Dynamic Orchestration of Disaggregated RAN functions in Open-RAN Architecture

Hiba Hojeij*, Guilherme I. Ricardo†, Mahdi Sharara*, Sahar Hoteit*,
Véronique Vèque*, Nancy Perrot‡ and Stefano Secci§
*Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire des Signaux et Systèmes (L2S), 91190, Gif-sur-Yvette, France
†IRIT/Toulouse INP - ENSEEIHT, University of Toulouse, France,
‡Orange Labs, Chatillon, France,
§CEDRIC, CNAM, France
Emails: firstname.lastname@centralesupelec.fr, guilherme.ricardo@irit.fr, nancy.perrot@orange.com, stefano.secci@lecnam.net

***Abstract*—In modern Open RAN architectures, the classic gNB radio protocol stack is disaggregated and implemented in different virtualized components, the Centralized Unit (CU), the Distributed Unit (DU), and the Radio Unit (RU). Each of these units is deployed throughout the RAN physical infrastructure in order to achieve users' required Quality of Service (QoS). Within this framework, our study is dedicated to optimizing the admission of User Equipments (UEs) into the system while ensuring guaranteed QoS. To do this, we focus on two key tasks: (i) establishing associations between UEs and RUs and (ii) strategically positioning the CUs and DUs of each UE across nodes in the network edge and regional clouds. We initially address these tasks by proposing a joint optimization problem of association and placement, subject to the system's available resources and QoS-related constraints. Then, we propose a sequential model for problem-solving. We formulate both scenarios as NP-Hard Integer Linear Programming (ILP) optimization problems and discuss their properties. We show through exhaustive simulations that our algorithms converge to good results in practice, outperforming other state-of-the-art techniques. Moreover, a comparative analysis between the joint and sequential models reveals a notable trade-off. The sequential model performs close to the performance of the joint model, showcasing a 6% reduction in user admittance, yet it excels with a remarkable 80% reduction in execution time.**

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

I. Introduction

Traditional radio access networks (RANs) have historically been characterized by proprietary, vertically integrated solutions, resulting in vendor lock-in and limited operational flexibility for network operators. However, in response to the rapidly evolving scenario of 5G cellular networks, a joint effort to promote *Open RAN* standardized architectural solutions was founded by *O-RAN Alliance* [1], gathering a vast range of academic and industrial partners. *Open RAN* offers a paradigm shift by advocating for disaggregation and standardization, fostering interoperability and vendor diversity [2]. This approach enables network operators to leverage a diverse ecosystem of hardware and software components from multiple vendors, promoting innovation and competition. In its most recent technical report [3], O-RAN consolidates the implementation (and extension) of the *3GPP 7.2x Split* architecture for gNB disaggregation. In 7.2x Split, the 3GPP's radio protocol stack implemented in classic gNBs is separated into three different functional components or units: (i) Open Radio Units (O-RU), (ii) Open Distributed Units (O-DU), and (iii) Open Centralized Units (O-CU). Each functional unit can either be implemented physically or virtually. O-RAN has already adopted virtualization and is considered the succeeding iteration of vRAN (Virtualized Radio Access Network) [1], offering enhanced capabilities by adding flexible RAN provisioning based on application needs and standardizing open interfaces among RAN components. As introduced in [3], the use of cloud architectures (generically referred to as O-Cloud) leverages different strategies to deploy functional components (e.g., scenario B and C among others [3]; where scenario B refers to the case where the DUs and CUs are situated at the edge cloud, and scenario C where the DUs and the CUs are at the edge and the regional cloud, respectively). This paradigm shift allows the dis-aggregated functional units to be deployed at various network nodes, which can encompass a range of locations such as cell sites, edge clouds, or regional clouds. By distributing these functional units across different nodes, network operators gain the flexibility to adapt to specific network demands and constraints. For instance, they can strategically position functional components closer to end-users at cell sites to reduce latency for applications with stringent delay requirements. Alternatively, they may choose to deploy these components in edge or regional clouds, where there might be more computing resources available to handle complex processing tasks. Thus, on the one hand, the disaggregation and distribution of functional units throughout the O-Cloud network provide flexibility, and on the other hand, its deployment must be carefully designed to satisfy challenging constraints. Moreover, part of the innovative characteristics of the O-RAN architecture is the implementation of controller components, which are called RAN Intelligent Controllers (RICs). Depending on the scope and time scale of which they operate, the RICs can be categorized as Non-Real Time (Non-RT) and Near-Real Time (Near-RT) RIC [2].

