On Flexible Placement of O-CU and O-DU Functionalities in Open-RAN Architecture

Hiba Hojeij, Mahdi Sharara, Sahar Hoteit, Véronique Vèque

On Flexible Placement of O-CU and O-DU Functionalities in Open-RAN Architecture

Hiba Hojeij, *Student Member, IEEE*, Mahdi Sharara, *Student Member, IEEE*, Sahar Hoteit, *Member, IEEE*, Véronique Vèque, *Member, IEEE*

***Abstract*—Open Radio Access Network (O-RAN) has recently emerged as a new trend for mobile network architecture. It is based on four founding principles: disaggregation, intelligence, virtualization, and open interfaces. In particular, RAN disaggregation involves dividing base station virtualized networking functions (VNFs) into three distinct components - the Open-Central Unit (O-CU), the Open-Distributed Unit (O-DU), and the Open-Radio Unit (O-RU) - enabling each component to be implemented independently. Such disaggregation improves system performance and allows rapid and open innovation in many components while ensuring multi-vendor operability. As the disaggregation of network architecture becomes a key enabler of O-RAN, the deployment scenarios of VNFs on O-RAN clouds become critical. In this context, we propose an optimal and dynamic placement scheme of the O-CU and O-DU functionalities on the edge or in regional O-clouds. The objective is to maximize users' admittance ratio by considering mid-haul delay and server capacity requirements. We develop an Integer Linear Programming (ILP) model for O-CU and O-DU placement in O-RAN architecture. Additionally, we introduce a Recurrent Neural Network (RNN) heuristic model that can effectively emulate the behavior of the ILP model. The results are promising in terms of improving users' admittance ratio by up to 10% when compared to baselines from state-of-the-art. Moreover, our proposed model minimizes the deployment costs and increases the overall throughput. Furthermore, we assess the optimal model's performance across diverse network conditions, including variable functional split options, link capacity bottlenecks, and channel bandwidth limitations. Our analysis delves into placement decisions, evaluating admittance ratio, radio and link resource utilization, and quantifying the impact on different service types.**

***Index Terms*—Open RAN, Resource Allocation, Operations Research, Simulation, Deep Learning, RNN**

I. Introduction

THE entire Telecoms industry is going through a profound transformation driving the move towards open architectures and software-based networks. This trend is moving ahead faster and gaining momentum thanks to open-source software and standards for communication infrastructure components. On the one hand, an open architecture approach can help operators to emancipate themselves from vendors' lock-in and the derived high operational and capital expenditures. On the other hand, vendors can then bypass complex and high-barrier hardware design and production lines, focusing instead on advanced functionalities, interfaces, and software life-cycle maintenance and licensing models. As an initiative to drive openness and intelligence for the next-generation wireless networks, Open-Radio Access Network (O-RAN) has recently emerged to break the last barrier in the development of fully softwarized radio access networks [2] [3]. The O-RAN Alliance is an industry consortium focused on reshaping the radio access network (RAN) ecosystem towards more open, intelligent, virtualized, and fully interoperable mobile networks. It was founded in 2018 by leading telecom operators and vendors, including China Mobile, Deutsche Telekom, and Orange [2]. The industry has several interests in O-RAN techniques because open, virtualized, and standardized interfaces will result in cost reductions, increased flexibility, better interoperability between vendors, and fast deployment of new technologies and services. Furthermore, embedding AI and machine learning may increase network performance.

A fundamental principle of O-RAN is the disaggregation of traditionally integrated RAN components. This disaggregation involves splitting RAN functionalities into three components: the Open-Central Unit (O-CU), the Open-Distributed Unit (O-DU), and the Open-Radio Unit (O-RU), each handling separate virtual network functions (VNFs).

Unlike commonly used static deployment strategies, our research delves into the potential benefits of dynamic deployment for O-CU and O-DU components in either edge or regional clouds. The primary goal is to satisfy users' quality of service (QoS) requirements while simultaneously enhancing the overall efficiency and performance of the network. This approach introduces a new level of flexibility and adaptability to the network's architecture. To achieve this, we formulate an Integer Linear Programming (ILP) model designed to optimally and dynamically place the O-CU and O-DU within the O-RAN architecture. The exploration spans various constraints, including the capacity of cloud servers, link capacity, and delay budget. Furthermore, we consider diverse service requirements such as enhanced Mobile Broadband (eMBB), Ultra-Reliable and Low Latency Communications (URLLC), and massive Machine-Type Communications (mMTC), aiming to optimize user satisfaction while meeting the unique needs of each service type. Our approach adopts a specific deployment scenario where O-RU is consistently located at the cell site. Meanwhile, the O-DU is deployed on the Edge cloud.

Manuscript sent 15 February 2024. This work was funded by the ANR HEIDIS (https://heidis.roc.cnam.fr/; ANR-21-CE25-0019) project. This work was partially presented at IEEE International Conference on Sensing, Communication, and Networking (SECON) [1].

Hiba Hojeij, Sahar Hoteit, and Véronique Vèque are with the Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire des signaux et systèmes, 91190 Gif-sur-Yvette, France (e-mail: hiba.hojeij@ centralesupelec.fr; sahar.hoteit@centralesupelec.fr; veronique.veque@centralesupelec.fr). Sahar Hoteit is also a junior member of Institut Universitaire de France (IUF), France. Mahdi Sharara is with Orange Labs, 92320 Châtillon, France (e-mail: mahdi.sharara@orange.com).

Finally, the O-CU has the flexibility to choose between edge or regional clouds to handle its functionalities. Our approach reflects a dynamic and flexible placement scenario between scenarios B and C, as illustrated in Figure 1. Our results showcase the feasibility of establishing multiple connections between an O-RU and several O-DUs in an O-RAN network, emphasizing the concept of *Shared O-RU*, defined in the *Shared-O-RU-Multi-O-DU* feature [4]. This feature is particularly beneficial for user dispatching, offering a flexible and efficient resource allocation within the network. Furthermore, our proposed solution yields significant benefits in terms of user admittance ratio and cost reduction when compared to three baseline solutions; these include a random placement of O-CU and O-DU functionalities among regional or edge clouds, and two static placements: one with both O-CU and O-DU on edge clouds, and the other with O-CU and O-DU on regional and edge clouds, respectively.

In Addition, we introduce a heuristic to efficiently solve the optimization problem. Leveraging a recurrent neural network (RNN) based model [5], [6], we tap into the potential of deep learning to provide less-complex alternatives to highly-complex optimal algorithms in terms of execution time such as the branch and bound algorithms used to find optimal solutions to ILP problems [7]. We propose an RNN-based model that uses a bidirectional LSTM architecture trained with the ILP model's output to effectively mimic the optimal placement of O-CU and O-DU, achieving the desired benefits. Our approach is novel because we integrate the LSTM model that considers the sequential nature of RAN disaggregation and long-term dependencies among users. This enables efficient VNF Placement across available computing resources.

Finally, we rigorously evaluate the optimal model's performance under diverse network conditions. This comprehensive assessment considers potential computational, radio, and link capacity bottlenecks and constraints imposed by channel bandwidth, as well as variations in functional split options. We analyze performance regarding admittance ratio, multi-resource utilization, and the impact on different service types. This exploration allows us to assess the model's adaptability to varying requirements of the O-RAN architecture.

The main contributions of this paper are summarized here:

1) We propose a flexible placement scenario of O-CU and O-DU functionalities between edge and regional clouds in O-RAN architecture.
2) We formulate an ILP model for the optimal placement of O-CU and O-DU, considering various constraints and service requirements, with the objective of maximizing users' admittance.
3) We introduce a heuristic solution using RNN-based models that achieves performance closely comparable to the ILP-based optimal algorithm but significantly reduces execution time.
4) We evaluate the model's performance under diverse network conditions, including computing resources, link capacities, and radio resources.

The rest of this paper is organized as follows. Section I overviews the O-RAN disaggregation concept and our work motivation. Section III provides an overview of the related work. Our proposed ILP-based model and deep learning-based heuristic are described in Section IV and Section V, respectively. The simulation framework is detailed in Section VI. Section VII quantifies the behavior of the proposed algorithms, and finally, Section VIII concludes the paper.

II. O-RAN Overview and motivation

One of the fundamental principles underlying O-RAN is the disaggregation process. Traditionally, RAN components were tightly integrated, with hardware and software provided by a single vendor. Using disaggregation, hardware and software are decoupled. Virtualization allows the RAN functions to run on general-purpose hardware rather than specialized, proprietary equipment. The disaggregation of RAN into three main distinct components (O-CU, O-DU, and O-RU) enables each to be implemented independently, thus promoting greater flexibility and interoperability within the network [8]. The disaggregation concept has altered the definition of the RAN and redirected the attention of resource allocation solutions towards the O-CU and O-DU, especially in terms of VNF deployment options. These options comprise several configurations for placing O-CU and O-DU at regional and edge cloud nodes. By deploying these VNFs across nodes at varying distances, network operators gain the ability to tailor their deployments to specific network requirements and meet minimum Quality of Service (QoS) standards, such as high throughput and low latency. For example, placing VNFs closer to end-users, such as at cell sites, can reduce end-to-end latency for applications with strict delay constraints. Conversely, deploying these components in more distant cloud hosts can leverage greater computational resources to manage higher packet processing rates. Therefore, while the disaggregation and distribution of functional units throughout the O-Cloud network offer significant flexibility, their deployment must be carefully planned to meet stringent constraints. Practical examples include scenarios where real-time applications like augmented or virtual reality demand low latency, necessitating edge deployments. Meanwhile, data-intensive applications such as video streaming or cloud gaming can benefit from higher processing power in regional clouds.

