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

Hiba Hojeij, Mahdi Sharara, Sahar Hoteit, Véronique Vèque

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

Hiba Hojeij, Mahdi Sharara, Sahar Hoteit, Véronique Vèque
Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire des Signaux et Systèmes (L2S), 91190, Gif-sur-Yvette, France.
Emails: {hiba.hojeij, mahdi.sharara, sahar.hoteit, veronique.veque}@centralesupelec.fr

***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 aims to improve system performance and allow 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 over O-RAN clouds become critical. In this context, we propose an optimal and dynamic placement scheme of the O-CU and O-DU functionalities either 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 VNF placement in O-RAN architecture. Additionally, we introduce a Recurrent Neural Network (RNN) heuristic model that can effectively replicate the behavior of the ILP model. We get promising results 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.**

I. Introduction

The entire telecom 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 [1] [2].

One of the fundamental principles underlying O-RAN is the splitting of RAN functionalities into three distinct components - the Open-Central Unit (O-CU), the Open-Distributed Unit (O-DU), and the Open-Radio Unit (O-RU) - through a process of disaggregation. This approach enables each component to be implemented independently, thus promoting greater flexibility and interoperability within the network [3]. Such a disaggregation yields a wide range of functional split options. The functional split in an Open RAN architecture refers to the separation of the baseband processing functions (PHY, MAC, RLC, etc.) between the O-CU and the O-DU in the radio access network. O-RAN Alliance has evaluated the different functional split options proposed by 3GPP; the selected RU/DU split option is the 7.2x split that strikes a balance between the simplicity of the radio unit and the data rates and latency required on the interface between the radio and the distributed units [3]. This recent functional split 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 split deployment options. These options comprise several configurations for placing O-CU, O-DU, and the Near-Real-Time Radio Intelligent Controller (Near-RT RIC) at regional and edge clouds. The Near-RT RIC is responsible for intelligent edge control of RAN nodes and resources. It controls RAN elements and their resources with optimization actions that typically take 10 milliseconds to one second to complete.

In this context, O-RAN envisions different strategies for deploying its functional splits on either regional or edge cloud locations, or at proprietary cell sites [4]. Fig. 1 depicts the different O-RAN Cloud deployment scenarios [5]. 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.

Unlike the commonly used static deployment strategies, our research explores the potential benefits of dynamic deployment of the O-CU and O-DU components in either edge or regional clouds. We aim at satisfying users' quality of service (QoS) requirements while improving the network's overall efficiency and performance. This approach offers more flexibility and adaptability to the network. We formulate an Integer Linear Programming (ILP) model for optimally and dynamically placing the O-CU and O-DU in the O-RAN architecture. We explore various constraints concerning the capacity of cloud servers, link capacity, and delay budget, as well as the diverse service requirements of users, including the enhanced mobile broadband (eMBB) services - which refers to services requiring high data rates, the ultra-reliable and low latency communications (URLLC) - which refers to mission-critical applications with low latency, and massive machine-type communications (mMTC) - which stands for massively

connected and energy-constrained services. Our objective is to optimize user satisfaction while meeting the needs of these services. We adopt the functional split 7.2x between O-RU and O-DU, the one selected by O-RAN alliance, and the option-2 split between O-DU and O-CU. We focus on scenarios in which the Radio Unit (O-RU) that handles the low-PHY layer functions is always located at the cell site. The Distributed Unit (O-DU), mainly responsible for high-PHY, MAC, and RLC functionalities, is deployed on the Edge cloud. Lastly, the Central Unit (O-CU) can choose either edge or regional clouds to handle RRC, SDAP, and PDCP functionalities. Our approach reflects a dynamic and flexible placement scenario between scenarios B and C of Figure 1. Our results highlight that it is possible to establish multiple connections between an O-RU and several O-DUs in an O-RAN network, under the concept known as *Shared O-RU*, defined in the *Shared-O-RU-Multi-O-DU* feature [6]. This feature is particularly useful for user dispatching, as it allows for more flexible and efficient resource allocation within the network. Our proposed solution outcomes significant benefits in terms of user admittance ratio and cost reduction compared to three baseline solutions, including a random placement of both O-CU and O-DU functionalities among regional or edge clouds and two static placements; one in which both O-CU and O-DU are on edge clouds, and the other in which the O-CU and O-DU are on regional and edge clouds, respectively. Finally, we present a heuristic to solve the optimization problem efficiently. We resort to a recurrent neural network (RNN) based model. Numerous studies have investigated the potential of deep learning (DL) to provide less-complex alternatives to highly-complex optimal algorithms [8] and [9]. DL networks not only exhibit comparable performance but also demonstrate significantly shorter execution times, resulting in being well-suited for practical deployments. Based on this perspective, we develop an RNN-based model using a bidirectional LSTM architecture trained with the output of the ILP model. The RNN-based model can mimic the optimal placement of O-CU and O-DU and achieve the desired benefits.