In this paper, we investigate how to leverage the distribution of functional units, mainly CUs and DUs, to increase the admissibility of users under different communication service requirements. We propose an optimization model for maximizing the number of users we can accommodate in the network while satisfying their communication requirements. We propose to control the admittance level by (i) establishing a coherent end-to-end data flow for each user throughout the O-Cloud network and (ii) assigning sufficient resource blocks (RBs). An O-Cloud can be either an edge cloud or a regional cloud. We initially formulate this problem of users-to-RU association and CU/DU to O-Cloud placement jointly as an Integer Linear Program (ILP) designed to optimize user admittance within the network during each transmission time interval (TTI). Furthermore, we propose another approach to address the optimization problem sequentially, where the UE-to-RUs association and CU/DU placement are tackled through separate sequential ILPs. This is motivated by the deployment of Non-RT RIC and Near-RT RIC in the O-RAN ecosystem, that promotes the flexibility of placing RAN operations based on their required time scale. We evaluate both proposed models, and we show the superiority of our approaches over other baselines that adopt a fixed placement at the cloud level and fixed user association at the radio level with respect to different evaluation metrics. Moreover, we investigate the advantage of adopting a sequential model in terms of computational complexity.

The rest of this paper is organized as follows: Section II provides a brief overview of the related work. The system model and our proposed ILP-based models 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

Various studies have tackled the optimization of radio function placement in the context of evolving RAN architectures. The emergence of O-RAN architecture [1], has paved the way for new research directions in this area.

In [4], authors combine RB allocation and DU selection to enhance energy efficiency and ensure low-latency traffic within the O-RAN architecture. They propose an energy-aware optimization model that jointly addresses RB allocation and DU selection. Moreover, authors in [5] introduce a dynamic DU placement approach, allowing flexibility in DU positioning throughout the network for the sake of minimizing O-RAN costs. However, these works maintain fixed CU locations, which may result in sub-optimal outcomes.

Authors in [6] propose a deep reinforcement learning method to determine optimal O-Cloud locations for O-DU and O-CU Virtualized/Cloud-native network functions (VNFs/CNFs) and establish optimal user equipment (UE) to O-RU associations. Their primary objective is to minimize latency and reduce deployment costs. However, their model lacks consideration of the diverse service requirements of different users. In [7], authors address the optimization problem of efficiently placing DUs and CUs, considering the distributed nature and limited capacity of processing pools with the aim of minimizing the number of active processing pools and total network latency. Notably, their work does not account for slice-specific requirements and does not specifically address the users' association with Radio Units (RUs).

Additionally, our prior work in [8] concentrates on the dynamic placement of CUs and DUs but overlooks the critical aspect of UE-to-RU association, which plays a crucial role in optimizing network performance. However, in this paper, we jointly address these challenges, aiming to maximize users' admittance ratio while meeting the diverse Quality of Service (QoS) requirements of different slices. The next section provides a detailed description of our system model.

III. System Model

The system consists of a set $\mathcal{R}$ of RUs located across a fixed squared area of side L, such that each RU $r \in \mathcal{R}$ has a position determined by coordinates $P_r = (X_r, Y_r) \in [0, L]^2$. We hereafter refer to the geographical deployment of RUs as the *cell site*. Consider a set $\mathcal{U}$ of UEs arbitrarily located across the cells site, such that the position of each $u \in \mathcal{U}$ is given by coordinates $P_u = (X_u, Y_u) \in [0, L]^2$.

