Multi-Resource Orchestration and Energy-Aware VNF Placement for Open RAN

Hiba Hojeij, Sahar Hoteit, Véronique Vèque

Multi-Resource Orchestration and Energy-Aware VNF Placement for Open RAN

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

***Abstract*—This paper summarizes the PhD research work conducted on resource orchestration in disaggregated Open RAN (O-RAN) architectures. In O-RAN, the gNB radio protocol stack is split into virtualized functions— Centralized Unit (CU), Distributed Unit (DU), and Radio Unit (RU)—that are dynamically deployed across edge and regional cloud infrastructures. This architectural transformation introduces new challenges in managing computing, radio, and transport resources while satisfying diverse Quality of Service (QoS) requirements.**

Our research is structured into three major contributions. First, the problem of CU/DU placement is addressed under fixed User Equipment (UE)-to-RU associations through an Integer Linear Programming (ILP) formulation and a Bi-directional Long Short-Term Memory (Bi-LSTM)-based Recurrent Neural Network (RNN) heuristic for scalable inference. The second contribution extends the model to jointly optimize CU/DU placement and UE-to-RU association, with a decomposition heuristic and an enhanced RNN-based solution. The proposed models significantly outperform conventional baselines in terms of user admittance and runtime. Finally, an energy-aware extension is introduced to minimize total system power consumption by consolidating workloads and managing radio and computing resources activation aiming to enhance the sustainability of O-RAN orchestration while maintaining QoS guarantees.

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

I. Introduction

A. Thesis Context and Motivations

The Radio Access Network (RAN) forms a foundational component of wireless communication systems, bridging user devices and the core network. 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]–[3].

O-RAN decomposes the traditional base station into three logical, virtualized components: the Centralized Unit (CU), Distributed Unit (DU), and Radio Unit (RU). These functions are dynamically deployed over a cloud-native infrastructure, referred to as the O-Cloud, which is composed of heterogeneous computing nodes at the edge and regional levels, as shown in Figure 1. This architectural flexibility enables multi-vendor interoperability and paves the way for intelligent, service-aware orchestration strategies. In this context, my PhD thesis objectives consist of designing dynamic deployment strategies for O-RAN's disaggregated components, mainly CUs and DUs, through coordinated orchestration of computing, radio, and transport resources. This paper highlights the main contributions of my PhD thesis, published in 4 papers and an ongoing contribution yet to be submitted. Our works in [4] and [5] address the dynamic placement of CU and DU functions in O-RAN environments under diverse network conditions and constraints. A further development integrating joint UE-RU association with CU/DU placement is introduced in [6], and a detailed RNN-based heuristic for solving the joint problem with extensive performance evaluation was presented in [7].

B. Major Challenges and Problem Statement

The disaggregated nature of O-RAN architecture introduces new challenges in orchestrating heterogeneous resources across a distributed and virtualized infrastructure. A key issue is the efficient placement of CU and DU functions on edge and regional cloud servers, 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).

As a first contribution, we address the CU/DU placement problem under the assumption of fixed UE-to-RU association. The goal is to determine optimal placements that maximize the number of admitted UEs, subject to computing and latency constraints. This problem is modeled as an Integer Linear Programming (ILP) formulation and further tackled through a Bi-LSTM-based RNN heuristic that provides real-time approximations of the optimal placement decisions.

In the second contribution, we extend our work to jointly optimize CU/DU placement and UE-to-RU association, reflecting the full flexibility enabled by disaggregated O-RAN. The resulting joint problem introduces higher complexity due to the coupling between radio access and cloud placement decisions. To solve this issue, we propose a sequential decomposition heuristic that separates association and placement stages. An enhanced Bi-LSTM RNN model is also developed, capable of predicting both association and placement decisions.

Finally, in the third and ongoing contribution of this PhD project, energy efficiency is incorporated into the orchestration process. By modeling both computing and radio power consumption, the extended framework aims to minimize total system energy usage while preserving service-level guarantees.