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

II. Related Work

Several works from the literature have addressed the placement problem of O-RAN components from different perspectives to optimize resource allocation mechanisms but with different objectives. In [10], 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 (UE) 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 [11] tackle the flexible placement of the

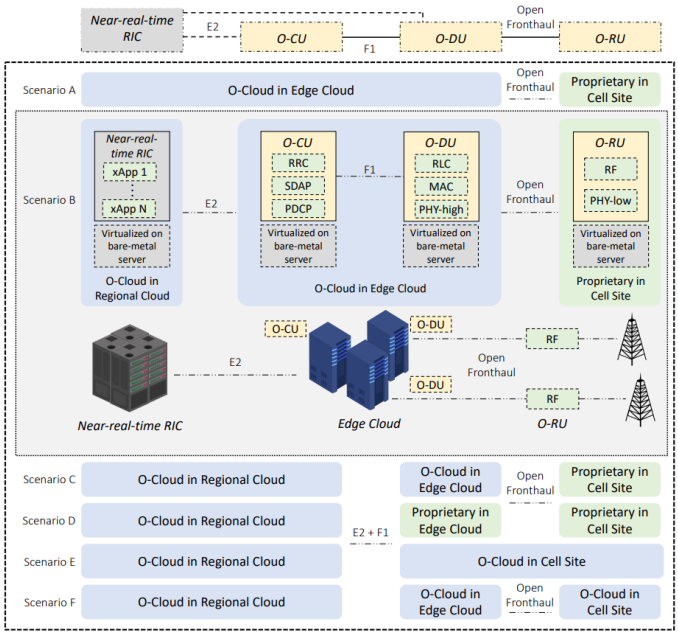


Figure 1: O-RAN Cloud deployment scenarios [5]

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 servers infrastructure is similar to their multi-tiered network infrastructure. However, they seek to maximize the profit of the infrastructure provider. Moreover, in [12], 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 CU and DU are always placed at an edge server while Near-RT RIC is located on the regional cloud. The work of [13] 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 to open-access edge servers for hosting 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 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 with the aim of maximizing 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. The next section of this paper provides a detailed description of the ILP-based proposed model.

III. 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

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 that user i chooses server s for its functionality FU
$z_{iss'}$	Binary variable indicating that user i chooses s for DU and s' for CU
$C_{ss'}$	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 the functionality FU of user i (in 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

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 [10]. 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' services 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 [14] and [17] 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 [18].

- 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 [11]:

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

  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 RU, DU, and 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 RU, 50% by DU, and 10% by CU as mentioned in [13]. 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' (that is equal to 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 [19].

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

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

To model the placement problem of O-DU and O-CU on the edge or regional clouds, we propose the following ILP-based

optimization problem:

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

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

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

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

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

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

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

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

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

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

$$\sum_{i \in \mathcal{I}} B_i^{mid} * z_{iss'} \leq C_{ss'}, s, s' \in \mathcal{S}, s = s' \tag{15}$$

$$\sum_{i \in \mathcal{I}} R_i^{CU} \theta_{is}^{CU} + R_i^{DU} \theta_{is}^{DU} \leq R_s, s \in \mathcal{S} \tag{16}$$

$$\delta_{ss'} * z_{iss'} \leq \delta_i^{mid}, i \in \mathcal{I}, s, s' \in \mathcal{S} \tag{17}$$

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 1 if $s \in S_{reg}$. This setup aims to encourage the O-DU functionality of each user 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, if this is not feasible, it will choose the nearest available server to its corresponding O-RU. Moreover, we add to the objective function a priority parameter ϵ_i as a function of the user's service type, which allows us to prioritize eMBB and uRLLC UEs over mMTC ones.

