Energy-Efficient Placement and Association in Disaggregated O-RAN

Hiba Hojeij*, Sahar Hoteit*†, Véronique Vèque*, Alexis I. Aravanis*
*Université Paris-Saclay, CentraleSupélec, CNRS, L2S, Gif-sur-Yvette, France
†Institut Universitaire de France (IUF), France
Name.surname@centralesupelec.fr

***Abstract*—In modern Open RAN architectures, the traditional gNB protocol stack is disaggregated into virtualized components, Centralized Unit (CU), Distributed Unit (DU), and Radio Unit (RU), which are flexibly deployed across the infrastructure to meet diverse QoS requirements. This paper proposes an energy-aware model that jointly optimizes user association and virtual network function (VNF) placement to minimize total system energy consumption. By dynamically consolidating workloads and selectively activating radio and compute resources, the model reduces energy usage without compromising service constraints. We formulate the problem as an integer linear problem (ILP) to obtain optimal solutions and introduce a Graph Neural Network (GNN)-based heuristic that closely approximates optimal placements in real time. Simulation results demonstrate up to 75% energy savings at low loads and show the GNN reduces execution time by over 99% while maintaining near-optimal performance.**

***Index Terms*—Open RAN, resource allocation, UE association, CU/DU placement, ILP, GNN, energy efficiency.**

I. Introduction

As 5G and Beyond 5G (B5G) technologies emerge, new architectural paradigms are required to meet increasingly stringent demands for high data rates, low latency, and massive device connectivity. In this context, the Open Radio Access Network (O-RAN) initiative has emerged as a transformative approach, grounded in the principles of openness, disaggregation, and intelligence [1].

O-RAN decomposes the traditional base station into Centralized Unit (CU), Distributed Unit (DU), and Radio Unit (RU), deployed flexibly over the O-Cloud at edge and regional levels 1. This introduces new orchestration challenges, notably energy-aware VNF placement and UE–RU association under strict QoS. A key issue is the efficient placement of virtual network functions (VNFs) on edge and regional cloud servers, as well as users to RU association, while ensuring that strict Quality of Service (QoS) constraints are satisfied for various service types such as enhanced Mobile Broadband (eMBB), Ultra-Reliable Low-Latency Communication (URLLC), and massive Machine-Type Communication (mMTC).

While maximizing user satisfaction is critical for efficient O-RAN operation, the ever-growing energy demands of dense, cloud-native RAN deployments now make sustainability a first-order concern for operators. With the RAN domain responsible for up to 80% of network-wide energy consumption [2], [3], improving energy efficiency has become a key objective in the design and orchestration of future Open RAN systems. Reducing energy consumption in O-RAN has gained attention at both the computing and transport layers. GreenRAN [4] introduces a scalable placement framework using metaheuristics, while [5] and [6] tackle compute and transport energy optimization under QoS constraints using ILP and DRL. The recent model in [7] proposes a comprehensive MILP framework incorporating VNF migration, server, and transport energy, yet it lacks user-level association. In this paper, we fill this gap by integrating energy-aware server and RU management within a full-stack QoS-compliant orchestration framework. We address the problem of joint energy-aware placement and association in O-RAN by jointly optimizing CU/DU placement and User equipment (UE)-to-RU association to minimize the overall system energy consumption. Our model incorporates both computing (CUs, DUs) and radio (RUs) power costs, providing a holistic framework for sustainable O-RAN deployments. We exploit (i) workload consolidation at the O-Cloud layer to reduce the number of active servers and idle resources, and (ii) Physical Resource Block (PRB) blanking at the radio access level, which allows RUs to deactivate unused RBs and enter energy-saving sleep modes without compromising user QoS. We formulate the joint energy minimization problem as an Integer Linear Programming (ILP) model that captures the interdependencies between placement, association, resource allocation, and energy consumption, subject to slice-specific QoS constraints. To overcome the high computational complexity of optimal ILP solutions and support scalable, real-time orchestration, we propose a Graph Neural Network (GNN)-based heuristic that closely approximates optimal performance at a fraction of the execution time.