To enhance the network efficiency, the disaggregation concept yields a wide range of functional split options, distributing baseband processing functions among the O-RU, O-DU, and O-CU. The O-RAN Alliance selects the functional split 7.2x that balances the simplicity of the radio unit and the data rates and latency required on the interface between the radio and the distributed units [8]. Specifically, in the 7.2x split, the O-RU handles Fast Fourier Transform (FFT) and cyclic prefix addition/removal operations, making the RU cost-effective and easy to deploy. The O-DU then manages the remaining physical layer functionalities and the Medium Access Control (MAC) and Radio Link Control (RLC) layers. Finally, the O-CU implements the higher layers of the 3GPP stack, including the Radio Resource Control (RRC) layer, the Service Data Adaptation Protocol (SDAP) layer, and the Packet Data Convergence Protocol (PDCP) layer. The Near-RT RIC is responsible for intelligent edge control of

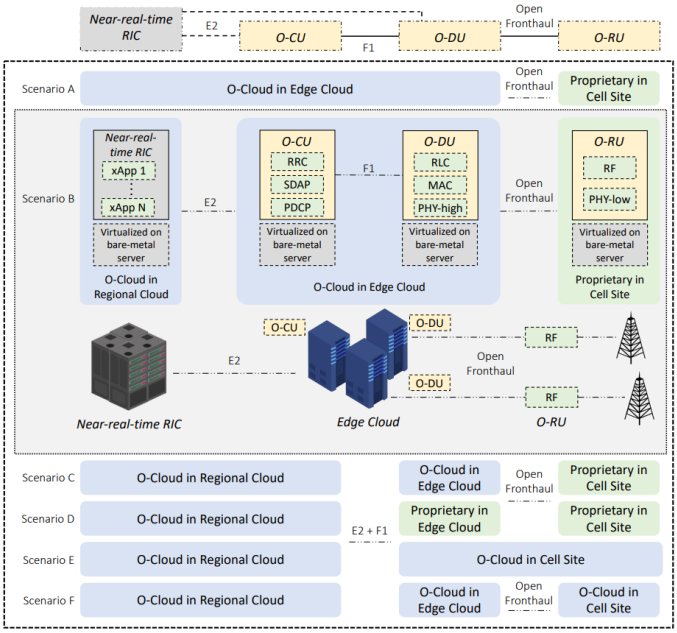


Figure 1: O-RAN Cloud deployment scenarios [10]

RAN nodes and resources. It controls RAN elements and their resources through microservices, called xApps, which typically run control loop tasks with durations ranging from 10 milliseconds to one second.

In this context, the O-RAN alliance envisions different strategies for deploying its functional splits on regional or edge cloud locations or at proprietary cell sites [9]. Fig. 1 depicts the different O-RAN Cloud deployment scenarios [10]. For instance, Scenario A refers to the case where all the network components except the O-RUs are deployed at the edge cloud of the network. Scenario B presents the case where the O-DU and O-CU functionalities are located in the edge cloud while the Near-RT RIC is in the regional cloud. Our proposal adopts a flexible deployment scenario between scenarios A and B, leveraging the computing resources utilization and enhancing the network performance. The interfaces among the O-CU, O-DU, and O-RU, shown in Fig. 1, are crucial for communication in disaggregated networks. The fronthaul connects the O-RU to the O-DU for user data transmission, the F1 interface links the O-CU and O-DU for control and user data exchange, and the E2 interface connects the O-CU to the Near-RT RIC for real-time RAN resource optimization through xApps.

III. Related Work

To address the challenges introduced by disaggregating the RAN, several works from the literature tackle the VNF placement problem in O-RAN to optimize resource allocation mechanisms but with different objectives. Moreover, the intelligence supported by the O-RAN boosts studies on the application of deep learning-based techniques in RAN.

A. VNF placement in O-RAN

Optimization objectives of the VNF placement problem are the most relevant characteristic when comparing the related work since they yield distinct problem formulations and provide diverse insights. In [11], the authors propose a deep reinforcement learning method that explores the best O-Cloud locations for O-DU and O-CU Virtualized/Cloud-native network functions (VNFs/CNFs), along with the optimal user equipment to O-RU associations. Their objective is to minimize the delay while reducing the deployment cost. According to their findings, the proposed algorithm outperforms the static allocation of the O-CUs and O-DUs, although they do not consider the diverse service requirements of different users. Authors in [12] tackle the flexible placement of the three-layer RAN slices (O-RU, O-DU, O-CU) over a multi-tier aggregation sites network topology while adopting flexible functional split options. Our consideration of edge-regional server infrastructure is similar to their multi-tiered network infrastructure. However, they seek to maximize the profit of the infrastructure provider. Moreover, in [13], the authors propose an optimization model that deploys O-RAN components within regional and edge clouds while minimizing the network outage. Scenario B is adopted in their work, where O-CU and O-DU are always placed at an edge server. The work of [14] suggests a framework that optimizes the number of instantiated RUs in a given area based on its long-term network statistics. Then, it associates these RUs with open-access edge servers to host the corresponding DUs and CUs. The main objective of their work is to minimize the overall deployment cost by installing the minimum number of RUs and open-access edge servers. In [15], authors present a dynamic DU placement strategy, which enhances flexibility in positioning DUs across the network to minimize O-RAN costs. However, these studies retain fixed CU locations, potentially leading to sub-optimal results. The authors of [16] explore the placement of VNFs on various nodes (access and aggregation nodes) while incorporating flexible functional split options. They account for each split option's latency, bandwidth, and computing resource requirements. Given these resources' availability and the split options' priority, their model places VNFs to achieve minimal computing resource utilization with maximum aggregation efficiency.

In conclusion, various studies in the literature have tackled the placement problem of O-RAN components from different perspectives. Some have focused on minimizing delay, reducing deployment cost, maximizing profit, minimizing network outage, and reducing overall deployment cost. Our work in this paper addresses the dynamic placement problem of the O-RAN components to maximize users' admittance ratio while satisfying the diverse QoS requirements of uRLLC, eMBB, and mMTC slices. Our proposed solution enables the optimal and dynamic allocation of O-CU and O-DU functions on either edge or regional clouds.

B. Deep Learning-based solutions in RAN

This section reviews existing Deep Learning-based works, primarily focusing on RNN applications in the 4G/5G RAN. Works in [17] and [18] mainly address radio resource scheduling while considering dynamic changes in radio access and services. In [17], the authors present a Deep Learning-based framework for intelligent radio resource assignment in 5G networks. This framework aims to predict traffic congestion

and the occupancy state of base stations, allowing for an adaptive uplink and downlink ratio to prevent congestion. The proposed framework utilizes a deep tree model and a long short-term memory (LSTM) network to forecast future traffic based on current and past data. Similarly, the authors in [18] address the traffic congestion issue using a deep LSTM learning algorithm to predict traffic load at the base station. The proposed algorithm executes appropriate action policies based on these predictions to avoid or reduce congestion intelligently. These works primarily employ LSTM networks due to the temporal dependency nature of traffic data.

Moreover, several works focus on managing user handovers and base station energy based on user mobility using RNN models. In [19], the authors first presented an analytical model of handover cost in 5G, considering factors such as signaling overhead, latency, call dropping, and radio resource wastage. Then, they propose a prediction scheme based on the RNN with the LSTM algorithm to minimize handover costs. Their study demonstrated that accurate handover predictions could significantly reduce user dissatisfaction, handover latency, resource wastage, and overhead. Similarly, the LSTM algorithm is utilized in [20] to learn each UE's mobility pattern from its historical trajectories and predict its future movements. Based on these mobility predictions, the corresponding base station determines whether a handover is necessary for the UE.

The work in [21] proposes a BiLSTM RNN-based approach to sub-optimally depict the performance of an ILP-based algorithm for optimal allocation of computing resources in a given centralized RAN architecture trained to select Modulation and Coding Scheme (MCS) index. Compared to this work, our RNN solution predicts the optimal placement of VNFs and depicts the performance of an ILP model that constitutes more constraints, making it challenging to be trained to mimic the optimal model.

Inspired by these RNN learning scheme capabilities used to deal with emerging challenges at the RAN level, we consider using the BiLSTM RNN model to solve our problem of optimal placement of O-CU and O-DU functionalities while maximizing the user's admittance. The system's set of users is considered a sequence of inputs that the RNN will process to detect dependencies among them. The RNN's capability to capture these dependencies makes it well-suited for our resource allocation problem, as it helps the model learn the optimal placement of O-CU and O-DU functionalities.

Our approach is novel because we integrate the LSTM model that considers the sequential nature of RAN disaggregation and long-term dependencies among users. This enables efficient placement of VNFs across available computing resources. Overall, our paper contributes to the research on dynamic VNF placement in O-RAN environments and highlights the importance of considering the interdependency among users of a given traffic condition on user admission.

This study extends our prior research [1], which focused solely on evaluating the performance of the proposed placement model under constraints related to radio and computing resources. The previous work lacked a comprehensive evaluation of the model's placement decisions across different network conditions. In our current study, we expand the assessment by considering additional constraints. We analyze the model's performance under limited channel bandwidth as well as varying link capacity. Understanding how the model behaves under these constraints and impacts different service types is crucial for real-world deployment scenarios.