Our ILP problem 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 CU and 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 DU and 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

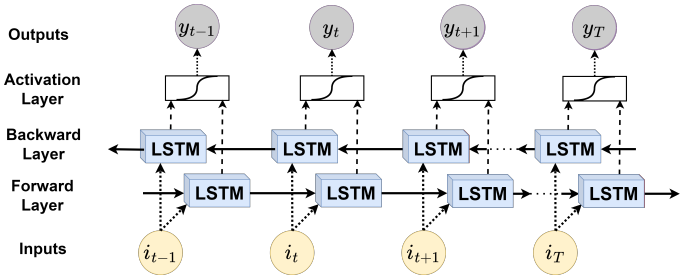


Figure 2: The Bi-LSTM layer

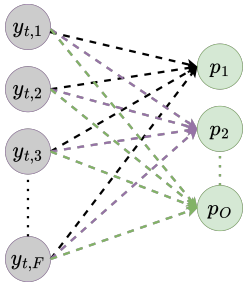


Figure 3: The fully connected layer

(8) and (9) ensure that the user's functionalities CU and DU are not allocated more than once. Constraints (10) and (11) ensure that $z_{iss'}$ is set to one only if DU and 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., if one of the DU or CU functionalities is not allocated, the whole user will be discarded. Constraint (13) enforces that the DU functionality is never allocated on a regional server. Constraints (14) and (15) ensure that the link capacity required for user i between the servers s and s' chosen for DU and 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 these latter two constraints. Server computational capacity is respected by constraint (16). And finally, the maximum link latency is guaranteed by constraint (17).

IV. Deep Learning-based Solution

Due to the high complexity of solving the NP-Hard ILP problem [21] 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) [20]

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 [20]. 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, inspired by the work in [7], that 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. The output of the BiLSTM layer is then passed through a sigmoid activation layer, producing an output vector of F elements. These outputs are 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 are the combination of locations of O-CU and O-DU among the available servers, plus an additional label to indicate that a user is dismissed. The neuron with the highest activation value corresponds to the decision.

To generate the training dataset, we conducted 25,000 experiments using the ILP model.

V. 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 [13], 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 we consider a number of 10 to 140 UEs spread following the distribution presented in [13] 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 [13].

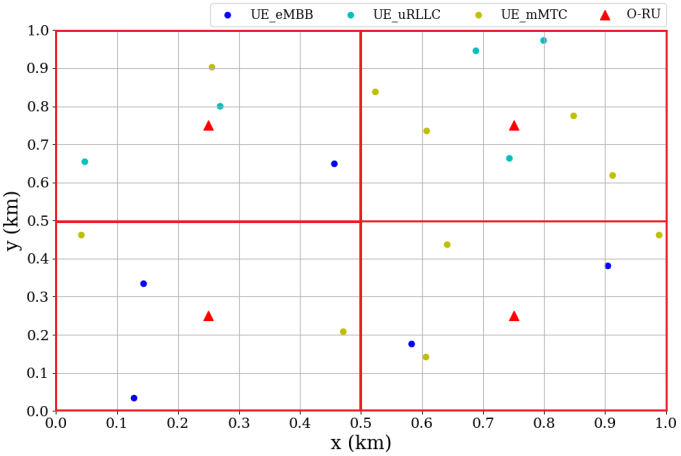


Figure 4: An example of network topology with 20 UEs.

TABLE II: Radio parameters used in our simulations

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

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 [11]. The mid-haul latency bounds δ_i^{mid} are considered as in [13] 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 [13] 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 [16]. 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 [11]. We note that the MCS index impacts the code rate and spectral efficiency, as referred to in 3GPP specification [18]. 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 [13]. Additional radio parameters used in the experiments are outlined in Table II. We note that our ILP-based problem is solved using IBM CPLEX software [15], a mathematical optimization solver, on a computer with 11th generation Intel® Core™ i9-11950H Processor and 16 GB RAM.