The rest of the paper is organized as follows: The system model and our proposed ILP-based solutions are described in Section II and III, respectively. Section V details the simulation framework and illustrates the performance evaluation of the proposed algorithms. Finally, Section VI concludes the paper.

II. System Model

Let $\mathcal{R}$ denote the set of RUs deployed across a geographical area of side L, where each RU $r \in \mathcal{R}$ is positioned at $P_r = (X_r, Y_r) \in [0, L]^2$. UEs, denoted by the set $\mathcal{U}$, are arbitrarily distributed within the same area, each located at $P_u = (X_u, Y_u) \in [0, L]^2$. The O-Cloud network is modeled as a graph $\mathcal{G} = (\mathcal{H}, \mathcal{E})$, with vertex set $\mathcal{H}$ representing cloud

978-3-903176-75-1 ©2025 IFIP

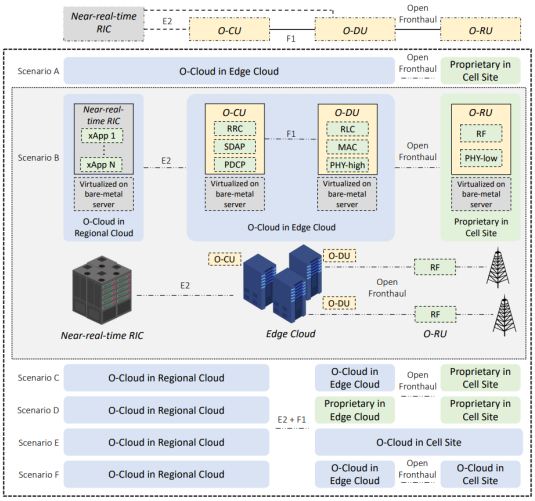

Fig. 1: O-RAN Cloud deployment scenarios [8]

hosts and edge set $\mathcal{E}$ the physical links. Hosts are classified into edge-cloud ($\mathcal{H}_E$) and regional-cloud ($\mathcal{H}_R$) nodes, such that $\mathcal{H} = \mathcal{H}_E \cup \mathcal{H}_R$. Each host h has a location P_h, a computational capacity G_h, and a static power consumption P^h_{static}. Each UE u requests a service from the set $\mathcal{S}$ (e.g., eMBB, uRLLC, mMTC), with slice-specific Quality of Service (QoS) requirements: required data rate $\lambda_{s(u)}$ and maximum tolerable end-to-end delay D^{E2E}_u. The system ensures the requested QoS for admitted UEs.

A. UE-RU Association

Let $x^{\text{RU}}_{u,r} \in \{0,1\}$ denote the association of UE u to RU r. Each RU r has M_r physical resource blocks (RBs) available, partitioned among slices via integer variables $\rho_{r,s} \in \mathbb{Z}_+$, representing the RBs allocated to slice s at RU r. The RB demand of UE u when linked to RU r is in (1), with $\eta_{u,r}$ the rate per RB. The total RBs assigned to a UE is in u (2).

$$\text{RB}_{u,r} = \left\lceil \frac{\lambda_{s(u)}}{\eta_{u,r}} \right\rceil, \tag{1}$$

$$\text{RB}_u(\mathbf{x}^{\text{RU}}) = \sum_{r \in \mathcal{R}} \text{RB}_{u,r} \cdot x^{\text{RU}}_{u,r}, \quad \forall u \in \mathcal{U}. \tag{2}$$

B. DU-CU Placement

We adopt a per-service, per-RU, per-host placement granularity. For each RU r, service s, and host h:

- $x^{\text{DU}}_{r,s,h} \in \{0,1\}$: 1 if a DU instance for service s at RU r is placed on host h; 0 otherwise. (DUs can be placed only on edge-cloud hosts.)
- $x^{\text{CU}}_{r,s,h} \in \{0,1\}$: 1 if a CU instance for service s at RU r is placed on host h; 0 otherwise. (CUs may be placed at either edge or regional cloud hosts.)