C. Thesis Contributions

The main contributions of this PhD are as follows:

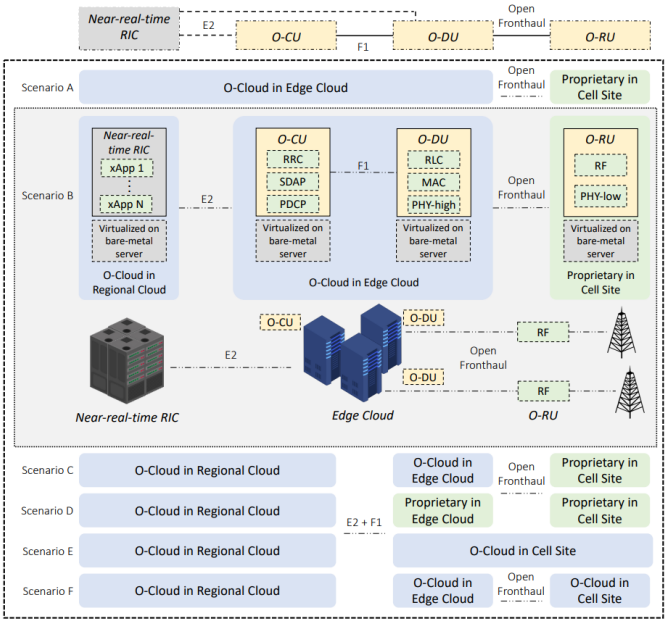

Fig. 1: O-RAN Cloud deployment scenarios [8]

- **CU/DU Placement Optimization:** An ILP model is formulated to optimize the placement of CU and DU functions under static UE-to-RU association. A Bi-LSTM-based RNN heuristic is developed to approximate optimal placement with real-time execution.
- **Joint Placement and Association:** The model is extended to jointly optimize CU/DU placement and UE-to-RU association. A decomposition heuristic and an enhanced RNN are proposed to address scalability and inference efficiency.
- **Energy-Aware Orchestration:** An ongoing extension integrates computing and radio energy consumption into the optimization framework to improve the sustainability of O-RAN deployments.

The remainder of this paper is organized as follows. Section II reviews related work on VNF placement, UE association, and energy efficiency in O-RAN. Section III details the first contribution. Section IV presents the extended joint placement and association model. Section VI evaluates the performance of the proposed models. Section VII describes the ongoing energy-aware orchestration work. Finally, Section VIII concludes the paper.

II. Related Work

A. VNF Placement and UE Association in O-RAN

Disaggregated O-RAN architectures have motivated extensive research on the placement of virtualized network functions (VNFs), such as Centralized Units (CUs) and Distributed Units (DUs), and the association of User Equipments (UEs) to Radio Units (RUs). Studies in [9], [10] optimize DU placement and radio resource allocation for energy efficiency and cost reduction but assume fixed CU locations, limiting flexibility. Reinforcement learning is used in [11] to jointly optimize CU/DU placement and UE-RU association, but it overlooks QoS differentiation. Other works, including [12]–[15], address functional split placement and deployment cost, yet they often lack dynamic user-level association or admission control mechanisms. While these models address parts of the orchestration problem, they do not jointly capture QoS-aware admission, multi-resource constraints, and dynamic UE association as comprehensively as our work.

B. Deep Learning-Based Solutions in RAN

Deep learning, particularly Recurrent Neural Networks (RNNs), has been applied to RAN for resource scheduling [16], [17], mobility prediction [18], and congestion control. In [19], a Bi-LSTM model is trained to approximate ILP solutions for centralized RAN placement. However, most works assume limited system constraints and focus on forecasting or control. In contrast, our RNN heuristic emulates an ILP model with full placement constraints in a disaggregated RAN setting, making it suitable for real-time orchestration under complex dependencies.

C. Energy Efficiency in Open RAN

Reducing energy consumption in O-RAN has gained attention at both the computing and transport layers. GreenRAN [20] introduces a scalable placement framework using metaheuristics, while [21] and [22] tackle compute and transport energy optimization under QoS constraints using ILP and DRL. The recent model in [23] proposes a comprehensive MILP framework incorporating VNF migration, server, and transport energy, yet it lacks user-level association. Our ongoing work fills this gap by integrating energy-aware server and RU management within a full-stack QoS-compliant orchestration framework.