TABLE I: Network Parameters and Notations

Parameters	***Definition***
$\mathcal{S}$	Set of all servers
$\mathcal{S}_{reg}$	Set of regional servers
$\mathcal{S}_{edge}$	Set of edge servers
$\mathcal{I}$	Set of all users
θ_{is}^{FU}	Binary variable indicating if server s hosts UE i's FU
$C_{ss'}$	Link available capacity between server s and s' (Gbps)
B_i^{mid}	Link capacity required by user i on mid-haul (Mbps)
R_s	Available computational capacity on server s (GOPS)
R_i^{FU}	Required server capacity for user i's FU (GOPS)
α_{CU}	Computational complexity of O-CU
α_{DU}	Computational complexity of O-DU
$\delta_{ss'}$	Latency between server s and s' (ms)
δ_i^{mid}	Maximum mid-haul latency for user i (ms)
W_i	Maximum achievable throughput by user i (Mbps)
C_F_{is}	Centralization factor of user i over server s
ϵ_i	Priority value for user i

IV. Proposed ILP-based Optimal Model

To solve the placement problem of the O-CU and O-DU in O-RAN architecture, our main intention is to optimize the usage of cloud resources, particularly computational resources. We develop an ILP-based model that maximizes users' admittance ratio while moving toward the regional cloud, considering the computational capacity at the O-cloud servers, the delay budget, and the available link capacity. It is worth mentioning that the processing costs at the edge O-Cloud nodes are higher than those on regional O-Cloud nodes [11]. Thus, we propose an optimal and dynamic allocation of the resources between edge and regional clouds, which encourages, for instance, the O-CU functionalities to be at the regional clouds if users' service requirements permit. We consider a set of $\mathcal{S}$ servers randomly distributed over the edge and regional clouds, where $\mathcal{S}_{edge}$ and $\mathcal{S}_{reg}$ define the sets of edge and regional servers, respectively. We define R_s as the available computational capacity on server $s \in \mathcal{S}$ in terms of Giga Operations Per Second (GOPS). Furthermore, we define the link latency between two servers s and s' by $\delta_{ss'}$. The network includes a set of $\mathcal{I}$ users, each belonging to one of the three service types (eMBB, uRLLC, or mMTC), with different service requirements. We denote the maximum allowed latency on the mid-haul link (i.e., the link between the O-CU and the O-DU) by each user i, by δ_i^{mid}. We recall that the O-DU is set in our scenario to be at the edge cloud, while O-CU can choose between the edge and regional clouds. Table I summarizes the notations used throughout the paper.

The link and computational capacity requirements as well as the delay budget are modeled using equations (1), (2), and

(3), respectively. The maximum achievable throughput by an admitted user is formulated in equation (4).

- The mid-haul link (i.e., the link between the O-DU and O-CU when adopting option-2 split) capacity B_i^{mid} needed for each user $i \in \mathcal{I}$ is modeled as referred to [22] and [23] by:

$$B_i^{mid}[Mbps] = \frac{TBS \cdot N_{TBS}(IP_{pkt} + H_{PDCP})}{(IP_{pkt} + H_{PDCP} + H_{RLC} + H_{MAC}) \cdot 1000} \tag{1}$$

where TBS represents the transport block (TB) size, N_{TBS} is the number of TBs per TTI, IP_{pkt} is the IP packet size, and lastly, H_{PDCP}, H_{RLC} and H_{MAC} the header size of PDCP, RLC, and MAC layers, respectively. These parameters are defined as in the standard specification in [24].

- The computational server capacity required by each user $i \in \mathcal{I}$ is modeled based on an estimation of the complexity in terms of Giga Operations Per Second (GOPS). To quantitatively determine the computational complexity R_i^{FU} of a functional unit FU for user i (FU refers to either O-CU or O-DU functional units), we use the computational model from [12]:

$$R_i^{FU}[GOPS] = \frac{\alpha_{FU}(3A + A^2 + M \cdot C \cdot L/3)RB_i}{10} \tag{2}$$

where α_{FU} is a scaling factor that represents the computational requirement of a specific functional unit FU with respect to the overall computational requirement. The total computational capacity is distributed among O-RU, O-DU, and O-CU based on the 'PHY split' and 'RLC-PDCP split'. With the considered split-7.2x (between O-RU and O-DU) and split-2 (between O-DU and O-CU), 40% of the processing is done by O-RU, 50% by O-DU, and 10% by O-CU as mentioned in [14]. Hence, α_{DU} and α_{CU} are respectively equal to 0.5 and 0.1. 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_i, the number of resource blocks assigned to user i.

- The link latency $\delta_{ss'}$ between servers $s, s' \in S$ is determined by the propagation delay in the fiber links, which is the ratio of the distance between servers, $dist(s, s')$ multiplied by the refractive index of the fiber optic cable (n = 1.5) over the speed of light in the fiber c.

$$\delta_{ss'} = \frac{dist(s, s') \cdot n}{c} \tag{3}$$

- The maximum achievable throughput of a given user $i \in \mathcal{I}$, denoted as W_i, is determined in equation (4) as specified in [25].

$$W_i[Mbps] = \frac{N_{sym} \cdot N_{SC} \cdot M \cdot C \cdot L(1 - 0.14)RB_i}{1000} \tag{4}$$

where N_{sym} is the number of symbols per sub-frame and N_{SC} is the number of subcarriers per RB.

We formulate the placement of O-DU and O-CU optimization problem as follows in Problem 1. The objective function in (5) aims at maximizing the number of admitted users. θ_{is}^{CU} and θ_{is}^{DU} are the binary decision variables indicating whether user $i \in I$ chooses server $s \in S$ for its O-CU and O-DU functionalities, respectively or not. Our objective function includes C_F_{is}, a distance-dependent centralization factor. It is determined as follows: C_F_{is} is set to be inversely proportional to the distance between the edge server s and the O-RU, to which user i is associated, if $s \in S_{edge}$, and set to be one if $s \in S_{reg}$. This setup encourages each user's O-DU functionality to select the nearest available edge server to its associated O-RU. As for the O-CU functionality, which can be hosted either on edge or regionally, it will prefer to choose the regional option, having the higher weight of C_F_{is}, if the latency requirements allow. However, it will choose the server nearest to its corresponding O-RU if this is not feasible. Moreover, we add a priority parameter ϵ_i to the objective function as a function of the user's service type, allowing us to prioritize eMBB and uRLLC UEs over mMTC ones.

Problem 1:

$$\text{maximize} \quad \sum_i \sum_s C\_F_{is} \cdot (\theta_{is}^{CU} \cdot \epsilon_i + \theta_{is}^{DU} \cdot \epsilon_i) \tag{5}$$

$$\text{subject to} \quad \theta_{is}^{CU}, \theta_{is}^{DU} \in \{0, 1\}, i \in \mathcal{I}, s \in \mathcal{S} \tag{6}$$

$$z_{iss'} \in \{0, 1\}, i \in \mathcal{I}, s, s' \in \mathcal{S} \tag{7}$$

$$\sum_{s \in \mathcal{S}} \theta_{is}^{CU} \leq 1, i \in \mathcal{I} \tag{8}$$

$$\sum_{s \in \mathcal{S}} \theta_{is}^{DU} \leq 1, i \in \mathcal{I} \tag{9}$$

$$z_{iss'} \leq (\theta_{is}^{DU} + \theta_{is'}^{CU})/2, s, s' \in \mathcal{S}, i \in \mathcal{I} \tag{10}$$

$$z_{iss'} \geq \theta_{is}^{DU} + \theta_{is'}^{CU} - 1, s, s' \in \mathcal{S}, i \in \mathcal{I} \tag{11}$$

$$\sum_{s \in \mathcal{S}} \theta_{is}^{DU} = \sum_{s \in \mathcal{S}} \theta_{is}^{CU}, i \in \mathcal{I} \tag{12}$$

$$\sum_{s \in \mathcal{S}_{regional}} \theta_{is}^{DU} = 0, i \in \mathcal{I} \tag{13}$$

$$\sum_{i \in \mathcal{I}} B_i^{mid}(z_{iss'} + z_{is's}) \leq C_{ss'}, s, s' \in \mathcal{S}, s \neq s' \tag{14}$$

$$\sum_{i \in \mathcal{I}} R_i^{CU} \theta_{is}^{CU} + R_i^{DU} \theta_{is}^{DU} \leq R_s, s \in \mathcal{S} \tag{15}$$

$$\delta_{ss'} \cdot z_{iss'} \leq \delta_i^{mid}, i \in \mathcal{I}, s, s' \in \mathcal{S} \tag{16}$$

Our ILP problem 1 has the following constraints: Constraint (6) defines θ_{is}^{CU} and θ_{is}^{DU} as binary integer variables. These variables are set to 1 if and only if the O-CU and O-DU functionalities of user i are admitted on the server s. Constraint (7) defines $z_{iss'}$, a binary decision variable that is set to 1 when O-DU and O-CU functionalities of a user i are allocated at servers s and s', respectively, i.e., $z_{iss'}$ represents the product of the two decision variables θ_{is}^{CU} and θ_{is}^{DU} of the model. Constraints (8) and (9) ensure that the user's functionalities O-CU and O-DU are not allocated more than once. Constraints (10) and (11) ensure that $z_{iss'}$ is set to one only if O-DU and O-CU of the user i are allocated at servers s and s', respectively. Constraint (12) guarantees that either both functionalities of the user are admitted or not, i.e.,

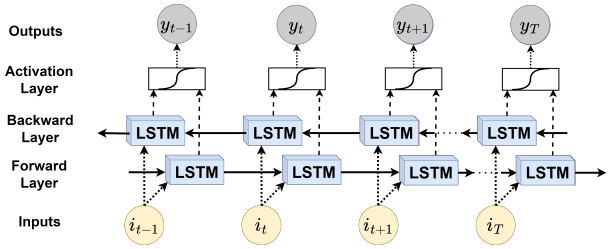


Figure 2: The Bi-LSTM layer

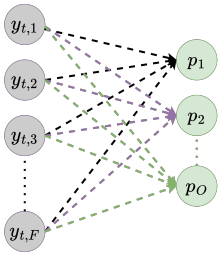


Figure 3: The fully connected layer

if one of the O-DU or O-CU functionalities is not allocated, the whole user will be discarded. Constraint (13) enforces that the O-DU functionality is never allocated on a regional server. Constraint (14) ensures that the link capacity required for user i between the servers s and s' chosen for O-DU and O-CU does not exceed the available capacity between these two servers $C_{ss'}$. The symmetry of link capacity between the servers is taken into account in this latter constraint. Server computational capacity is respected by constraint (15). Finally, the maximum link latency is guaranteed by constraint (16).