C. O-Cloud Computation Model

Each cloud host has limited computational capacity and can host a finite number of functional unit instances. The computational cost of each DU and CU instance is measured in GOPS, as in [9], and depends on system parameters such as modulation, coding rate, antennas, and MIMO layers as shown in equation 3. Based on 3GPP functional splits and workload distribution, we assign 50% and 10% of the processing load to DU and CU, respectively. The total computational utilization per host is computed by aggregating the GOPS required by all active functionalities placed on it.

$$g^{DU/CU}_{r,s} = \frac{\alpha_{DU/CU}(3A + A^2 + M \cdot C \cdot L/3)}{10} \cdot \rho_{r,s}, \tag{3}$$

where α^{CU} and α^{DU} reflect the split of processing load based on functional split 7.2x and 2, as in [10]. We denote by M the modulation bits (i.e., the number of bits per symbol), C the coding rate, L the number of MIMO layers, A the number of antennas, and RB_u the number of resource blocks assigned to user u.

The total computing load on host h is:

$$g_h(\mathbf{x}^{\text{RU}}) = \sum_{r \in \mathcal{R}} \sum_{s \in \mathcal{S}} \left(g^{\text{DU}}_{r,s} x^{\text{DU}}_{r,s,h} + g^{\text{CU}}_{r,s} x^{\text{CU}}_{r,s,h} \right). \tag{4}$$

D. End-to-End (E2E) Delay Model

The end-to-end (E2E) delay experienced by a given UE is mainly determined by the propagation delay between the deployed functional units, which comprises two major components: Midhaul (MH) delay and Fronthaul (FH) delay.

For each UE u, associated to RU r and requesting service s, we define the E2E delay as:

$$d^{\text{E2E}}_u = d^{\text{FH}}_u + d^{\text{MH}}_u, \tag{5}$$

where, for each UE u, the FH delay, i.e., from the DU to the RU, is defined as:

$$d^{\text{FH}}_u = \sum_{h \in \mathcal{H}_E} \frac{\|P_r - P_h\|}{v_{\text{Fiber}}} \cdot x^{\text{DU}}_{r,s,h}, \tag{6}$$

Similarly, for each UE u, the MH delay is measured between the CU to the DU, and is given by:

$$d^{\text{MH}}_u = \sum_{h \in \mathcal{H}_E} \sum_{h' \in \mathcal{H}} \frac{\|P_h - P_{h'}\|}{v_{\text{Fiber}}} \cdot x^{\text{DU}}_{r,s,h} \cdot x^{\text{CU}}_{r,s,h'}, \tag{7}$$

where v_{Fiber} is the propagation speed of light in fiber.

E. Computing-level Energy Model

To capture server-level energy consumption, we introduce binary activation variables $I_h \in \{0,1\}$ for each host h, denoting whether the server is powered on or off. The total computing power consumption P_{comp} comprises both dynamic and static energy components as in [4]:

$$P_{\text{comp}} = \sum_{h \in \mathcal{H}_E} P^{\text{edge}} \cdot \frac{1}{G_h} \cdot \sum_{r \in \mathcal{R}} \sum_{s \in \mathcal{S}} \left(g^{DU}_{r,s} x^{DU}_{r,s,h} + g^{CU}_{r,s} x^{CU}_{r,s,h} \right) + \sum_{h \in \mathcal{H}_R} P^{\text{regional}} \cdot \frac{1}{G_h} \cdot \sum_{r \in \mathcal{R}} \sum_{s \in \mathcal{S}} g^{CU}_{r,s} x^{CU}_{r,s,h} + \sum_{h \in \mathcal{H}} P^h_{\text{static}} \cdot I_h, \tag{8}$$

where, $P^{\text{edge}}$and P^{regional} represent the dynamic energy consumption per GOPS of processing load at edge and regional servers, respectively, and P^{h}_{static} is the static power consumption term that accounts for the baseline energy required to maintain host h operationally active. It is worth noting that in (8), the placement is per-service, which better reflects deployment. For instance, $x^{CU}_{r,s,h}$ denotes the placement of a CU instance handling service s at RU r on host h.

