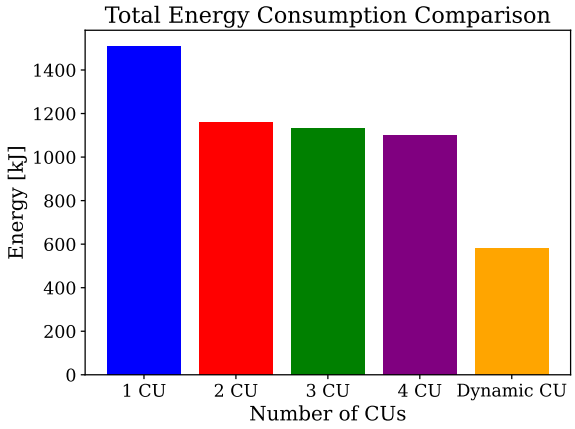

Fig. 6: Benchmark

energy consumption by approximately 53% compared to the static deployment with 4 CUs and up to around 60% compared to the worst-case scenario (the static deployment with 1 CU).

IV. Conclusions & Future Work

In this paper, we delve deeper into addressing the fundamental question: *How can resources be optimally allocated within the O-RAN 5G framework to minimize energy consumption?* We also presented a novel approach to minimizing energy consumption in 5G disaggregated RANs by dynamically scaling the CU components. Through experimental evaluation on a real testbed, we demonstrated that energy consumption is directly influenced by the number of CUs deployed and the data volume requested, which can significantly vary at runtime. By developing a dynamic CU allocation policy, we showed that energy savings of up to 60% can be achieved compared to static configurations. In the future, we plan to expose our system to more complex scenarios with less controlled environments while integrating an AI-driven method to forecast traffic patterns and predict the transmitting duration of the UEs. This will enable proactive scaling of the CU component, further enhancing energy optimization.

References

[1] Charlotte Freitag et al., "The real climate and transformative impact of ICT: A critique of estimates, trends, and regulations," 2021.

[2] L. M. P. Larsen, H. L. Christiansen, S. Ruepp, and M. S. Berger, "Toward greener 5g and beyond radio access networks—a survey," *IEEE Open Journal of the Communications Society*, 2023.

[3] A. M. Abdalla, J. Rodriguez, I. Elfergani, and A. Teixeira, *Energy Efficiency in the Cloud Radio Access Network (C-RAN) for 5G Mobile Networks*, 2019.

[4] L. M. P. Larsen, A. Checko, and H. L. Christiansen, "A Survey of the Functional Splits Proposed for 5G Mobile Crosshaul Networks," *IEEE Communications Surveys & Tutorials*, 2019.

[5] "Ieee standard for local and metropolitan area networks part 16: Air interface for fixed broadband wireless access systems," *IEEE Std 802.16-2004 (Revision of IEEE Std 802.16-2001)*, 2004.

[6] M. e. a. Ghoraishi, "Begreen: Beyond 5g energy efficient networking by hardware acceleration and ai-driven management of network functions," 2023.

[7] A. El-Amine, M. Iturralde, H. A. Haj Hassan, and L. Nuaymi, "A distributed q-learning approach for adaptive sleep modes in 5g networks," in *2019 IEEE Wireless Communications and Networking Conference (WCNC)*, 2019.

[8] F. Kooshki, A. G. Armada, M. M. Mowla, A. Flizikowski, and S. Pietrzyk, "Energy-efficient sleep mode schemes for cell-less ran in 5g and beyond 5g networks," *IEEE Access*, 2023.

[9] L. M. Moreira Zorello, M. Sodano, S. Troia, and G. Maier, "Power-efficient baseband-function placement in latency-constrained 5g metro access," *IEEE Transactions on Green Communications and Networking*, 2022.

[10] H. Hojeij, M. Sharara, S. Hoteit, and V. Vèque, "Dynamic placement of o-cu and o-du functionalities in open-ran architecture," in *2023 20th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*, 2023.