V. Deep Learning-based Solution

Due to the high complexity of solving the NP-Hard ILP problem [26] defined in the previous section, finding a solution to our ILP problem would take an impractical amount of time. Thus, we need to consider alternatives with lower computational complexity. Deep Learning has demonstrated its potential to tackle complex tasks by learning a function that maps the input to the desired output. A Recurrent Neural Network (RNN) is a branch of deep learning that can handle sequences of interdependent elements such as weather prediction and language translation. In our task that involves multiple users sharing common resources, RNN would be beneficial (i.e., the placement decision of the user's O-CU and O-DU functionalities made at a specific time step will affect the availability and suitability of O-Cloud resources for other users in the network in the subsequent time steps, and this dependency needs to be taken into account in order to ensure optimal allocation). Long-Short-Term-Memory (LSTM) [27] is a well-known architecture of RNN that can deal with long-term dependencies in sequential data. The LSTM architecture includes memory cells and gates, such as input, output, and forget gates, that control the flow of information. At each time step, the LSTM receives input and hidden state vectors to update the memory cell and generate an output vector. A traditional LSTM RNN architecture variant is the bidirectional Long Short-Term Memory (BiLSTM) RNN model [27]. The BiLSTM RNN model extends the traditional LSTM architecture by simultaneously processing input sequences in both forward and backward directions. By incorporating information from both directions, the BiLSTM model can capture more complex dependencies between input elements and thus achieve higher accuracy.

In our study, we propose a heuristic approach, which involves utilizing an RNN model to learn and predict the optimal placement of O-CU and O-DU among available servers. The model uses a sequence-to-sequence classification, where each element in the sequence corresponds to a user and produces an output that represents a decision on the placement of O-CU and O-DU functionalities for that user. The model is composed of a BiLSTM RNN layer and a fully connected layer, as illustrated in Figures 2 and 3, respectively. The BiLSTM layer receives a sequence of users as input of size T, where each user is represented by a feature vector that includes several parameters, such as its relative position with respect to the O-RU, number of RBs, MCS index, associated O-RU, user requirements (i.e., maximum latency, GOPS required), slice type, priority, etc. For an input i_t, the BiLSTM produces an output y_t, which is a vector of size F containing the elements $[y_{t,1}, y_{t,2}, ..., y_{t,F}]$, with $F = 2H$, where H is a hyperparameter representing the number of hidden layers in an LSTM. Each element in the output vector has a value between -1 and 1. This output vector is then fed into the fully connected layer that uses the softmax activation function for multi-class classification. The classification layer includes O neurons, where O represents the number of possible decisions or labels. The labels combine O-CU and O-DU locations among the available servers, plus an additional label to indicate that a user has been dismissed. The neuron with the highest activation value corresponds to the decision. We recall that our optimal problem deals with the placement of O-CU and O-DU functionalities on a per-user basis.

To generate the training dataset, we performed 25,000 simulations, each representing a distinct network configuration instance with varying parameters, including the number of users, user locations, service requirements, and resource availability. The data gathered from each simulation serve as inputs to the BiLSTM model, with a sequence of users and their associated features. We derive the optimal placement decisions for each simulation scenario by solving the joint ILP model. These optimal solutions are then used as labels in our training dataset.

VI. Simulation Framework

We consider a network topology composed of 4 O-RUs distributed over an area of 1 km^2. This assumption is based on a real traffic profile for hourly UEs density variation in a 1x1 km industrial area, as described in [14], in which the optimal number of O-RUs to be instantiated was determined to be 4. UEs are randomly distributed in the considered area; an example of the network topology with 20 UEs is depicted in Fig.4. The system uses a 20 MHz bandwidth so that each O-RU has 100 RBs available per transmission time interval (TTI). UEs are associated with the nearest O-RU, and

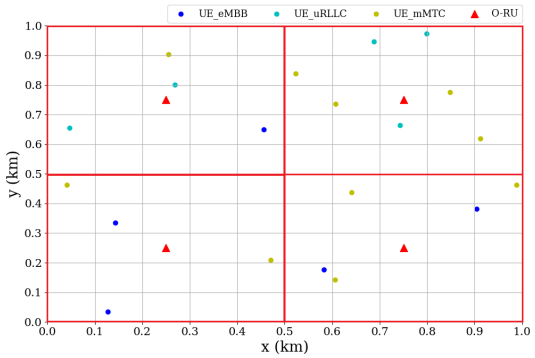


Figure 4: Example of the network topology with 20 UEs

we consider a number of 10 to 140 UEs spread following the distribution presented in [14] for an industrial area that has 25% eMBB users, 25% uRLLC users, and 50% mMTC users. We assume the existence of three edge servers located approximately 10 km from the O-RUs and one regional server located between 40 to 80 km from the O-RUs [14]. Each O-RU at the cells site is fully connected to the three edge-cloud servers. Moreover, we assume a mesh connectivity of fiber links among all servers in the system. This topology particularly considers that the 10 km and 80 km distance limit at the fronthaul and midhaul, respectively, is not exceeded. It is worth noting that testing our solution on one km² area does not impact its scalability, as we evaluate the model with an increasing number of users. Additionally, for larger areas, considering that the ILP model becomes significantly complex for larger systems—rendering it impractical for any solver to reach the optimal solution—we suggest deploying our solution in clusters of smaller areas rather than a single large area.

The computational capacity R_s of edge servers follows a uniform random distribution ranging from 100 to 200 GOPS, while the regional server's capacity ranges from 1000 to 2000 GOPS, as stated in [12]. The mid-haul latency bounds δ_i^{mid} are considered as in [14] a random value in the range of 100 to 300 µsec for uRLLC users, 500 µsec for eMBB, and 1000 µsec for mMTC. The radio resource allocation follows an approach inspired from [14] that consists of allocating 50% of the total available resource blocks (RBs) to eMBB users and 25% to each of the uRLLC and mMTC users with no resource waste. In addition, eMBB users are assigned a random number of RBs between 10 to 20, while uRLLC and mMTC users are assigned between 1 to 5 RBs, as in [28] [29]. The MCS index for each user is set as a random number between 17 to 28, with all users assumed to have a 64-QAM modulation scheme, as in [12]. We note that the MCS index impacts the code rate and spectral efficiency, as referred to in 3GPP specification [24]. The available bandwidth of the mid-haul link between edge-edge servers is a random value ranging from 1 to 10 Gbps, while the bandwidth between edge-regional servers is randomly chosen between 10 and 20 Gbps. Accordingly, these values are selected so that the mid-haul link can support the throughput demand of all admitted users as in [14]. Additional radio parameters used in the experiments are outlined in Table II. We note that our ILP-based problem is solved

TABLE II: Summary of radio parameters

Parameter	***Value***
A	4 Antennas
N_{Sym}	14 symbols per sub-frame
N_{SC}	12 subcarriers
L	2 MIMO layers
M	$log_2(64)$

TABLE III: Summary of RNN Parameters

Parameter	***Value***
Optimizer	Adam
Learning Rate	0.005
Hidden Size (LSTM)	128
Number of LSTM Layers	1
Batch Size	60
Input Size	92
Sequence Length	100

using IBM CPLEX software [7], a mathematical optimization solver, on a computer with 11th generation Intel® Core™ i9-11950H Processor and 16 GB RAM. Finally, for the RNN model implementation, we utilized Python with the PyTorch library. The input size of our dataset is 92, corresponding to the number of features per user. The Adam optimizer was employed to optimize the training parameters, with a learning rate of 0.005. We configured the model with 1 LSTM layer and a hidden size H of 128, which is determined by trial and error. The batch size was also chosen by trial to be 60. For detailed simulation parameter settings of the RNN, please refer to Table III. Finally, for the RNN model implementation, we utilized Python with the PyTorch library. Each entry in our dataset has an input size of 92, corresponding to the number of features per user. The sequence length of the trained model is set to 100 (padding is used for scenarios with fewer users in the system). We employed the Adam optimizer to optimize the training parameters, starting with a learning rate of 0.005. The model configuration includes 1 LSTM layer with a hidden size H 128, determined by trial and error. The batch size, chosen empirically, is set to 60. For detailed simulation parameter settings of the RNN, please refer to Table III. Our code includes customization for dataset reading and generation.[1]

VII. Performance Evaluation

In this section, we compare the performance of our proposed algorithm, referred to as the *Optimal scenario*, along with the heuristic based on *RNN* model with respect to three baselines defined as follows:

- An *All_Edge scenario*; in which O-CUs and O-DUs are always on the edge servers (i.e. scenario B of Fig. 1).
- A *Static scenario*; where O-CUs are always placed on the regional servers while the O-DUs are always on the edge servers (i.e. scenario C of Fig. 1).
- A *Random* scenario; servers are placed randomly between edge and regional for both O-DUs and O-CUs.

[1]The code is publicly available at https://github.com/HibaHojeij/CU-DU-placement-in-O-RAN.git.

A. Optimal Model versus Baselines

The performance metrics used in this subsection are:

- Average admittance ratio: It reports the average number of admitted users among all users present in the network at each transmission time interval (TTI).
- Throughput: It evaluates the average throughput of all admitted users. The throughput of an admitted user $i \in \mathcal{I}$, W_i, is determined based on Equation (4).
- Deployment Cost: This metric quantifies the average cost of deploying O-CUs at the selected servers. It is computed as the cost of running the computational operations on a server (in GOPS). The regional server has more processing capacity and uses less energy than the edge server; thus, running VNFs in regional servers is less expensive than in edge servers [11]. At the edge server, according to [14] and [22], 1 GOPS costs 1.59$, while at the regional cloud, it costs 0.5$/GOPS.
- Fairness Index: For measuring how fair the users are being admitted over the three service types (eMBB, uRLLC, mMTC), Jain's fairness index is used as formulated in Equation (17) as follows:

$$\zeta = (\sum_{j=1}^{N} AAR_j)^2 / (N \cdot \sum_{j=1}^{N} AAR_j^2) \tag{17}$$

where $N = 3$ refers to the number of heterogeneous service types, AAR_j is the average admittance ratio of users of service type j.