VI. Performance Evaluation

In this section, we compare the performance of our proposed algorithm, referred to as the *Optimal scenario*, along

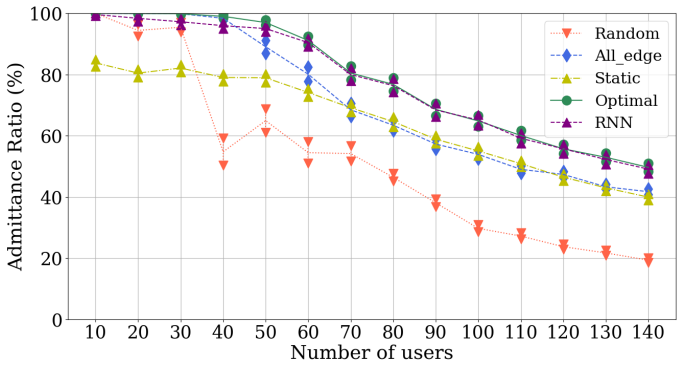


Figure 5: Average admittance ratio as a function of the number of users in the system

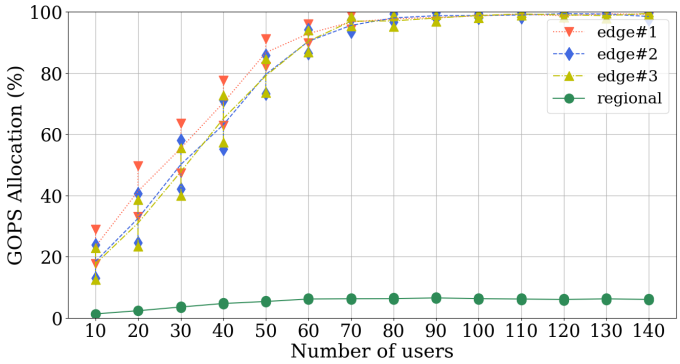


Figure 6: GOPS allocation per server in Optimal deployment scenario

with the heuristic based on *RNN* model with respect to three baselines defined as follows:

- An *All_Edge scenario*; in which only the edge servers are available (i.e., O-CUs and O-DUs are always on the edge servers). This corresponds to 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. This refers to scenario C of Fig. 1.
- A *Random* scenario; where a random selection of servers between edge and regional is adopted for both O-DUs and O-CUs.

The performance metrics used in this paper are as follows:

- 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, whether they are regional or edge servers. It is computed as the cost of running the computational operations on a server (in GOPS). As the regional server has more processing capacity and uses less energy than the edge server, running VNFs in regional servers is less expensive than in edge servers [10]. At the edge server, according [13] [14], 1 GOPS costs 1.59$, while at the regional cloud, it costs 0.5$/GOPS.

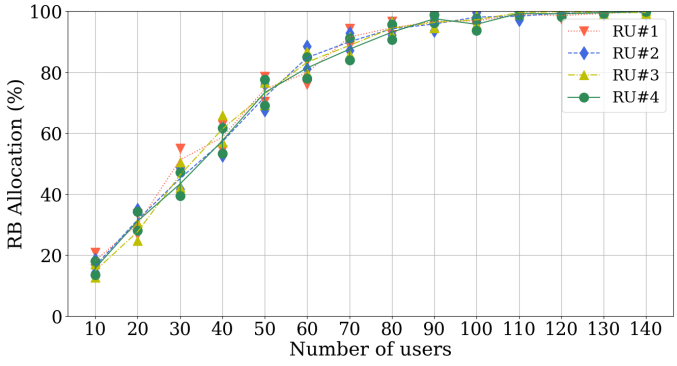


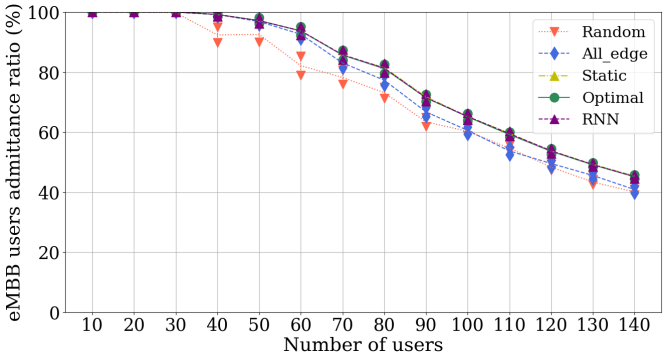
Figure 7: RB allocation as increasing the number of users