The O-Cloud network is modeled as a graph $\mathcal{G} = (\mathcal{H}, \mathcal{E})$, where $\mathcal{H}$ is the set of vertices, representing the cloud hosts, and $\mathcal{E}$ is the set of edges, representing the physical links connecting two neighboring hosts. $\mathcal{H}$ is further partitioned in two domains: the set of edge-cloud hosts $\mathcal{H}_E$ and the set of regional-cloud hosts $\mathcal{H}_R$, such that $\mathcal{H} = (\mathcal{H}_E \cup \mathcal{H}_R)$. For each host h, if $h \in \mathcal{H}_E$, then it is located at $P_h \in [L, L']^2$, otherwise, if $h \in \mathcal{H}_R$, its location is $P_h \in [L'', L''']^2$. Each RU in the cells site is fully connected to edge-cloud hosts.

Each UE requests the provision of a communication service from the set of slices $\mathcal{S}$. Each slice has Quality of Service (QoS) requirements in terms of (i) achieved data rate (throughput) and (ii) end-to-end (E2E) delay[1]. If the system is currently able to meet all the QoS requirements of a given UE's slice, then it *admits* the UE and provides the requested communication service. In the following, we discuss the system's characteristics that impact the provision of UEs' required QoS and consequently determine its admittance.

A. UE-RU Association

Each UE is within reach of potentially multiple RUs simultaneously and, if it is admitted to the system, it must be associated with one of its neighboring RUs. The UE-RU association decision is captured by variables $x^{\text{RU}}_{u,r} \in \{0, 1\}, \forall u \in \mathcal{U}, \forall r \in \mathcal{R}$,indicating whether UE u is associated with RU r, $x^{\text{RU}}_{u,r} = 1$, or not, $x^{\text{RU}}_{u,r} = 0$.

The vector of association variables is denoted by $\mathbf{x}^{\text{RU}} = [x^{\text{RU}}_{u,r} : \forall u \in \mathcal{U}, \forall r \in \mathcal{R}]$.

We assume classic OFDMA scheduling, such that RUs' time-bandwidth is split into radio resource blocks (RBs) that can be assigned to associated UEs. Each RU $r \in \mathcal{R}$ has a total of $M_r \in \mathbb{Z}_+$ RBs that are further distributed among all slices in $\mathcal{S}$. We introduce variables $\rho_{r,s} \in \mathbb{Z}_+, \forall r \in \mathcal{R}, \forall s \in \mathcal{S}$, to capture the number of RBs dedicated to slice s at RU r.

The number of RBs $\text{RB}_{u,r}$ required by user u if associated to RU r is computed as following

$$\text{RB}_{u,r} \triangleq \left\lceil \frac{\lambda_{s(u)}}{\eta_{u,r}} \right\rceil, \forall u \in \mathcal{U}, \forall r \in \mathcal{R}\,. \tag{1}$$

where $s(u) \in \mathcal{S}$ is the slice requested by UE u, $\lambda_s \in \mathbb{R}_+$ is the data rate required by slice $s \in \mathcal{S}$ and $\eta_{u,r} \in \mathbb{R}_+$ is the (wireless) link capacity per RB measured using the principles of Shannon theory as in [9]. We note that a user u assigned to an RU r is supposed to get its required number of RBs in order to transmit at its required data rate. The number of RBs assigned to a UE u is determined as follows

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

B. DU-CU Placement

We consider as in [8] a hybrid deployment scenario between scenarios B and C, defined in [10], where the DU functions are implemented at the edge cloud, while the CU functions can be on either the edge or regional clouds. Firstly, we introduce variables $x^{\text{DU}}_{u,h} \in \{0, 1\}, \forall u \in \mathcal{U}, \forall h \in \mathcal{H}$, indicating whether UE u's DU is placed at cloud host h (i.e., $x^{\text{DU}}_{u,h} = 1$) or

[1] We consider a scenario with symmetric requirements for both uplink and downlink. For the sake of presentation and without loss of generality, we define our model exclusively in terms of downlink.