We note that 100 simulations were performed, and confidence intervals of 95% are provided in the following results. We start our evaluation by analyzing the average admittance ratio as a function of the number of users for each considered scenario of the *Optimal* scenario in comparison to the baselines from the state-of-the-art. The results, as depicted in Fig. 5, demonstrate that the *Optimal* scenario outperforms all other scenarios in terms of the average admittance ratio. The *All_Edge* scenario follows the same trend, but with a 10% lower admittance ratio, due to the limited computational resources of edge clouds in meeting the diverse users' requirements, namely eMBB users, which are computationally more demanding. On the other hand, the *Random* and *Static* scenarios have the poorest average admittance ratio, which can be interpreted by the fact that uRLLC users have low latency requirements; hence, placing O-CUs in a regional cloud, whether randomly or statically, increases link latency, leading to a lower probability of user admission. Furthermore, we present the performance of our proposed RNN model, illustrated in purple on the same graph of Fig. 5. The results indicate that the RNN model can closely replicate the optimal model's admittance ratio, with a difference of no more than 2% compared to the optimal solution. This proves the RNN's ability to capture dependencies and learn from the large dataset of network scenarios enables it to provide near-optimal solutions, addressing the limitations of traditional baselines from state-of-the-art that fail to give a close performance to the ILP-based one as the RNN does. Additionally, the system starts experiencing a decline in the admittance ratio when the number of users exceeds 50. To better understand this behavior, Figures 6 and 7 report the GOPS and RB allocation, respectively, in the system. As shown in Fig. 6, the computational resources at the three edge servers become utilized at more than 80% when the number of users in the system exceeds 50, indicating that, at and beyond this point, the capacity of the edge servers becomes the bottleneck for more demanding users (i.e., eMBB users) as we will show later on. On the other hand, as seen from Fig. 7, all RUs become fully loaded when the number of users reaches 100, resulting in extra users not being assigned by RBs and, therefore, not being admitted. Despite the system becoming overloaded with more than 50 users, the *Optimal* model gives the best performance in terms of admittance ratio, as mentioned earlier; that is, the model strikes a balance between available resources and users' demands, taking into account their priorities. To further investigate the admittance ratio for different service types, we plot the admittance ratio for each service type for the different placement scenarios in Figure 8. Comparing the *Optimal* scenario with the *All_Edge* scenario, we notice that the former scenario admits more eMBB and mMTC users (Fig. 8a and Fig. 8c) and almost the same number of uRLLC users (Fig. 8b). This can be explained by the fact that the eMBB and uRLLC services are prioritized over mMTC services. Cloud computational resources are allocated accordingly when they are available and satisfy their latency requirements. Moreover, moving to the regional cloud provides more abundant resources, allowing for more eMBB and mMTC users to be admitted without penalizing the uRLLC user, as is the case in the *Optimal* scenario. Additionally, compared to the performance of the *Optimal* scenario, the RNN model shows that fewer uRLLC users are admitted while slightly more mMTC users are admitted. This highlights the reason for the 2% gap in the total average admittance ratio, seen in Fig. 5. The RNN model is suboptimal in predicting the placement of VNFs for uRLLC users. Finally, to draw a connection with the limited system capacity, we focus on the *Optimal* scenario of plots of Figure 8. The admittance ratio of eMBB and mMTC users reveals that they become not fully admitted when the number of users in the system exceeds 50, while uRLLC users are fully admitted at that stage. This means that the uRLLC users, having high priority and less GOPS demand, are prioritized over other users when the load on servers becomes more critical to meet the model objective of maximizing the admittance ratio. These results are consistent with the earlier analysis of the GOPS load presented in Figure 6, highlighting the limitation of the edge server capacity in our system.

In terms of throughput, Fig. 9 illustrates the overall throughput achieved by deploying different placement scenarios. It is evident that the *Optimal* scenario outperforms all other scenarios in terms of throughput. This result is consistent with the higher average admission ratio achieved by the *Optimal* scenario, as shown in Fig. 5. Moving to the computational cost evaluation, Fig. 10 presents the cost of deploying O-CUs for admitted users as a function of the total number of users for different placement scenarios. The *Optimal* placement scenario achieves up to 50% cost reduction compared to *All_Edge* scenario, as the former utilizes more regional

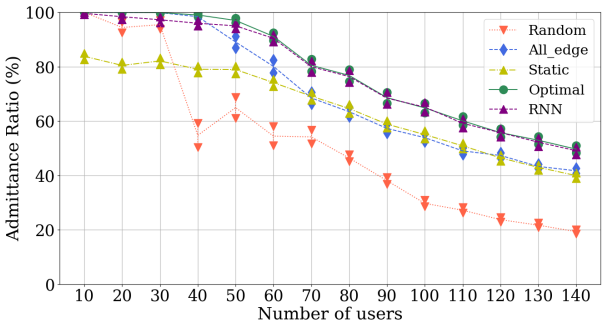


Figure 5: Average admittance ratio as a function of number of users

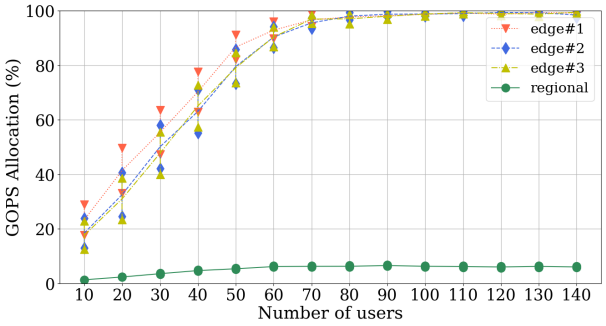


Figure 6: GOPS allocation per server in Optimal deployment scenario

servers, which are less expensive. The *Static* scenario has the lowest cost due to admitting fewer users and having the only possibility to choose regional clouds for hosting O-CUs. Moreover, an important consideration is the fairness of the admittance ratio among the three service types as the number of users in the system changes. The results are shown in Fig. 11 for the different scenarios. The *Optimal* scenario offers a better fairness index among users compared to the other scenarios. The RNN model achieves better fairness than the optimal scenario by admitting more mMTC users at the cost of admitting fewer uRLLC users, as we have seen before.

In addition to the performance improvements achieved by our proposed model, it is important to state that the RNN heuristic offers a significant advantage in terms of execution time when compared to the ILP model. As shown in Fig. 12, the RNN model achieves a remarkable 97% reduction in execution time, even when the number of users increases. The reduction in execution time becomes more significant as the number of users increases because the ILP model is

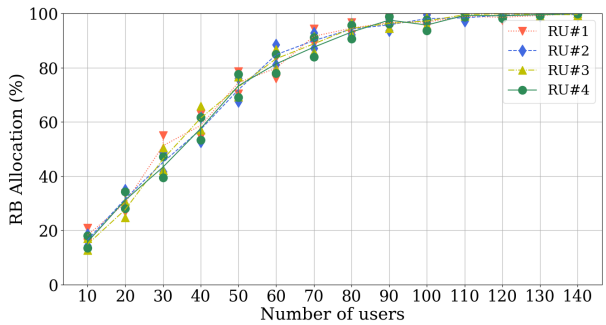
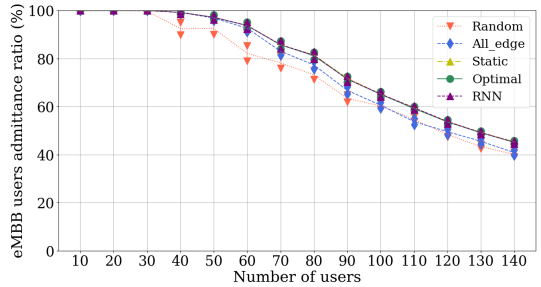
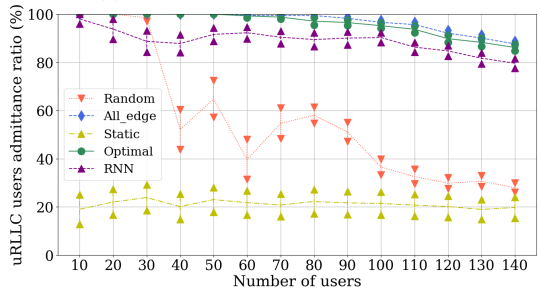


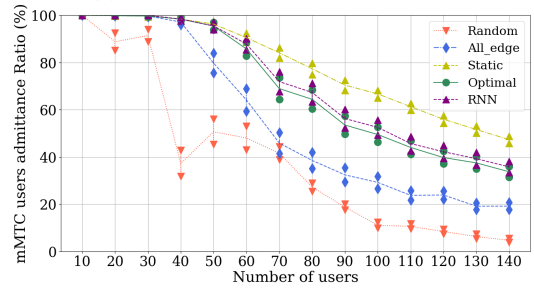
Figure 7: RB allocation as increasing the number of users