- 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 (18) as follows:

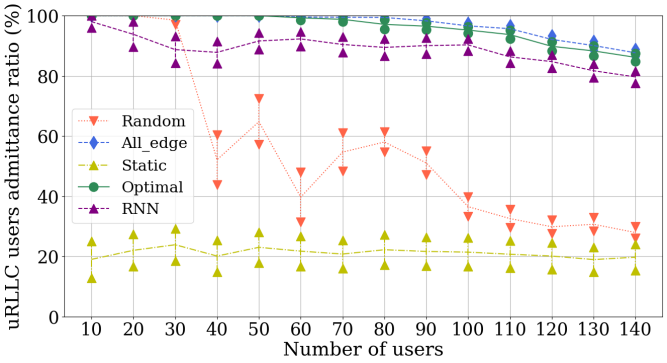
$$\zeta = (\sum_{j=1}^{N} AAR_j)^2 / (N \cdot \sum_{j=1}^{N} AAR_j^2) \qquad (18)$$

where $N = 3$ refers to the number of heterogeneous service types, AAR_j is the average admittance ratio of users of service type j.
- Scaling factor ratio: It is defined by 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 III. 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.

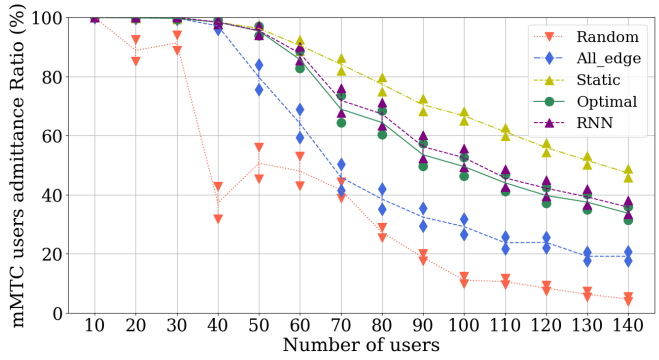
We note that 100 simulations were performed, and confidence intervals of 95% are provided in the following results. We start our evaluation by analyzing the average admittance ratio as a function of the number of users for each considered scenario. The results, as depicted in Fig. 5, demonstrate that 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 is able to



(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 as a function of the total number of users in the system

closely replicate the optimal model's admittance ratio, with a difference of no more than 2% compared to the optimal solution. Additionally, we clearly see from the figure that 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

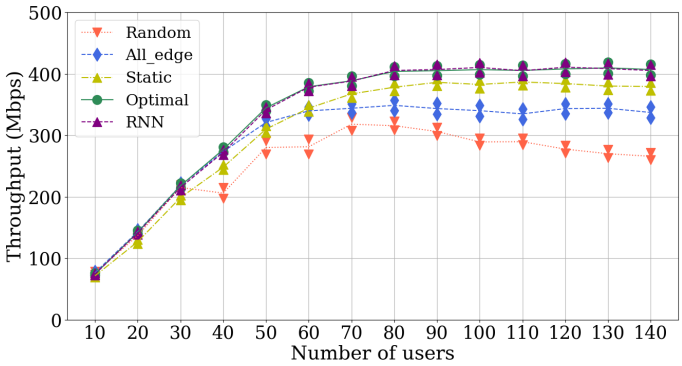


Figure 9: Total throughput as the function of the number of users in the system

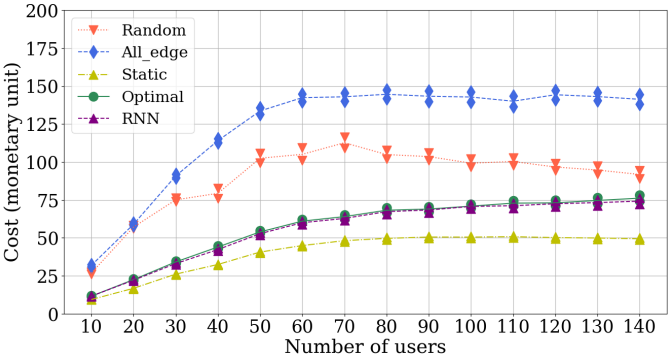


Figure 10: O-CU deployment costs for each scenario

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. In order 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 given higher priority over mMTC services, and when cloud computational resources are available and satisfy their latency requirements, they are allocated accordingly. 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