III. Contribution 1 – CU/DU Placement Optimization

Our first contribution focuses on the flexible placement of CU and DU functions in an O-RAN infrastructure. In this work, UE-to-RU associations are assumed to be fixed and based on proximity. The goal of this contribution is to optimize CU and DU deployment decisions across heterogeneous edge and regional O-Cloud resources in order to maximize user admittance while respecting computing and delay constraints. This work is published in [4], expanded and comprehensively evaluated in [5].

A. System Topology

We consider a disaggregated Open RAN deployment composed of a set of UEs $\mathcal{U}$, a fixed set of RUs $\mathcal{R}$, and a hierarchy of O-Cloud servers $\mathcal{H}$, divided into edge ($\mathcal{H}_E$) and regional ($\mathcal{H}_R$) hosts. Hosts $\mathcal{H}$ are connected via physical links $\mathcal{E}$ forming the network graph $G(\mathcal{H}, \mathcal{E})$.

B. CU/DU Placement Model

DU functions are exclusively deployed at edge hosts, whereas CU functions can be placed at either edge or regional hosts. Binary decision variables $x^{DU}_{u,h}$ and $x^{CU}_{u,h}$ indicate whether DU and CU for UE u are deployed on host h, respectively.

The following constraint ensures that for each UE, exactly one DU and one CU are placed:

$$\sum_{h \in \mathcal{H}_E} x^{DU}_{u,h} = \sum_{h \in \mathcal{H}} x^{CU}_{u,h} = 1, \quad \forall u \in \mathcal{U}. \tag{1}$$

C. Computational Resource Constraints

Each host h has a limited processing capacity G_h (in GOPS). The processing load in GOPS of DU and CU placements per user u is computed as:

$$g_u^{FU} = \frac{\alpha^{FU}(3A + A^2 + M \cdot C \cdot L/3)}{10} \cdot RB_u, \quad (2)$$

where $FU \in \{CU, DU\}$ and $\alpha^{CU} = 10\%$, $\alpha^{DU} = 50\%$ reflect the split of processing load based on functional split 7.2x and 2, as in [15]. RB_u is the fixed number of Resource Blocks allocated to user u. 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 resource used at each host must not exceed its capacity:

$$\sum_{u \in \mathcal{U}} \left(g_u^{CU} \cdot x_{u,h}^{CU} + g_u^{DU} \cdot x_{u,h}^{DU} \right) \leq G_h, \quad \forall h \in \mathcal{H}. \quad (3)$$

D. Midhaul Delay Model

We consider that the delay experienced by a UE is attributed to midhaul (MH) (link between CU and DU) propagation delay.

For each UE u, the MH delay is computed as:

$$d_u^{\text{MH}}(x) \triangleq \sum_{h,h' \in \mathcal{H}} \frac{\|P_h - P_{h'}\|}{v_{\text{Fiber}}} \cdot x_{u,h}^{CU} \cdot x_{u,h'}^{DU}, \quad (4)$$

where v_{Fiber} denotes the propagation speed over fiber, and $\|\cdot\|$ is the Euclidean distance between the CU and DU hosts.

E. Our Proposed Solutions

Our goal is to maximize the number of admitted UEs, conditioned on the system's ability to meet their computing and delay requirements under fixed UE-to-RU associations. We formulate the CU/DU placement optimization as an ILP problem defined as:

$$\max_{\mathbf{x}} \sum_{u \in \mathcal{U}} a_u = \sum_{h,h' \in \mathcal{H}} \epsilon_{s(u)} \cdot x_{u,h}^{DU} \cdot x_{u,h'}^{CU} \quad (5)$$

subject to (1), (3) and (4)

where a_u indicates the admission of UE u, $\epsilon_{s(u)}$ is a priority weight defined for each slice requested by UE u, $s(u)$.

The placement optimization problem is first formulated as an ILP model to maximize user admittance while satisfying computing and delay constraints across the O-Cloud. While the ILP provides optimal solutions, its computational complexity grows rapidly with the number of users, limiting scalability. To address this issue, we develop a Recurrent Neural Network (RNN) based heuristic, trained to approximate the ILP output. The model focuses on predicting CU and DU placements per user based on system context.