F. RU-level energy model

At the RAN level, we extend the energy model by integrating RU activation. A binary variable $I_r \in \{0,1\}$ is defined for each RU r, indicating whether it is active. RU activation depends on the association of at least one user to RU r. The RU energy model is formulated, inspired by [11], as follows:

$$P_{\text{RU}} = \sum_{r\in\mathcal{R}} \left(P^{\text{active}}_{\text{RU}} \cdot I_r + P^{\text{sleep}}_{\text{RU}} \cdot (1 - I_r) \right), \tag{9}$$

G. Transpot-level Energy model

We model the transport network's energy consumption considering both midhaul (MH) and fronthaul (FH) links, which are essential in disaggregated O-RAN deployments. Specifically, the energy consumption includes two components: a fixed component due to link activation, and a variable component proportional to the actual bandwidth (BW) utilization of each activated link [12]. The total transport network energy consumption, P_{link}, is defined as follows:

$$P_{\text{MH}} = \sum_{h\in\mathcal{H}_E} \sum_{h'\in\mathcal{H}_R} \left(u^{\text{MH}}_{h,h'} P^{\text{fix}}_{\text{net},h,h'} + \frac{b^{\text{MH}}_{h,h'}}{C^{\text{MH}}_{h,h'}} P^{\text{max}}_{\text{net},h,h'} \right), \tag{10}$$

$$P_{\text{FH}} = \sum_{r\in\mathcal{R}} \sum_{h\in\mathcal{H}_E} \left(u^{\text{FH}}_{r,h} P^{\text{fix}}_{\text{net},r,h} + \frac{b^{\text{FH}}_{r,h}}{C^{\text{FH}}_{r,h}} P^{\text{max}}_{\text{net},r,h} \right), \tag{11}$$

$$P_{\text{link}} = P_{\text{MH}} + P_{\text{FH}}, \tag{12}$$

where $u^{\text{MH}}_{h,h'}$ and u^{FH}_{h} are binary variables equal to 1 if the midhaul and fronthaul links are activated, respectively. The terms $P^{\text{fix}}_{\text{net},h,h'}$, $P^{\text{fix}}_{\text{net},h}$, $P^{\text{max}}_{\text{net},h,h'}$, and $P^{\text{max}}_{\text{net},h}$ represent the fixed and maximum load-dependent power consumption of midhaul and fronthaul links. The midhaul and fronthaul bandwidth utilizations $b^{\text{MH}}_{h,h'}$ and $b^{\text{FH}}_{r,h}$, are determined as in equations (13) and (14), respectively, where, considering Option-2 split, we model the MH link capacity required by UE u as follows

$$b^{\text{MH}}_{u}(\mathbf{x}) = \frac{(\text{IP} + H_{\text{PDCP}}) \cdot \text{TBS} \cdot N_{\text{TBS}} \cdot \text{RB}_u(\mathbf{x})}{(\text{IP} + H_{\text{PDCP}} + H_{\text{RLC}} + H_{\text{MAC}}) \cdot 1000}, \tag{13}$$

where, TBS represents the transport block size, N_{TBS} is the number of TBs per TTI, IP is the datagram size, and lastly, H_{PDCP}, H_{RLC} and H_{MAC} are the header size of PDCP, RLC, and MAC layers, respectively.

and considering Option-7.2x split, we define the FH link capacity required by UE u as

$$b^{\text{FH}}_{u}(\mathbf{x}^{RU}) = \frac{N_{\text{SYM}} \cdot N_{\text{SC}} \cdot N_{\text{IQ}} \cdot RB_u(\mathbf{x})}{1000}, \tag{14}$$

where N_{SYM} is the number of symbols per sub-frame, N_{SC} is the number of subcarriers per RB, and N_{IQ} is the number of I and Q bits.