(a) Admittance ratio of eMBB users



(b) Admittance ratio of uRLLC users



(c) Admittance ratio of mMTC users

Figure 8: Admittance ratio for each service type

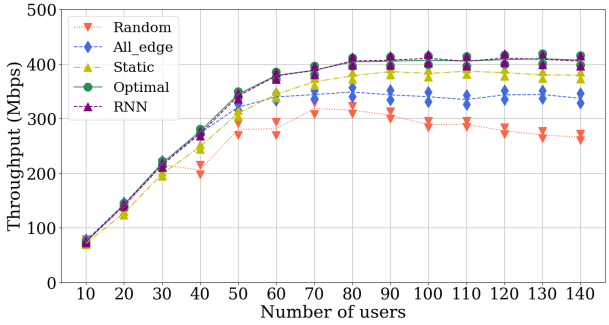


Figure 9: Total throughput as the function of number of users

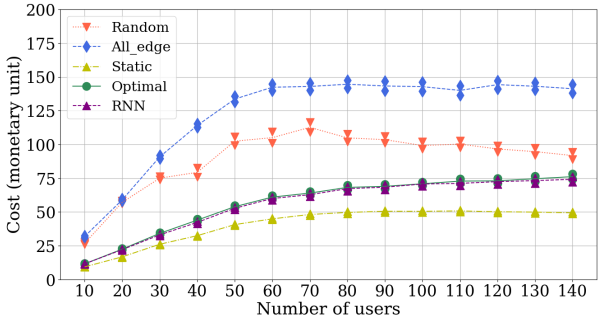


Figure 10: O-CU deployment costs for each scenario

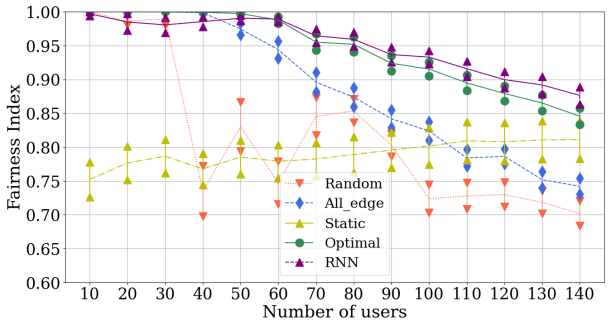


Figure 11: Fairness among all users

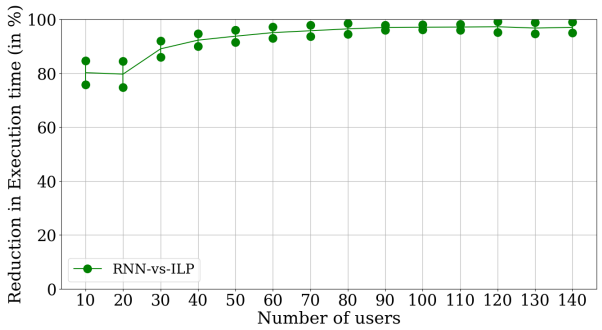
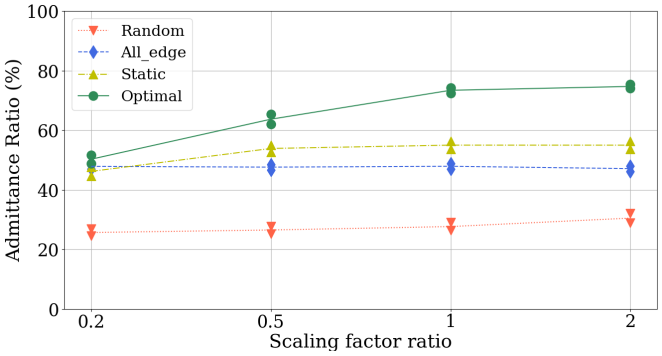
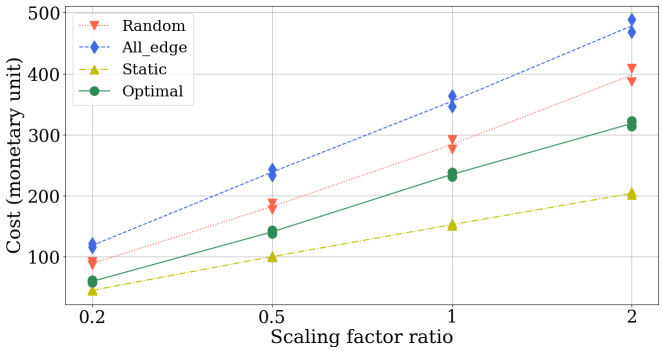


Figure 12: Reduction in execution time of the RNN model compared to the ILP model



(a) Average admittance ratio vs. scaling factor ratio



(b) O-CU deployment Cost vs. scaling factor ratio

Figure 13: Performance evaluation of different metrics vs. scaling factor ratio with 100 UEs

relatively faster when the number of users is small. Both models were executed in the exact same framework and on the same computer, emphasizing the superior efficiency of the RNN heuristic approach. Therefore, the RNN model offers a practical solution with fast execution times (around 10 milliseconds) that can be implemented in the Near RT-RIC component. It can serve as an xAPP that manages the network resources through standardized interfaces and service models in an O-RAN-compliant deployment.

B. Optimal Model under different functional split options

In this section, we evaluate the impact of the functional split option on the performance of the *Optimal* placement scenarios. For this purpose, we introduce a scaling factor ratio parameter defined as the ratio of α_{CU} over α_{DU}. As previously defined, α_{CU} and α_{DU} reflect the computational requirement (in %) of both O-CU and O-DU depending on their assigned functionalities, respectively. Increasing the scaling factor ratio signifies transferring more functionalities from O-DU to O-CU. This distribution of functionalities between O-CU and O-DU can be seen as having different functional split options. We recall that 40% of processing is done at O-RU, as earlier specified in Section IV. Thus, 60% of processing remains for both O-CU and O-DU (i.e., α_{CU} + α_{DU} = 0.6). Keeping that in mind, we test our optimal model performance with an increasing scaling factor ratio. In all our previous results, we set α_{CU} and α_{DU} to 0.1 and 0.5, respectively, resulting in a scaling factor ratio of 0.2. It is worth noting that adding more network functions to the O-CU increases the mid-haul bandwidth demand but reduces the computational demand on the O-DUs. Nonetheless, the link bandwidth is not a limiting factor in our system. Therefore, altering the functional split option can improve the efficiency of our model by encouraging centralization, as we will demonstrate later on. Fig. 13a and 13b display the average admittance ratio and the deployment cost, respectively, as a function of the scaling factor ratio for a system with 100 UEs. We note that the RNN model is not evaluated in this study as it is only trained for the scaling factor ratio of 0.2. The results clearly demonstrate the advantages of our *Optimal* placement scenario over other scenarios as the scaling factor ratio increases. This is interpreted by the fact that as the scaling factor ratio increases, the O-CU becomes more resource-demanding, making it more challenging to be placed at the edge clouds. The *Optimal* scenario solves this issue by giving the possibility for the O-CU to be hosted at the regional cloud if the latency constraints are met. The *Static* scenario has the lowest cost because it simply allows users to choose regional clouds to host O-CU and admits fewer users. This is in contrast to *All_edge* and *Random* scenarios that exhibit the lowest admittance ratio, but higher costs because they choose edge clouds to host O-CUs more often. We remark that the increase in the cost shown in Fig. 13b for all scenarios is a consequence of having more functionalities at the O-CU as the scaling factor ratio increases, and our calculations only consider the deployment cost of the O-CU. The difference in cost between all scenarios becomes more significant as the scaling factor increases; this is because the cost doubles as the scaling factor increases while the admittance remains the

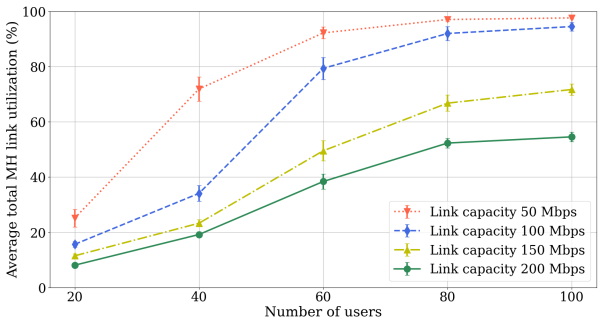


Figure 14: Average MH link utilization vs. number of users

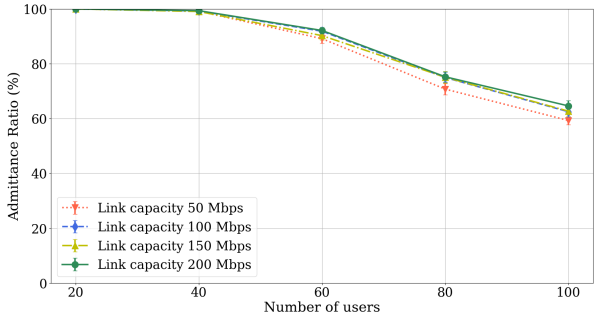


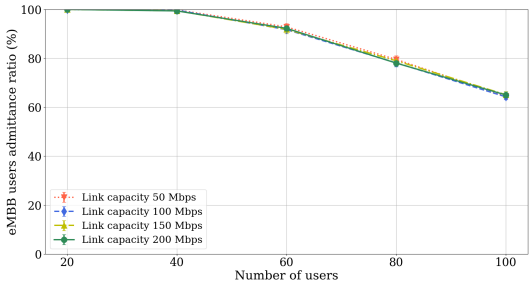
Figure 15: Average admittance ratio vs. number of users

same after the scaling factor of 1.

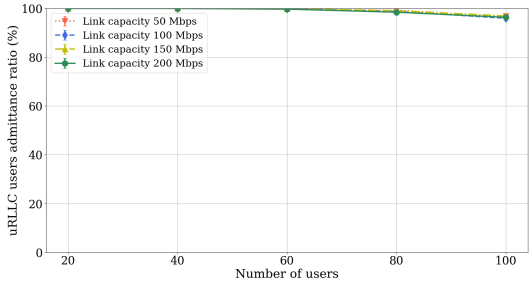
C. Optimal Model evaluation with variable link resources

In the previous sections, the computational and radio resources were the bottleneck of the system. In this section, we evaluate the performance of the optimal model with a variable link capacity. We focus on the midhaul link capacity as we deal with O-CU and O-DU placement only.