The RNN is built using a Bi-directional Long Short-Term Memory (Bi-LSTM) architecture to capture per-user sequence dependencies. This RNN model is trained using optimal solutions generated by the ILP and tested over varied network instances. Each UE is encoded as a feature vector (position, service type, QoS requirements,..) while outputs are structured as one-hot encoded decisions for DU and CU placement.

IV. Contribution 2 – Joint Placement and UE-RU Association

In our second contribution, the placement model is extended to jointly optimize both CU/DU placement and UE-to-RU association. This extension reflects a more realistic and flexible orchestration scenario where user attachment to the RU is not static as before but driven by performance-aware decisions. The goal remains to maximize user admittance under the combined constraints of computing, radio, and transport resources. This work is published in [6] and further expanded with an RNN-based heuristic in [7].

A. System Model

The system topology remains the same as described in Section III, with a set of UEs $\mathcal{U}$, RUs $\mathcal{R}$, and O-Cloud hosts $\mathcal{H} = \mathcal{H}_E \cup \mathcal{H}_R$.

1) UE-RU Association Model: Each UE $u \in \mathcal{U}$ must be associated with exactly one RU $r \in \mathcal{R}$, indicated by binary variables $x_{u,r}^{RU}$. The number of Resource Blocks (RBs) required by UE u on RU r is determined by:

$$RB_{u,r} = \left\lceil \frac{\lambda_{s(u)}}{\eta_{u,r}} \right\rceil, \quad (6)$$

where $\lambda_{s(u)}$ is the data rate requirement for the service slice of user u, and $\eta_{u,r}$ is the achievable spectral efficiency from Shannon's theory [24].

2) Radio Resource Constraints: Each RU r can allocate at most M_r RBs per TTI. The newly added constraints are:

Single RU association per UE:

$$\sum_{r \in \mathcal{R}} x_{u,r}^{RU} = 1, \quad \forall u \in \mathcal{U} \quad (7)$$

Slice-based RB allocation per RU:

$$\sum_{s \in \mathcal{S}} \rho_{r,s} \leq M_r, \quad \forall r \in \mathcal{R} \quad (8)$$

RB sufficiency constraint:

$$RB_{u,r} \cdot x_{u,r}^{RU} \leq \rho_{r,s(u)}, \quad \forall u \in \mathcal{U}, r \in \mathcal{R} \quad (9)$$

3) Updated Placement Constraints: The association and placement variables are now dependent. Each user must have exactly one CU and DU placement, and one RU association:

$$\sum_{r \in \mathcal{R}} x_{u,r}^{RU} = \sum_{h \in \mathcal{H}_E} x_{u,h}^{DU} = \sum_{h \in \mathcal{H}} x_{u,h}^{CU}, \quad \forall u \in \mathcal{U} \quad (10)$$

The computing resource and delay models remain unchanged from Contribution 1 but now depend on the selected RU as well.

4) E2E Delay Model: The End-to-End (E2E) delay considered in this contribution for each user is the sum of midhaul (CU–DU) and fronthaul (DU–RU) propagation delays over fiber. The MH delay is already defined in (4). The FH delay is defined as follows:

$$d_u^{\text{FH}} = \sum_{r \in \mathcal{R}} \sum_{h \in \mathcal{H}} \frac{||P_r - P_h||}{v_{\text{Fiber}}} \cdot x_{u,h}^{DU} \cdot x_{u,r}^{RU}, \quad (11)$$

The total E2E delay must respect the maximum latency thresholds set per service.

B. Our Proposed Solutions

In this contribution, the orchestration framework is extended to jointly determine the UE-to-RU association and the placement of CU/DU functions per user. Our objective is to maximize the number of admitted UEs whose QoS and resource constraints can be satisfied. We formulate this by the following MILP model:

$$\max_{\mathbf{x},\rho} \sum_{u\in\mathcal{U}} a_u = \sum_{r\in\mathcal{R}} \sum_{h\in\mathcal{H}_E} \sum_{h'\in\mathcal{H}} \epsilon_{s(u)} \cdot x^{RU}_{u,r} \cdot x^{DU}_{u,h} \cdot x^{CU}_{u,h'} \quad (12)$$

subject to (3), (4), (7), (8), (9), (10), and (11).