not (i.e., $x^{\text{DU}}_{u,h}=0$)[2]. We denote the vector of DU-placement variables by $\mathbf{x}^{\text{DU}}=[x^{\text{DU}}_{u,h}:\forall u\in\mathcal{U},\forall h\in\mathcal{H}]$. Correspondingly, we introduce variables $x^{\text{CU}}_{u,h}\in\{0,1\},\forall r\in\mathcal{R},\forall h\in\mathcal{H}$, indicating whether UE u's CU, specifically its User Plane component (CU-UP), is placed at cloud host h (i.e., $x^{\text{CU}}_{u,h}=1$) or not (i.e., $x^{\text{CU}}_{u,h}=0$). The vector of CU-placement variables is denoted by $\mathbf{x}^{\text{CU}}=[x^{\text{CU}}_{u,h}:\forall u\in\mathcal{U},\forall h\in\mathcal{H}]$. The vector of association-placement variables is denoted by $\mathbf{x}=[\mathbf{x}^{\text{RU}},\mathbf{x}^{\text{DU}},\mathbf{x}^{\text{CU}}]$.

C. O-Cloud Computation Model

Each cloud host has enough computational capacity (RAM and CPU) to run a limited number of functional unit instances. Each instance of functional unit (FU), i.e., O-DU and O-CU, has an associated computational cost [11], given in *Giga Operations Per Second* (GOPS), that is defined as follows

$$g^{\text{FU}}_u(\mathbf{x}^{\text{RU}})\triangleq\frac{\alpha_{\text{FU}}\cdot(3A+A^2+M\cdot C\cdot L/3)}{10}\cdot\text{RB}_u(\mathbf{x}^{\text{RU}}),\tag{3}$$

where FU is replaced with either CU or DU, M represents the modulation bits (i.e., the number of bits per symbol), C denotes the coding rate, L is the number of MIMO layers, A corresponds to the number of antennas, and $\text{RB}_u(\mathbf{x})$ is the number of resource blocks assigned to user u, as defined in (2). The constants α_{DU} and α_{CU}, defined for each FU, serve as a scaling factor representing the average computational load of O-DUs and O-CUs, respectively, with respect to their total computational requirements. Specifically, in our system, we adopt the Split-7.2x between O-RU and O-DU and the Split-2 between O-DU and O-CU, and based on the computational load distribution described in [12], we assign $\alpha_{DU}=50\%$ and $\alpha_{CU}=10\%$ of the computational workload to the DU and the CU, respectively (the O-RU is in charge of the remaining 40%). We formalize the total computational utilization in node h as

$$g_h(\mathbf{x}^{\text{RU}})\triangleq\sum_{u\in\mathcal{U}}g^{\text{CU}}_u(\mathbf{x}^{\text{RU}})\cdot x^{\text{CU}}_{u,h}+g^{\text{DU}}_u(\mathbf{x}^{\text{RU}})\cdot x^{\text{DU}}_{u,h},\tag{4}$$

where the computational cost functions $g^{\text{CU}}_u(\cdot)$ and $g^{\text{DU}}_u(\cdot)$ are defined in (3).

D. E2E Delay Model

We consider that the E2E delay experienced by a given UE is primarily affected by the total propagation delay at Midhaul (MH) and Fronthaul (FH) links. For each UE u, the MH delay is measured between the CU to the DU, is given by:

$$d^{\text{MH}}_u(\mathbf{x})\triangleq\frac{||P_h-P_{h'}||}{v_{\text{Fiber}}}\cdot x^{\text{CU}}_{u,h}\cdot x^{\text{DU}}_{u,h'},\tag{5}$$

where $v_{\text{Fiber}}\in\mathbb{R}_+$ is the propagation speed over fiber, and $\|\cdot\|$ represents the Euclidean distance between two hosts. Similarly, for each UE u, the FH delay, i.e., from the DU to the RU, is defined as

$$d^{\text{FH}}_u(\mathbf{x})\triangleq\frac{||P_r-P_h||}{v_{\text{Fiber}}}\cdot x^{\text{DU}}_{u,h}\cdot x^{\text{RU}}_{u,r}.\tag{6}$$

IV. Problem Definition

Our goal is to optimize the system performance by maximizing the admittance of UEs that is conditioned to the system's capability to satisfy their requested services' requirements.