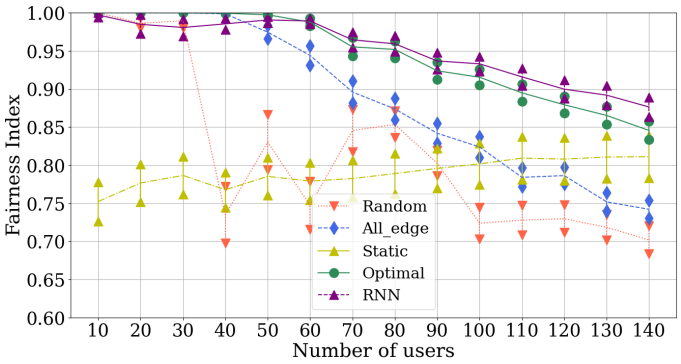


Figure 11: Fairness among all users as a function of the number of users

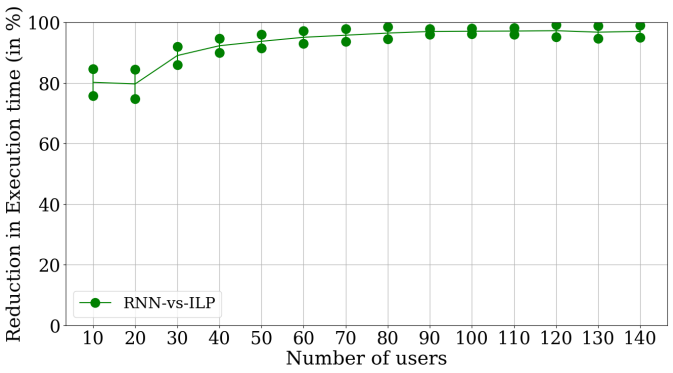


Figure 12: Reduction in Execution time (in %) of the RNN model compared to the ILP model as a function of the number of users

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.

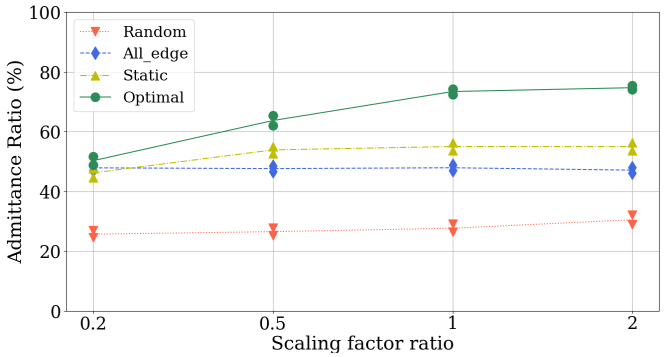
In terms of throughput, Fig. 9 illustrates the overall throughput achieved by deploying different placement scenarios. It is evident that the *Optimal* scenario outperforms all other scenarios in terms of throughput. This result is consistent with the higher average admission ratio achieved by the *Optimal* 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 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.

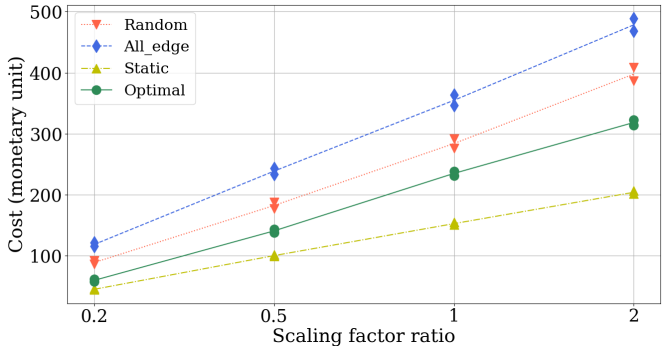
An important consideration is the fairness of the admittance ratio among the three service types as the number of users in the system changes. The results are shown in Fig. 11 for the different scenarios. The *Optimal* scenario offers a better fairness index among users compared to the other scenarios. 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 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.