To address its computational hardness, we implement the following solutions:

- An ILP formulation solved using CPLEX for optimal benchmarking.
- A sequential decomposition heuristic that solves the association and placement decisions in two stages.
- An extended Bi-LSTM Recurrent Neural Network (RNN) model that jointly predicts the UE-to-RU association and CU/DU placement per user. This model, presented in [7], uses a multi-label Bi-LSTM architecture to infer both association and placement decisions from the same input feature sequence.

V. Simulation Framework

The simulation setup, used across both contributions 1 and 2, comprises $|\mathcal{R}| = 4$ RUs distributed over a square area of 1 km^2. UEs are uniformly distributed within this area. The system operates with a 20 MHz bandwidth, yielding 100 RBs per Transmission Time Interval (TTI) at each RU. Radio parameters include 4 antennas, 2 MIMO layers, and 64-QAM modulation.

The number of UEs varies from 20 to 100, covering both underloaded and overloaded scenarios. UEs are assigned to network slices following the distribution in [15], with 25% eMBB, 25% uRLLC, and 50% mMTC users. The required data rates are set to 20 Mb/s, 5 Mb/s, and 1 Mb/s for eMBB, uRLLC, and mMTC slices, respectively. The number of required RBs per UE is calculated using Shannon's capacity formula, accounting for a distance-based path-loss model and a transmission power of 30 dBm. The O-Cloud infrastructure consists of $|\mathcal{H}_E| = 3$ edge cloud nodes located 5–10 km from the RUs, and one regional cloud node placed 40–80 km away. The computational capacity of edge hosts is uniformly sampled from 100 to 200 GOPS, while regional hosts range from 1000 to 2000 GOPS, in line with [13]. Latency constraints are slice-dependent, midhaul delay thresholds are randomly sampled from [100, 300] µs for uRLLC, set to 500 µs for eMBB, and 1000 µs for mMTC users. Fronthaul delay bounds are fixed at 100 µs across all slices, as in [15]. The ILP formulations are solved using IBM CPLEX [25] on a machine with an Intel® Core™ i9-11950H CPU and 16 GB RAM.

VI. Evaluation and Discussion

This section summarizes key evaluation results for the models developed throughout the aforementioned contributions. We assess the performance of the proposed ILP formulations and heuristic approximations under a variety of network conditions, resource limitations, and service requirements.

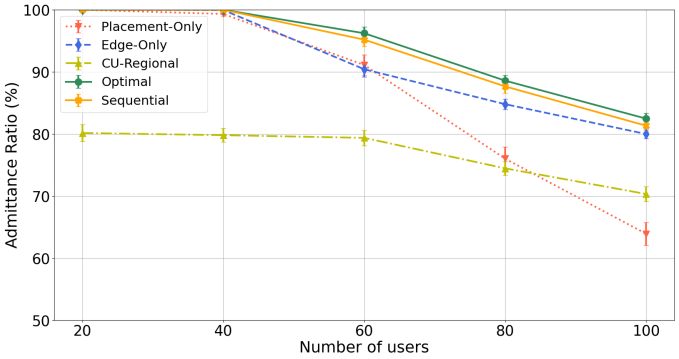

Fig. 2: Average admittance rate under varying number of users

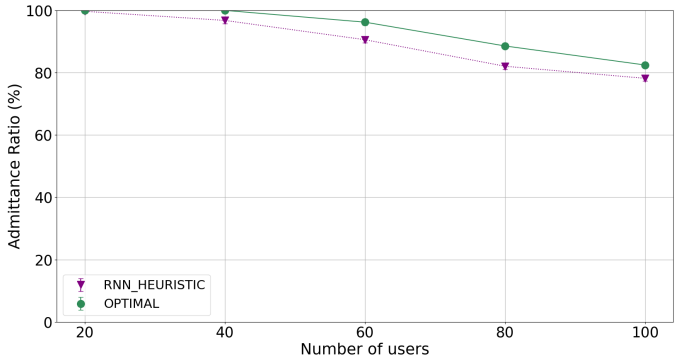

Fig. 3: Comparison of admittance ratio between the Optimal model and the RNN heuristic under varying number of users.