III. Problem Definition

Our goal is to optimize O-RAN orchestration by minimizing the total energy consumption of the system, jointly considering the activation and placement of RUs, DUs, and CUs, subject to service requirements, resource capacities, and QoS constraints. We formulate the optimization problem as follows:

$$\min_{\mathbf{x},\,\rho,\,I_h,\,I_r,\,u} \quad P_{\text{total}} = P_{\text{comp}} + P_{\text{RU}} + P_{\text{link}} \tag{15}$$

$$\text{s.t.} \quad \sum_{r\in\mathcal{R}} x^{\text{RU}}_{u,r} = 1, \forall u \in \mathcal{U} \tag{16}$$

$$\sum_{h\in\mathcal{H}} x^{\text{CU}}_{r,s,h} = \sum_{h\in\mathcal{H}_E} x^{\text{DU}}_{r,s,h} = 1, \forall r \in \mathcal{R},\ \forall s \in \mathcal{S} \tag{17}$$

$$x^{\text{DU}}_{r,s,h} = 0, \forall r, s,\ h \in \mathcal{H}_R \tag{18}$$

$$x^{\text{RU}}_{u,r} \le I_r, \forall u \in \mathcal{U},\ \forall r \in \mathcal{R} \tag{19}$$

$$\sum_{s\in\mathcal{S}} \rho_{r,s} \le M_r, \forall r \in \mathcal{R} \tag{20}$$

$$\sum_{u\in\mathcal{U}_s} x^{\text{RU}}_{u,r}\, \text{RB}_{u,r} \le \rho_{r,s}, \forall r \in \mathcal{R},\ \forall s \in \mathcal{S} \tag{21}$$

$$g_h(\mathbf{x}) \le G_h, \forall h \in \mathcal{H} \tag{22}$$

$$d^{\text{E2E}}_{u} \le D^{\text{E2E}}_{u}, \qquad \forall u \in \mathcal{U} \tag{23}$$

$$I_h \ge x^{\text{DU}}_{r,s,h} + x^{\text{CU}}_{r,s,h}, \forall r, s, h \tag{24}$$

$$I_r \ge \sum_{u\in\mathcal{U}} x^{\text{RU}}_{u,r} / |\mathcal{U}|, \forall r \tag{25}$$

$$x^{\text{CU}}_{r,s,h} \in \{0,1\},\ x^{\text{DU}}_{r,s,h} \in \{0,1\}, \forall r, s, h \tag{26}$$

$$I_r,\ I_h,\ x^{\text{RU}}_{u,r},\ u^{\text{MH}}_{h,h'},\ u^{\text{FH}}_{r,h} \in \{0,1\}, \forall r, h, h' \tag{27}$$

$$\rho_{r,s} \in \mathbb{Z}_+, \forall r, s \tag{28}$$

First, constraint (16) guarantees that each UE is associated with exactly one radio unit (RU). The uniqueness of network function instances is assured by constraint (17), which ensures that for every RU and service pair, there is exactly one CU (placed on any host) and one DU instance (placed on an edge host). Constraint (19) requires that a UE may only be associated with an RU if that RU is active. Radio resource allocation is guaranteed by two constraints: (20) ensures that the total number of physical RBs allocated to all network slices at each RU does not exceed the RU's capacity, while (21) enforces that the RBs used by UEs in a particular slice do not exceed the RB allocation for that slice at each RU. Constraint (22) enforces that the aggregated computing load from CU and DU instances placed on any host does not surpass the host's processing capacity, and this is only considered if the host is active. Constraint (18) restricts DU placement to edge-cloud hosts, reflecting architectural requirements. The system must also guarantee that the end-to-end delay for each

UE does not exceed the user-specific maximum, as imposed by constraint (23). Host activation logic is formalized in constraint (24), where a host is considered active if any DU or CU is placed on it. Similarly, constraint (25) ensures that an RU is marked as active if it serves at least one UE. Variable domains are enforced by constraints (26), (27), and (28), specifying that placement, activation, and link usage variables are binary, while RB allocation variables are integer variables.