The last research question we raise in this paper is the impact of the functional split option on the performance of the *Optimal* placement scenarios. To address this question, we refer to the previously defined scaling factor ratio (introduced as a ratio of α_{CU} over α_{DU}). Varying this ratio changes the percentage of computational capacity required by the O-CU and O-DU, which reflects different functional split options. 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



(a) Average admittance ratio versus scaling factor ratio



(b) O-CU deployment Cost versus scaling factor ratio

Figure 13: Performance evaluation of different metrics as a function of scaling factor ratio when having 100 UEs

scaling factor increases; this is because the cost doubles as the scaling factor increases while the admittance remains the same after the scaling factor of 1.

VII. 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 CU-DU placement mechanism that aims to maximize the number of admitted UEs while minimizing the deployment cost of 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 CU and 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. 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. We also consider taking into account the radio resource allocation for users connected to the different O-RUs and the front-haul link capacities for an adaptive placement of O-RAN components.

Acknowledgment

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

References

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

[2] *L. Bonati, S. D'Oro, M. Polese, S. Basagni, and T. Melodia, "Intelligence and Learning in O-RAN for Data-Driven NextG Cellular Networks," in IEEE Communications Magazine, October 2021.*

[3] *M. Polese, L. Bonati, S. D'Oro, S. Basagni and T. Melodia, "Understanding O-RAN: Architecture, Interfaces, Algorithms, Security, and Research Challenges," in IEEE Communications Surveys & Tutorials.*

[4] *O-RAN Alliance White Paper, O-RAN use cases and deployment scenarios, 2020.*

[5] *L. Bonati, M. Polese, S. D'Oro, S. Basagni, T. Melodia, Open, programmable, and virtualized 5G networks: State-of-the-art and the road ahead, Comput. Netw. 182 (2020) 1–28.*

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

[7] *M. Sharara, S. Hoteit and V. Vèque, "A Recurrent Neural Network Based Approach for Coordinating Radio and Computing Resources Allocation in Cloud-RAN," 2021 IEEE 22nd International Conference on High Performance Switching and Routing (HPSR), Paris, France, 2021.*

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

[9] *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.*

[10] *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," in IEEE Trans. on Network and Service Management, 2022.*

[11] *E. Sarikaya and E. Onur, "Placement of 5G RAN Slices in Multi-tier O-RAN 5G Networks with Flexible Functional Splits," 2021 17th Intern Conference on Network and Service Management (CNSM), 2021.*

[12] *I. Tamim, A. Saci, M. Jammal and A. Shami, "Downtime-Aware O-RAN VNF Deployment Strategy for Optimized Self-Healing in the O-Cloud," 2021 IEEE Global Communications Conference (GLOBECOM), 2021.*

[13] *S. Mondal and M. Ruffini, "Optical Front/Mid-Haul With Open Access-Edge Server Deployment Framework for Sliced O-RAN," in IEEE Tran on Network and Service Management, Sept. 2022.*

[14] *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 Trans on Network and Service Management, 2020.*

[15] *Cplex, I. I. (2009). V12. 1: User's Manual for CPLEX. International Business Machines Corporation, 46(53), 157.*

[16] *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 Inter. Conf. on Cloud Networking (CloudNet), Paris, France, 2022.*

[17] *"Small cell virtualization functional splits and use cases," Tech. Rep. SCF159.07.02, Small Cell Forum, January 2016.*

[18] *Technical Specification Group Radio Access Network; NR; Physical Layer Procedures for Data, V16.0.0, Release 16, 3GPP Standard TS 38.214, Dec. 2019.*

[19] *NR; User Equipment (UE) radio access capabilities". User Equipment (UE) radio access capabilities, V15.18.0, Release 15, 3GPP TS 38.306, Oct. 2022.*

[20] *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.*

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