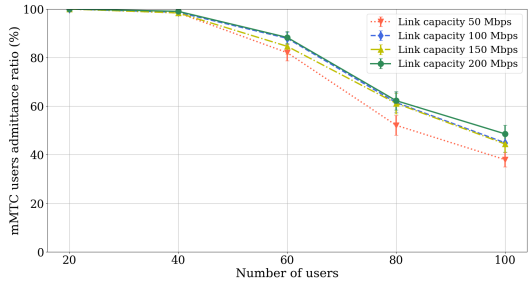
Figure 14 shows the bandwidth capacity utilization of links connecting edge servers and the regional server for different link bandwidths. The link utilization is computed as the average ratio of the utilized link bandwidth over the available link bandwidth. The link utilization becomes close to 100% when the number of users exceeds 80 UEs and with a link capacity below 100 Mbps. Beyond this value, e.g., for a link capacity of 200 Mbps, the link is no longer a bottleneck in the system with up to 55% utilization for 100 UEs. The performance of the optimal model regarding user admittance is illustrated in Figure 15. This figure demonstrates an average reduction of 6% in the admittance ratio when decreasing the link capacity from 200 Mbps to 50 Mbps for UE = 100, thus emphasizing the significant impact of introducing a bottleneck in midhaul link capacity to the system. Plots of Figure 16 show the admittance ratio for each service type. Figures 16a and 16b illustrate no difference in performance for eMBB and uRLLC users, respectively, while Figure 16c shows that the most affected users by the link capacity variation are the mMTC UEs, with 20% degradation in admittance when comparing link capacity 50 to 200 Mbps for a UE number of 100. We recall that mMTC UEs have lower priority over other services; thus, when having competence in resources, eMBB and uRLLC users are given priority to be admitted over the mMTC, and that is the case when the link is the bottleneck.



(a) Admittance ratio of eMBB users



(b) Admittance ratio of uRLLC users



(c) Admittance ratio of mMTC users

Figure 16: Admittance ratio for each service type as a function of the total number of users in the system for different link capacities

Moreover, to study the effect of varying the link capacity on the chosen server placement of the O-CUs, Figure 17 shows the percentage of admitted users' O-CUs placed at the regional cloud. As the MH link capacity increases from 50 to 200 Mbps, we notice that O-CUs tend to migrate more towards the regional server with an increasing number of users. However, in instances where the link capacity becomes a bottleneck, such as the case with 50 Mbps and more than 40 users, the Optimal model is obliged to place the O-CUs on the edge server due to the constraints imposed by the lower MH link capacity. Illustrating more the placement of the CUs of users belonging to each service type, the bar graphs of figures 18a and 18b show the optimal average percentage of CUs placement location among edge and regional servers for MH link capacity of 50 Mbps and 200 Mbps respectively. Comparing both plots emphasizes that the more link capacity is available at the MH, the more the eMBB and mMTC services mainly move to the regional server. For example, with a number of 100 users in the system, 98% of eMBB users' CUs are placed at the regional server when the link capacity is 200 Mbps % while for a link capacity of 50 Mbps, only 50% of eMBB and users' CUs are placed at the regional server.

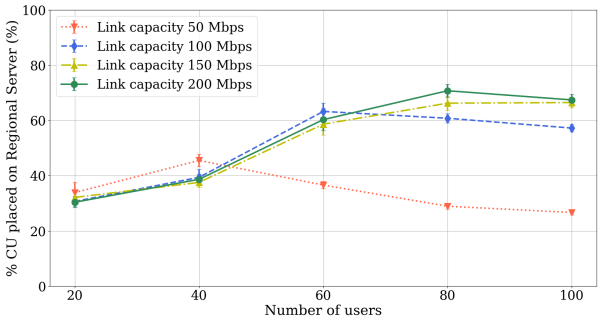


Figure 17: Percentage of admitted users' CUs placed at regional server for different link capacities

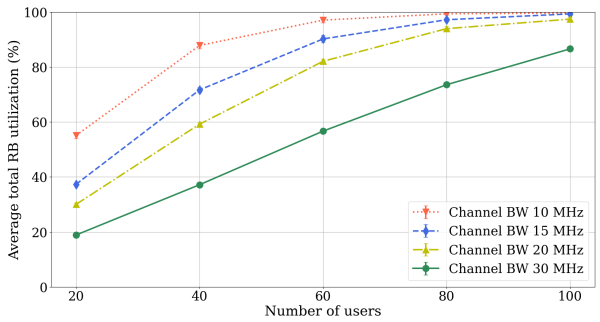
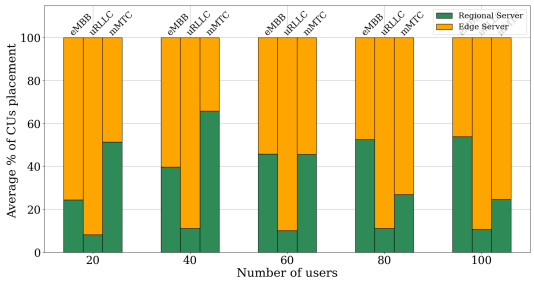
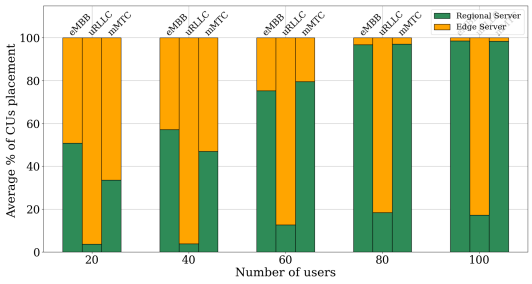


Figure 19: Average RB utilization with the number of users



(a) MH link capacity = 50 Mbps



(b) MH link capacity = 200 Mbps

Figure 18: Average ratio of CUs placed at the regional server per service type for MH link capacity of different link capacities

D. Optimal Model with variable Channel Bandwidth

In this section, we evaluate the performance of the optimal model while varying the channel bandwidth from 10 MHz to 30 MHz. Throughout our prior analyses, the system operated with a channel bandwidth fixed at 20 MHz. Increasing the operating bandwidth entails increasing the radio resources available per O-RU (more resource blocks RBs). Note that in this study, the link capacity is not a system bottleneck and is set to 1000 Mbps. Figures 19 and 20 report the utilization of radio resources (RBs) and edge servers' GOPS, respectively. The overall average RB utilization is calculated across all O-RUs in the system, representing the ratio of allocated RBs to the available RBs per O-RU. To assess the performance of the optimal model, we will interchangeably analyze these plots together, considering their correlation. It is essential to recall that user admittance requires the allocation of necessary RBs for transmission and identifying suitable placements for O-CUs and O-DUs on the available servers, respecting corresponding latency KPIs. Figure 21 depicts the average users' admittance ratio under different BW conditions, and Figure 22 showcases the admittance ratio per service type.

From Figure 21, the least admittance occurs for a BW of 10 MHz, where radio resources pose a bottleneck, with over 80% RB utilization for a user count exceeding 20 (Fig. 19). Edge servers are underutilized in this scenario, with 65% of GOPS utilization (Fig. 20). With a bandwidth of 15 MHz, significantly higher admittance is achieved (20% more than the 10 MHz case). The RBs are more abundant and sufficient for up to 80 users (Fig. 19), beyond which RUs become overloaded. However, servers remain underloaded at this stage, with around 90% GOPS utilization (Fig. 20). In the case of a BW of 20 MHz (as studied in Section VII-A), the highest admittance is attained, slightly surpassing the 15 MHz case by 5% for N = 60 users. GOPS gets overloaded faster as more users are being allocated with RBs. Both RB and GOPS become bottlenecks, with GOPS overloading at N = 60 users before RUs overload at N = 80 users, as previously analyzed in Section VII-A. For a 30 MHz BW, the same admittance as for the 20 MHz scenario is observed up to 60 users, at which point GOPS overloads (Fig. 20), preventing further user admissions. Past this threshold, we observe an 8% decrease in the overall average total admittance ratio. This decline is specifically linked to the per-slice admittance as depicted in the curves of Figure 22. Under a 30 MHz BW, all users receive adequate RBs for transmission, with RB underutilization seen in Figure 19 (85% utilization for N = 100 users). Consequently, more eMBB and uRLLC users attaining RB allocations, having already assigned higher placement priorities, will be admitted, leading to a 16% increase in eMBB user admittance for the 30 MHz BW compared to the 20 MHz BW scenario and 4% increase for that of uRLLC, as in Figures 22a and 22b respectively. However, this comes at the expense of admitting 20% fewer mMTC users, as in Figure 22c. Sacrificing the mMTC users' admission will intuitively reduce the fairness of the model, wherein Figure 23, for 30 MHz BW, the optimal model has the least fairness index of 0.82.

The last metric to evaluate pertains to the waste of radio resource utilization resulting from increasing the channel BW. We quantify RB waste as the average percentage across all O-RUs of allocated RBs for users who are not admitted. Figure 24 shows an RB waste reaching up to 30% in a system loaded with 100 users and a bandwidth of 30 MHz.

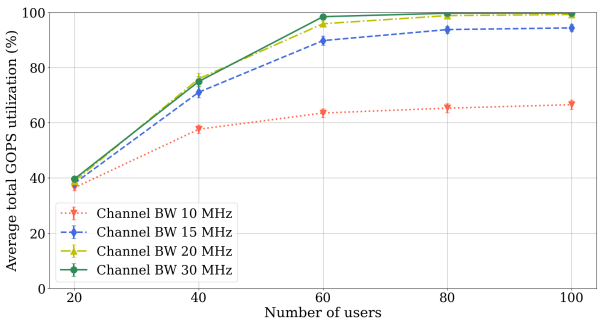


Figure 20: Average GOPS utilization with the number of users

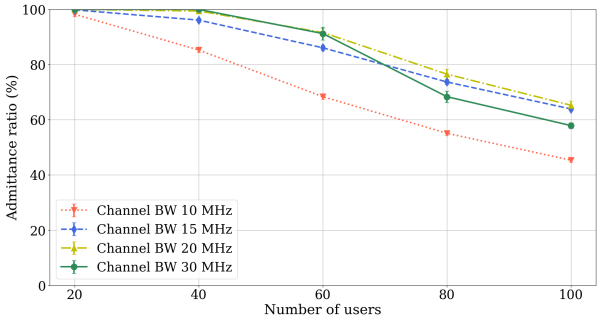
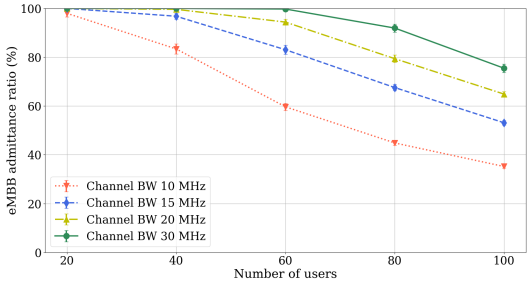
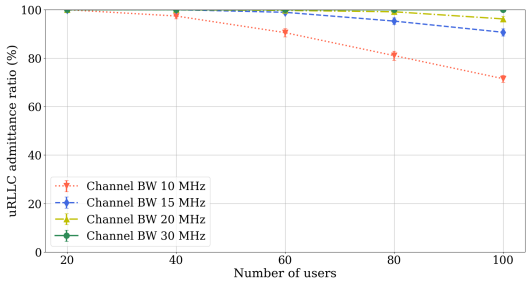