A. Joint ILP Model

We formulate the ILP-based optimization problem as follows:

Problem 1 (Joint Problem)**.**

$$\underset{\mathbf{x},\boldsymbol{\rho}}{\text{maximize}}\quad a_{\textit{Joint}}(\mathbf{x})=\sum_{u\in\mathcal{U}}\sum_{r\in\mathcal{R}}\sum_{h,h'\in\mathcal{H}}\epsilon_{s(u)}\cdot x^{\text{RU}}_{u,r}\cdot x^{\text{DU}}_{u,r}\cdot x^{\text{CU}}_{u,r}\tag{7}$$

$$\text{subject to}\quad\sum_{h,h'\in\mathcal{H}}x^{\text{DU}}_{u,h}\cdot x^{\text{CU}}_{u,h'}=\sum_{r\in\mathcal{R}}x^{\text{RU}}_{u,r},\forall u\in\mathcal{U}\tag{8}$$

$$\sum_{h\in\mathcal{H}_R}x^{\text{DU}}_{u,h}=0,u\in\mathcal{U}\tag{9}$$

$$\sum_{s\in\mathcal{S}}\rho_{r,s}\leq M_r,r\in\mathcal{R}\tag{10}$$

$$\textit{RB}_u(\mathbf{x}^{\text{RU}})\leq\rho_{r,s},r\in\mathcal{R},s\in\mathcal{S}\tag{11}$$

$$g_h(\mathbf{x}^{\text{RU}})\leq G_h,\forall h\in\mathcal{H}\tag{12}$$

$$d^{\textit{MH}}_u(\mathbf{x})\leq D^{\textit{MH}}_u,\forall u\in\mathcal{U}\tag{13}$$

$$d^{\textit{FH}}_u(\mathbf{x})\leq D^{\textit{FH}}_u,\forall u\in\mathcal{U}\tag{14}$$

$$x^{\textit{RU}}_{u,r}\in\{0,1\},\forall u\in\mathcal{U},\forall r\in\mathcal{R}\tag{15}$$

$$x^{\text{CU}}_{u,h},x^{\textit{DU}}_{u,h}\in\{0,1\},\forall u\in\mathcal{U},\forall h\in\mathcal{H}\tag{16}$$

$$\rho_{r,s}\in\mathbb{Z}_+,\forall r\in\mathcal{R},\forall s\in\mathcal{S}\tag{17}$$

The objective function (7) aims to maximize the number of admitted UEs weighted by a priority $\epsilon_{s(u)}$ defined for each slice requested by UE u, $s(u)$. Constraints (8) ensures that an admitted UE u has exactly one functional unit of each type associated to it. Additionally, due to delay limitations [10], we consider that DUs must be deployed at the vicinity of the cell sites, so they can only be placed in the edge-cloud domain. On the other hand, CUs can be deployed in both edge or reginal domains. This limitation is captured by (9).

In (10), we guarantee that the total amount of resources of RU r assigned to each slice does not exceed its total number of resource blocks M_r. Considering that every slice has different RB requirements, the number of UEs of slice s that RU r can accommodate is limited to its maximum amount of RBs $\rho_{r,s}$ dedicated to that slice. We represent these constraints in (11). In (12), we ensure that the computational utilization at each node h does not exceed its available computational capacity G_h. Finally, in (13) and (14), we enforce that both the MH and FH delays satisfy their tolerance values D^{MH}_u and D^{FH}_u, respectively. We refer to the (optimal) solution of Problem 1 as $\mathbf{x}^*$ and $\boldsymbol{\rho}^*$ formalized in Algorithm 1.

Proposition IV.1. *Problem 1 is NP-Hard.*

Proof. To prove the NP-hardness of our problem, we perform a reduction to the well-known NP-hard 0-1 knapsack problem [13]. A restricted instance of our Problem 1 is inherently as challenging as the knapsack problem, a known NP-hard problem. Hence, our problem is also NP-hard. □

Although Problem 1 is NP-Hard, we discuss in section IV-C how we can linearize it to find exact solutions in practice.

[2] We consider the "Shared-RU" framework introduced in [10, Chapter 14], where each RU may have multiple associated DUs. We further assume that DUs belonging to the same RU context coordinate to perform UE scheduling.