We evaluate the models presented in both Contribution 1—the placement optimization only referred to as **Placement-Only**, and Contribution 2—the joint placement and association and the sequential solution referred to as **Optimal** and **Sequential**, respectively, against the following baselines:

- **Edge-Only Model:** All CU and DU functions are statically placed on edge-cloud hosts. This setup mimics a fully distributed RAN (D-RAN) architecture.
- **CU-Regional Model:** CU functions are placed on the regional cloud host, while DU functions remain on edge hosts. UE-to-RU associations are dynamic. This model resembles a centralized RAN (C-RAN) setup.

We consider 100 instances of the previously described models by randomly varying UEs' (i) location and (ii) type of requested service. The averages are accompanied by error bars based on confidence intervals of 90%.

Figure 2 illustrates the average UE admittance ratio under varying user densities. The *Optimal* solution achieves the highest admission rate, while the *Sequential* model performs closely, with a gap of about 6%. The *Placement-Only* model (Contribution 1), lacking dynamic RU association, performs well under low load but degrades as system load increases. The *Edge-Only* and *CU-Regional* baselines show weaker performance due to resource overload and latency violations, respectively. The *RNN heuristic* model introduced in Contribution 2 is evaluated in terms of admission. Figure 3 shows that the *RNN heuristic* maintains a maximum optimality gap of 9% in admittance compared to the ILP solution.

In terms of computational efficiency, Figure 4 highlights

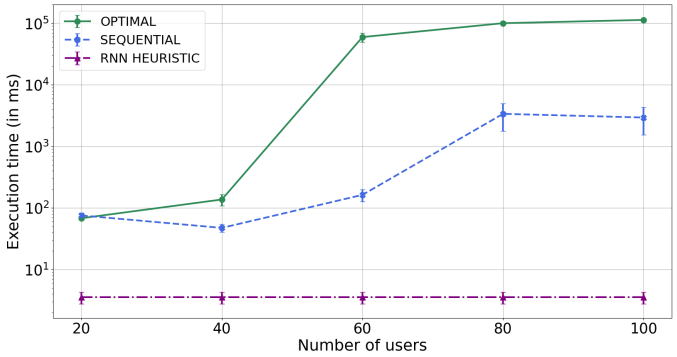

Fig. 4: Execution time (in ms) of the *Optimal*, *Sequential*, and *RNN heuristic* models as a function of the number of users

the advantage of the proposed RNN. While the ILP incurs a high runtime (over 100 seconds), and the *Sequential* approach reduces this to a few seconds, the *RNN heuristic* executes in just a few milliseconds regardless of the system load, making it suitable for integration within O-RAN for real-time decisions.

To conclude, *Contribution 1* enables flexible cloud deployment and improves over static baselines. *Contribution 2* further enhances performance through joint resource allocation, yielding higher admission. The *RNN heuristic* shows excellent scalability, offering real-time decisions with minimal optimality gap.

VII. Contribution 3 – Energy-Efficient Orchestration (Ongoing Work)

Our third contribution, currently in progress, focuses on enhancing the sustainability of O-RAN orchestration through energy-aware placement strategies. Unlike previous models that aim at maximizing user admittance, this work aims to minimize the total system energy consumption by jointly optimizing computing and radio energy costs. Given that RAN accounts for 70-80% of the total energy consumption in telecom networks [26], improving energy efficiency is crucial in O-RAN standardization. Accurate energy models can guide operators in selecting optimal deployment and configuration strategies while ensuring QoS guarantees.

Our current work aims to address energy reduction at both the server and RU levels. At the O-Cloud layer, energy savings are achieved by consolidating workloads to minimize active hosts and idle processing. At the radio access level, energy efficiency can be further enhanced through Physical Resource Block (PRB) blanking, allowing RUs to deactivate unused RBs and conserve power. Coupled with our dynamic UE-to-RU association model (Contribution 2), this technique can maximize RB aggregation and promote RU sleep modes without degrading QoS.

While our final objective is to account for all major components of energy consumption (including link transport energy, VNF migration cost, etc.), we currently focus on the integration of server-level and RU-level energy models.