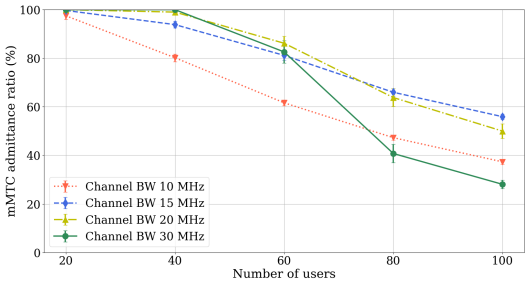
Figure 21: Average admittance ratio as a function of number of users



(a) Admittance ratio of eMBB users



(b) Admittance ratio of uRLLC users



(c) Admittance ratio of mMTC users

Figure 22: Admittance ratio for each service type as a function of the total number of users in the system for different Channel BW

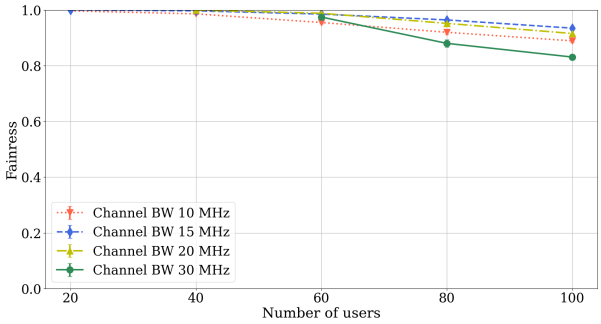


Figure 23: Fairness among all users as a function of number of users

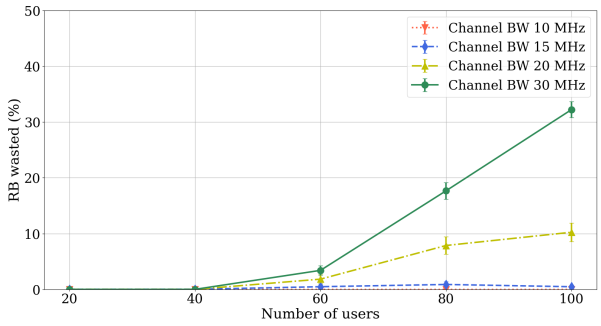


Figure 24: Average RB waste with increasing channel BW

VIII. Conclusion

Access networks are evolving toward Open RAN architecture, pushing them into a new era marked by greater openness, flexibility, and intelligence. This paper contributes significantly to solving one of the Open RAN design problems by focusing on the deployment scenarios of disaggregated network elements O-CUs and O-DUs over the edge and regional clouds. The objective is to find the optimal placement of the network functions of DUs and CUs in the O-Cloud nodes (i.e., edge and regional clouds) by considering mid-haul link delay and server capacity requirements. We propose *Optimal* model for the O-CU-DU placement mechanism that aims to maximize the number of admitted UEs while minimizing the deployment cost of O-CU by moving it towards the regional cloud. We compare our proposed optimal solution with three benchmarks, two of which are found in the literature with fixed O-CU and O-DU placement. The simulation results show that our proposed model outperforms the benchmarks. Additionally, we develop an RNN-based model that successfully mimics the *Optimal* model in a time-efficient fashion. Lastly, a comprehensive assessment of the optimal model's placement decisions allows us to quantify the efficiency and effectiveness of the proposed solution in different network conditions, including limitations over computing resources, link capacities, and radio resources. As a future work, we aim to develop a joint optimization problem for the placement problem and functional split selection, considering more dynamic scenarios and diverse service types. Additionally, for a more realistic scenario, we plan to integrate queueing and scheduling delay models, alongside propagation delay, into our model in future work. Maintaining the problem as an ILP

while incorporating these delays presents a certain challenge.

REFERENCES

[1] 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.

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

[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] O-RAN Alliance, "Technical Specification; O-RAN Control, User and Synchronization Plane Specification 11.0; O-RAN.WG4.CUS.0-R003-v11.00," Tech. Rep., March 2023.

[5] E. Bjornson and P. Giselsson, "Two Applications of Deep Learning in the Physical Layer of Communication Systems [Lecture Notes]," *IEEE Signal Processing Magazine*, 2020.

[6] M. Lee, G. Yu, and G. Y. Li, "Accelerating resource allocation for d2d communications using imitation learning," in *2019 IEEE 90th Vehicular Technology Conference (VTC2019-Fall)*, 2019.

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

[8] 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.

[9] O-RAN Alliance, "O-RAN WhitePaper - O-RAN use cases and deployment scenarios," 2020.

[10] 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.

[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] 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.

[13] 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.

[14] 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.

[15] 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.

[16] F. Z. Morais, G. M. F. de Almeida, L. Pinto, K. V. Cardoso, L. M. Contreras, R. d. R. Righi, and C. B. Both, "Placeran: Optimal placement of virtualized network functions in beyond 5g radio access networks," *IEEE Transactions on Mobile Computing*, 2023.

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

[18] 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.

[19] M. Ozturk, M. Gogate, O. Onireti, A. Adeel, A. Hussain, and M. A. Imran, "A novel deep learning driven, low-cost mobility prediction approach for 5g cellular networks: The case of the control/data separation architecture (cdsa)," *Neurocomputing*, 2019.

[20] 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.

[21] 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.

[22] Y. Xiao, J. Zhang, and Y. Ji, "Can fine-grained functional split benefit to the converged optical-wireless access networks in 5g and beyond?" *IEEE Transactions on Network and Service Management*, vol. 17, 2020.

[23] Small Cell Forum, "Small Cell Virtualization Functional Splits and Use Cases," Technical Report SCF159.07.02, January 2016.

[24] 3GPP, "Technical Specification Group Radio Access Network; NR; Physical Layer Procedures for Data," Technical Report TS 38.214, December 2019, v16.0.0, Release 16.

[25] 3GPP, "NR; User Equipment (UE) radio access capabilities," Technical Report TS 38.306, October 2022, v15.18.0, Release 15.

[26] B. Korte and J. Vygen, *Combinatorial Optimization: Theory and Algorithms*, 5th ed. Springer Publishing Company, Incorporated, 2012.

[27] R. Dhumal Deshmukh and A. Kiwelekar, "Deep learning techniques for part of speech tagging by natural language processing," in *Inter. Conf. on Innovative Mechanisms for Industry Applications (ICIMIA)*, 2020.

[28] M. Sharara, T. Pamuklu, S. Hoteit, V. Vèque, and M. Erol-Kantarci, "Policy-gradient-based reinforcement learning for computing resources allocation in o-ran," in *2022 IEEE 11th International Conference on Cloud Networking (CloudNet)*, 2022.

[29] M. Sharara, S. Hoteit, and V. Vèque, "Reinforcement learning based model for maximizing operator's profit in open-ran," in *NOMS 2023-2023 IEEE/IFIP Net. Operations and Management Symposium*, 2023.

Hiba Hojeij is currently a Ph.D. student at Paris-Saclay Univesity/CentraleSupélec, France. She received her Master's degree in Communication and Computer Networks Engineering from Politecnico di Torino, Italy, in 2021 and her Bachelor's degree in Electrical, Electronics, and Telecommunication Engineering from the Lebanese University, Lebanon, in 2019. Her research interests include resource allocation in wireless networks, Open-RAN, and machine learning applications for network optimization.

Mahdi Sharara received a diploma in electrical, electronics, computer, and telecommunications engineering from Lebanese University, Beirut, Lebanon, in 2018 and a master's degree in telecom and network from the Lebanese University and Saint-Joseph University in 2018. He received his PhD degree from Université Paris-Saclay in 2023. In 2023, he was a postdoctoral researcher at CentraleSupélec. Currently, he is a postdoctoral researcher at Orange. His research interests include Resource Allocation in Mobile Networks, Cloud-RAN, and Open-RAN, in addition to Machine Learning-based algorithms using deep learning and reinforcement learning.

Sahar Hoteit is currently an associate professor at Paris Saclay University/CentraleSupélec, France. She is also a junior member of "Institut Universitaire de France" since October 2024. She received the M.S.and PhD degree in network and computer science from the University of Pierre and Marie Curie (now Sorbonne University). Her research interests cover mobile networking, Internet of Things, game theory, Open-RAN and Cloud-RAN architectures. She has published several papers in leading international conferences (IFIP/IEEE IM, IEEE LCN, IEEE ICC, IEEE Globecom, IFIP/IEEE NOMS) and in peer-reviewed journals (i.e., IEEE Transaction on Mobile Computing, IEEE Transaction on Networking, IEEE Transaction on Network and Service Management, Elsevier Computer Networks, Elsevier Computer Communications).

Véronique Vèque obtained her PhD degree in communication networks in 1989 from University Pierre et Marie Curie - France. In 1990, she was an Associate Professor at University of Paris-Sud (Paris 11), and in 2000 to present, she worked as a full Professor at University of Paris-Sud/University Paris-Saclay. She is currently a research member of Laboratory of Signals and Systems. Her research interests lie in the field of both wireless, mobile networks, resource allocation and quality of service techniques.