[11] A. Mudvari, N. Makris, and L. Tassiulas, "Ml-driven scaling of 5g cloud-native rans," in *2021 IEEE Global Communications Conference (GLOBECOM)*, 2021.

[12] R. Singh, C. Hasan, X. Foukas, M. Fiore, M. K. Marina, and Y. Wang, "Energy-efficient orchestration of metro-scale 5g radio access networks," in *IEEE INFOCOM 2021 - IEEE Conference on Computer Communications*, 2021.

[13] Z. Zhu, H. Li, Y. Chen, X. Wen, Z. Lu, and L. Wang, "Joint base station sleeping and functional split orchestration in crosshaul-based v-ran," in *2023 IEEE Wireless Communications and Networking Conference (WCNC)*, 2023.

[14] R. Joda, T. Pamuklu, P. E. Iturria-Rivera, and M. Erol-Kantarci, "Deep reinforcement learning-based joint user association and cu–du placement in o-ran," *IEEE Transactions on Network and Service Management*, 2022.

[15] H. Li, P. Li, K. D. Assis, A. Aijaz, S. Shen, R. Nejabati, S. Yan, and D. Simeonidou, "NetMind: Adaptive RAN Baseband Function Placement by GCN Encoding and Maze-solving DRL," in *2024 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2024.

[16] A. A. A. Rage, N. Wang, and R. Tafazolli, "Nfscaler: Ai-powered 5g-and-beyond network function scaler for qos assurance and energy efficiency," in *2024 IEEE 10th International Conference on Network Softwarization (NetSoft)*, 2024.

[17] H. Li, A. Emami, K. D. R. Assis, A. Vafeas, R. Yang, R. Nejabati, S. Yan, and D. Simeonidou, "Drl-based energy-efficient baseband function deployments for service-oriented open ran," *IEEE Transactions on Green Communications and Networking*, 2024.

[18] N. K. Shankaranarayanan, Z. Li, I. Seskar, P. Maddala, S. Puthenpura, A. Stancu, and A. Agarwal, "Poet: A platform for o-ran energy efficiency testing," in *2024 IEEE 100th Vehicular Technology Conference (VTC2024-Fall)*, 2024.

[19] "Slices, a scientific instrument for the networking community," *Computer Communications*, vol. 193, pp. 189–203, 2022.

[20] N. Nikaein, M. K. Marina, S. Manickam, A. Dawson, R. Knopp, and C. Bonnet, "Openairinterface: A flexible platform for 5g research," 2014.

[21] B. Petit, "scaphandre," 2023. [Online]. Available: https://github.com/hubblo-org/scaphandre

[22] V. Gudepu, B. Chirumamilla, R. R. Tella, A. Bhattacharyya, S. Agarwal, L. Malakalapalli, C. Centofanti, J. Santos, and K. Kondepu, "Earnest: Experimental analysis of ran energy with open-source software tools," in *2024 16th International Conference on COMmunication Systems NETworkS (COMSNETS)*, 2024.

[23] M. Jay, V. Ostapenco, L. Lefevre, D. Trystram, A.-C. Orgerie, and B. Fichel, "An experimental comparison of software-based power meters: focus on cpu and gpu," in *2023 IEEE/ACM 23rd International Symposium on Cluster, Cloud and Internet Computing (CCGrid)*, 2023.

[24] D. López-Pérez, A. De Domenico, N. Piovesan, G. Xinli, H. Bao, S. Qitao, and M. Debbah, "A survey on 5g radio access network energy efficiency: Massive mimo, lean carrier design, sleep modes, and machine learning," *IEEE Communications Surveys Tutorials*, 2022.

[25] D. Raca, D. Leahy, C. J. Sreenan, and J. J. Quinlan, "Beyond throughput, the next generation: a 5G dataset with channel and context metrics," 2020.