A. Energy-Aware Optimization 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 [20]:

$$P_{\text{comp}} = \sum_{h\in\mathcal{H}_E} P^{\text{edge}} \cdot \frac{1}{G_h} \cdot \sum_{r\in\mathcal{R}} \sum_{s\in\mathcal{S}} \left(g^{DU}_{r,s}\, x^{DU}_{r,s,h} + g^{CU}_{r,s}\, x^{CU}_{r,s,h}\right) + \sum_{h\in\mathcal{H}_R} P^{\text{regional}} \cdot \frac{1}{G_h} \cdot \sum_{r\in\mathcal{R}} \sum_{s\in\mathcal{S}} g^{CU}_{r,s}\, x^{CU}_{r,s,h} + \sum_{h\in\mathcal{H}} P^{h}_{\text{static}} \cdot I_h, \tag{13}$$

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 (13), the placement decisions are defined on a per-service basis rather than per-user. This is more realistic from a deployment perspective, as server activation decisions can be made at the service level. For instance, $x^{CU}_{r,s,h}$ denotes the placement of a CU instance handling service s at RU r on host h.

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 [27], as follows:

$$P_{\text{RU}} = \sum_{r\in\mathcal{R}} \left(P^{\text{active}}_{\text{RU}} \cdot I_r + P^{\text{sleep}}_{\text{RU}} \cdot (1 - I_r)\right), \tag{14}$$

Our new objective is to minimize the total energy consumption of the system P_{total}:

$$\min \quad P_{\text{total}} = P_{\text{comp}} + P_{\text{RU}}. \tag{15}$$

subject to the constraints defined in previous contributions, refined to make placement per service per RU, ensuring service requirements and resource limits are still respected.

TABLE I: Power Consumption Parameters

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

B. Preliminary Framework and Evaluation

The evaluation of our energy-aware orchestration framework is currently in progress. To establish a baseline for performance, we consider the same network topology and simulation parameters described in Section V, while ensuring that both radio and computing resources are provisioned to have an under-loaded system, where all UEs can be admitted. Otherwise, no energy gains can be achieved. We use parameter values inspired by [20] and [27]. The power-related settings are summarized in Table I.

Initial evaluation results demonstrate significant energy savings when using the proposed energy-aware orchestration model, compared to a baseline scenario in which all servers and RUs are kept active. As illustrated in Figure 5, the energy saving decreases as the system load increases. The load is

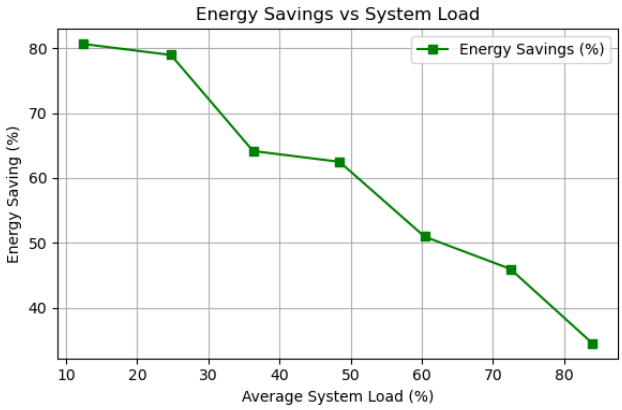

Fig. 5: Energy savings versus load.

computed in terms of the combined utilization of computing resources (GOPS) and radio resources (RBs). However, even at 85% average system load, our approach achieves more than 30% energy savings, thanks to the joint optimization of resource allocation, user association, and selective activation of infrastructure components.

Ongoing simulations will further quantify the energy savings under varying load conditions and extend the model to integrate transport link energy components and VNF migration costs, where we anticipate further improvements.