IV. Graph Neural Network based Heuristic

We design a Graph Neural Network (GNN) model to predict energy-efficient user associations and server placements in an O-RAN system [13]. The motivation for using a GNN comes from the natural graph-like structure of O-RAN networks [14]. In our setup, users are connected to RUs, and these RUs are connected to servers that can host their CU and DU functions. This creates a system of nodes and edges that is well-suited for a graph-based learning approach. he graph has UE, RU, and server nodes, connected via UE–RU and RU–server edges. Each node is assigned a feature vector. User nodes include features like location, service type, delay requirement, and computing demand. BS and server nodes include their location, server type (edge or regional), and computing capacity. To enrich each node with global information about the whole network, we include values from the graph's Laplacian spectrum as part of every node's input features.

The GNN consists of a single heterogeneous convolution layer that updates each node's features based on its neighbors. We use the SAGEConv operator to allow flexible and scalable learning across the different types of nodes. After the convolution, we apply batch normalization and a non-linear activation to each node's updated features. The model is trained to solve three tasks: (1) predict which RU each user should connect to, (2) predict the placement of CU for each service at each RU, and (3) predict the same for the DU. For these tasks, we use separate linear classifiers. The training dataset is obtained from our optimal solutions, having the objective of energy consumption minimization. This GNN learns to approximate the optimal resource allocation policy with much lower computation time. Thanks to the graph structure and the use of spectral features, it can generalize well across different user distributions and network loads, making it a practical tool for energy-aware orchestration in Open RAN [15].

V. Simulation Framework and Evaluation

We build our simulation setup based on the same network topology proposed in [16], [17], while ensuring that both radio and computing resources are provisioned to have an underloaded system, where all UEs can be admitted. Otherwise, no energy gains can be achieved. It consists of 4 RUs, distributed across a square area of side 1 km. The UEs are scattered within the defined area uniformly at random. The system employs a 20-MHz bandwidth, resulting in 100 RBs available per TTI at each RU. Additional radio parameters include four antennas, two MIMO layers, and 64-QAM modulation. The number of UEs varies from 20 to 100. Users belong to different slices, including enhanced Mobile Broadband (eMBB), ultra-Reliable Low Latency Communication (uRLLC), and massive Machine Type Communication (mMTC), following the distribution in [10] for an industrial area where 25% of users are eMBB users, 25% are uRLLC users, and 50% are mMTC users, with data rate requirements of 20 Mb/s, 5 Mb/s, and 1 Mb/s, respectively. The MH delay bounds D_u^{MH} are drawn from $[100, 300]$ µs for uRLLC, 500 µs for eMBB, and 1000 µ for mMTC. The FH delay bounds are set to 100 µs for all service types [10]. We consider a set of 3 edge-cloud nodes, such that the distance between any pair of edge-cloud nodes and RUs is between 5-10 km. Moreover, we consider 1 regional-cloud node randomly located within 40-80 km away from the edge-cloud nodes. The O-Cloud setup is in line with the specification in [18]. The computational capacity G_h is set to 350 GOPS for edge-cloud servers, and 1000 GOPS for the regional-cloud node. We use parameter values inspired by [4] and [11]. The power-related settings are summarized in Table I.

TABLE I: Power Consumption Parameters

Parameter	Value (W)
Static power per edge server ($P_{\text{static}}^{\text{edge}}$)	120
Static power per regional server ($P_{\text{static}}^{\text{regional}}$)	200
RU power (active mode) ($P_{\text{RU}}^{\text{active}}$)	397
RU power (sleep mode) ($P_{\text{RU}}^{\text{sleep}}$)	40

Initial evaluation results demonstrate significant energy savings achieved through our proposed energy-aware orchestration strategies, as compared to a naïve baseline where all RUs and servers (both edge and regional) remain constantly active.