VIII. Conclusion

This paper summarizes the PhD dissertation and the associated four publications that address key challenges in resource allocation and orchestration for disaggregated Open RAN (O-RAN) networks. The first contribution has focused on optimizing the dynamic placement of CU and DU functions across edge and regional cloud nodes, formulated as an ILP problem and approximated via a Bi-LSTM-based heuristic to ensure real-time applicability. In the second contribution, the model was extended to jointly include UE-to-RU association. A sequential decomposition heuristic and an enhanced Bi-LSTM learning model were proposed to address scalability while preserving performance. These models were shown to significantly outperform baseline strategies in terms of user admittance and execution time. In our ongoing work, an energy-aware orchestration extension is introduced to minimize both computing and radio power consumption by consolidating workloads and activating only necessary infrastructure components. This work aims to contribute towards energy-aware network management in future O-RAN deployments.

IX. 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] M. Polese, L. Bonati, S. D'Oro, S. Basagni, and T. Melodia, "Understanding o-ran: Architecture, interfaces, algorithms, security, and research challenges," *IEEE Com. Surveys Tutorials*, 2023.

[3] L. Bonati, S. D'Oro, M. Polese, S. Basagni, and T. Melodia, "Intelligence and learning in o-ran for data-driven nextg cellular networks," *IEEE Communications Magazine*, 2021.

[4] H. Hojeij, M. Sharara, S. Hoteit, and V. Vèque, "Dynamic placement of o-cu and o-du functionalities in open-ran architecture," in *IEEE Inter. Conf on Sens, Comm, and Netw (SECON)*, Madrid, Spain, Sep. 2023.

[5] ——, "On flexible placement of o-cu and o-du functionalities in open-ran architecture," *IEEE Transactions on Network and Service Management*, 2025.

[6] H. Hojeij, G. I. Ricardo, M. Sharara, S. Hoteit, V. Vèque, and S. Secci, "Flexible association and placement for open-ran," in *IEEE INFOCOM 2024 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, 2024.

[7] ——, "On flexible association and placement in disaggregated ran designs," *Computer Communications*, 2025.

[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] T. Pamuklu, S. Mollahasani, and M. Erol-Kantarci, "Energy-efficient and delay-guaranteed joint resource allocation and DU selection in o-RAN," in *2021 IEEE 4th 5G World Forum (5GWF)*. IEEE, oct 2021.

[10] A. Ndao, X. Lagrange, N. Huin, G. Texier, and L. Nuaymi, "Optimal placement of virtualized dus in o-ran architecture," in *2023 IEEE 97th Vehicular Technology Conference (VTC2023-Spring)*, 2023, pp. 1–6.

[11] 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 Trans, on Net. and Service Mngmt*, 2022.

[12] M. Klinkowski, "Latency-aware du/cu placement in convergent packet-based 5g fronthaul transport networks," *Applied Sciences*, vol. 10, 2020.

[13] 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.

[14] I. Tamim, A. Saci, M. Jammal, and A. Shami, "Downtime-aware o-ran vnf deployment strategy for optimized self-healing in the o-cloud," in *2021 IEEE Global Communications Conference (GLOBECOM)*, 2021.

[15] 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.

[16] M. S. Hossain and G. Muhammad, "A deep-tree-model-based radio resource distribution for 5g networks," *IEEE Wireless Comm*, 2020.

[17] Y. Zhou, Z. M. Fadlullah, B. Mao, and N. Kato, "A deep-learning-based radio resource assignment technique for 5g ultra dense networks," *IEEE Network*, 2018.

[18] C. Wang, Z. Zhao, Q. Sun, and H. Zhang, "Deep learning-based intelligent dual connectivity for mobility management in dense network," in *2018 IEEE 88th Vehicular Technology Conference (VTC-Fall)*, 2018.

[19] M. Sharara, S. Hoteit, and V. Vèque, "A recurrent neural network based approach for coordinating radio and computing resources allocation in cloud-ran," in *2021 IEEE 22nd International Conference on High Performance Switching and Routing (HPSR)*, 2021.

[20] 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, pp. 1–10.

[21] 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.

[22] 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.

[23] 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

[24] B. Ojaghi, F. Adelantado, and C. Verikoukis, "So-ran: Dynamic ran slicing via joint functional splitting and mec placement," *IEEE Transactions on Vehicular Technology*, vol. 72, no. 2, pp. 1925–1939, 2023.

[25] Cplex, I. I., *V12.1: User's Manual for CPLEX*, International Business Machines Corporation, 2009.

[26] 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

[27] 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.