To quantify the benefits of energy-aware orchestration, we first consider the *Optimal* model, which jointly optimizes user association and VNF placement while minimizing total energy consumption. The energy saving gain is computed as:

$$\text{Gain} = 100 \cdot \frac{E_{\text{Baseline}} - E_{\text{scheme}}}{E_{\text{Baseline}}} \tag{29}$$

As shown in Figure 2, the *Optimal* model achieves substantial energy savings across all system loads. The *Optimal* model achieves significant energy savings, with up to 75% reduction at low load (20 users). The observed savings decrease with an increasing number of users, reaching 35% savings even under full system load (100 users). We then evaluate the performance of our proposed *GNN heuristic* by comparing its energy savings against the *Optimal* model. The GNN is designed to approximate the optimal placement and association decisions in real time, with minimal computational overhead. Figure 2 also shows the percentage of energy savings achieved by the GNN heuristic. The *GNN heuristic* effectively mimics this energy-saving behavior, closely replicating the optimal performance at all user densities. Specifically, it reaches 67% saving at 20 users, with only a marginal optimality gap of under 5%. Even with higher user densities (80–100 users), the GNN sustains a 25% energy reduction. The GNN model exhibits a notable increase in energy saving at a number of users equal to 80. This refers to the fact that the GNN may

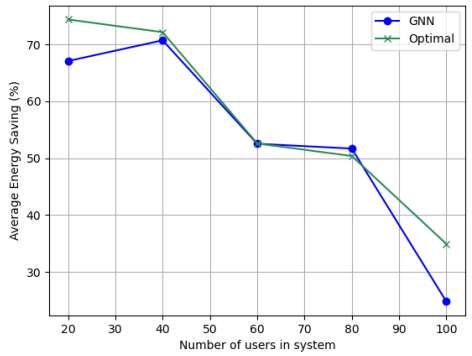

Fig. 2: Average energy saving (%) of the Optimal and GNN heuristic models compared to the baseline

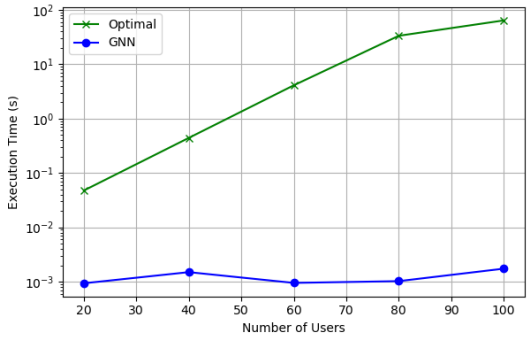

Fig. 3: Execution time of *Optimal* and *GNN heuristic* models

admit fewer users and underutilize certain RUs and servers due to suboptimal association or placement decisions. This phenomenon illustrates a trade-off between energy efficiency and user admittance, where the GNN sacrifices service performance to favor energy conservation. Despite this, the GNN model still demonstrates effective energy-aware behavior.

Figure 3 illustrates the execution time in seconds of the *Optimal* model versus the proposed *GNN heuristic* as the number of users increases. The results show the increase in computation time for the *Optimal* ILP solver, growing from milliseconds at low user counts to over 64 seconds for 100 users. This exponential increase reflects the complexity of solving the joint optimization problem. In contrast, the GNN-based heuristic maintains a nearly constant execution time across all user loads, with inference times order of 2 milliseconds. This lightweight inference cost makes the GNN heuristic especially attractive for real-time network control in O-RAN systems. The time reduction exceeds 99.9% compared to the optimal solver at high loads, confirming the GNN's practicality as a real-time dApp for scalable orchestration. Although the GNN may yield suboptimal or slightly infeasible decisions, its execution efficiency enables frequent model updates within the 1-ms TTI interval, crucial for dynamic and dense network environments.

VI. Conclusion

This paper presents an energy-efficient model for Open RAN, optimizing both placement and user association. Our ILP-based solution establishes an optimal benchmark, while the GNN heuristic achieves real-time approximation with minimal overhead. Simulation results confirm the effectiveness of our approach, with significant energy savings and fast execution. These findings support the feasibility of scalable, energy-aware orchestration in future O-RAN deployments.

VII. Acknowledgment

This PhD was funded by the ANR HEIDIS project (nb: ANR-21-CE25-0019; https://heidis.roc.cnam.fr)

References

[1] O-RAN Alliance, "O-RAN WhitePaper - Building the Next Generation RAN," https://www.o-ran.org/resources, October 2018.

[2] K. Technologies, "Energy efficiency of radio units in next-generation open radio access networks," Keysight Technologies, USA, Tech. Rep., September 2023. [Online]. Available: https://www.keysight.com

[3] M. Dryjański, "The o-ran whitepaper 2023–energy efficiency in o-ran," *O-RAN Alliance*, 2023.

[4] 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.

[5] N. Sen and A. F. A, "Towards energy efficient functional split and baseband function placement for 5g ran," in *2023 IEEE 9th International Conference on Network Softwarization (NetSoft)*, 2023, pp. 237–241.

[6] E. Amiri, N. Wang, M. Shojafar, and R. Tafazolli, "Energy-aware dynamic vnf splitting in o-ran using deep reinforcement learning," *IEEE Wireless Communications Letters*, vol. 12, no. 11, pp. 1891–1895, 2023.

[7] W. T. Pires, G. Almeida, S. Correa, C. Both, L. Pinto, and K. Cardoso, "Optimizing energy consumption for vran placement in o-ran systems with flexible transport networks," Jan. 2025. [Online]. Available: http://dx.doi.org/10.36227/techrxiv.173611601.16245000/v1

[8] L. Bonati, M. Polese, S. D'Oro, S. Basagni, and T. Melodia, "Open, programmable, and virtualized 5g networks: State-of-the-art and the road ahead," *Computer Networks*, vol. 182, p. 107516, 2020.

[9] E. Sarikaya and E. Onur, "Placement of 5g ran slices in multi-tier o-ran 5g networks with flexible functional splits," in *2021 17th International Conference on Network and Service Management (CNSM)*, 2021.

[10] S. Mondal and M. Ruffini, "Optical front/mid-haul with open access-edge server deployment framework for sliced o-ran," *IEEE Trans. on Network and Service Mngmt*, vol. 19, no. 3, 2022.

[11] M. Q. Usman, C. J. Sreenan, M. Dryjanski, and A. O'Driscoll, "Power modeling of the o-ran o-ru amp; application of advanced sleep modes for enhanced energy efficiency," Nov. 2024.

[12] N. Fryganiotis, E. Stai, I. Dimolitsas, A. Zafeiropoulos, and S. Papavassiliou, "Dynamic, reconfigurable and green network slice admission control and resource allocation in the o-ran using model predictive control," in *IFIP Networking Conference*. IEEE, 2024.

[13] W. Jiang, "Graph-based deep learning for communication networks: A survey," *Computer Communications*, vol. 185, pp. 40–54, 2022.

[14] P. Tam and S. Kim, "Graph-based learning in core and edge virtualized o-ran for handling real-time ai workloads," *IEEE Transactions on Network Science and Engineering*, 2025.

[15] Q. Siyu, L. Shuopeng, L. Shaofu, C. Ken *et al.*, "Energy-efficient vnf deployment for graph-structured sfc based on graph neural network and constrained deep reinforcement learning," in *Asia-Pacific Network Operations and Management Symposium (APNOMS)*. IEEE, 2021.

[16] H. Hojeij, G. I. Ricardo, M. Sharara, S. Hoteit, V. Vèque, and S. Secci, "On flexible association and placement in disaggregated ran designs," *Computer Communications*, 2025.

[17] ——, "Flexible association and placement for open-ran," in *IEEE INFOCOM WKSHPS*, 2024.

[18] 3GPP, "Technical Specification Group Radio Access Network; NR; Physical Layer Procedures for Data," Tech. Rep., 2019.