Algorithm 1 JOINT

```
input : $G = (\mathcal{H}, \mathcal{E})$, $\mathcal{R}, \mathcal{U}, \mathcal{S}$,
        Functions $RB_{u,r}(\cdot)$, $g_u^{\text{FU}}(\cdot)$, $d_u^{\text{MH}}(\cdot)$, $d_u^{\text{FH}}(\cdot)$, and
        Parameters $P_r, P_u, P_h, M_r$ , $\boldsymbol{\alpha}_{CU}$, $\boldsymbol{\alpha}_{DU}$.
output: Feasible Admission Setup $(\mathbf{x}^*, \boldsymbol{\rho}^*)$
$(\mathbf{x}^{*\text{RU}}, \mathbf{x}^{*\text{DU}}, \mathbf{x}^{*\text{CU}}, \boldsymbol{\rho}^*) \leftarrow \underset{\mathbf{x},\boldsymbol{\rho}}{\text{argmax}}\{a_{\text{Joint}(\mathbf{x})} : (8), (9), (10),$
 $(11), (12), (13), (14), (15), (16), (17)\}$
 return $x^*$
```

Algorithm 2 SEQUENTIAL

```
input : $G = (\mathcal{H}, \mathcal{E})$, $\mathcal{R}, \mathcal{U}, \mathcal{S}$,
        Functions $RB_{u,r}(\cdot)$, $g_u^{\text{FU}}(\cdot)$, $d_u^{\text{MH}}(\cdot)$, $d_u^{\text{FH}}(\cdot)$, and
        Parameters $P_r, P_u, P_h, M_r$ , $\boldsymbol{\alpha}_{CU}$, $\boldsymbol{\alpha}_{DU}$.
output: Feasible Admission Setup $(\hat{\mathbf{x}}, \hat{\boldsymbol{\rho}})$
$(\hat{\mathbf{x}}^{\text{RU}}, \hat{\boldsymbol{\rho}}) \leftarrow \underset{\mathbf{x},\boldsymbol{\rho}}{\text{argmax}}\{a_{\text{SP1}(\mathbf{x})} : (10), (11), (15), (17)\}$
 $(\hat{\mathbf{x}}^{\text{DU}}, \hat{\mathbf{x}}^{\text{CU}}) \leftarrow \underset{\mathbf{x}}{\text{argmax}}\{a_{\text{SP2}(\mathbf{x})} : (9), (13), (14), (16), (20),$
 $(21)\}$
 return $\hat{x}$
```

B. Sequential model

Problem 1 could be further decomposed into two sub-problems that can be solved in different time scales. We define the first sub-problem in Problem 2. We consider a fixed group of UEs that we need to determine their optimal association with the RUs by (i) distributing Resource Blocks (RBs) of each RU across the three service types and (ii) determining the most suitable RU for each user, aiming to maximize the number of users successfully associated with an RU.

Problem 2 (Primary Sub-Problem)**.**

$$\underset{\mathbf{x},\boldsymbol{\rho}}{\text{maximize}} \quad a_{\mathit{SP1}}(\mathbf{x}) \triangleq \sum_{u \in \mathcal{U}} \epsilon_u \cdot x_{u,r}^{\text{RU}} \tag{18}$$

$$\text{subject to} \quad (10), (11), (15), (17)$$

We define the second optimization problem in Problem 3. We address the placement of CUs and DUs for the UEs, taking into account their respective RU associations determined in Problem 2, i.e, $\hat{\mathbf{x}}^{\text{RU}}$ and $\hat{\boldsymbol{\rho}}$. Problem 3's objective (19) is to maximize the number of admitted UEs (among the associated ones), by finding a valid CU-DU placement in terms of the remaining constraints.

Problem 3 (Secondary Sub-Problem)**.**

$$\underset{\mathbf{x}}{\text{maximize}} \quad a_{\mathit{SP2}}(\mathbf{x}) \triangleq \sum_{u \in \mathcal{U}} \epsilon_{s(u)} \cdot x_{u,h}^{\text{DU}} \cdot x_{u,h}^{\text{CU}} \tag{19}$$

$$\text{subject to} \quad \sum_{h,h' \in \mathcal{H}} x_{u,h}^{\text{DU}} \cdot x_{u,h'}^{\text{CU}} \leq \sum_{r \in \mathcal{R}} \hat{x}_{u,r}^{\text{RU}}, \forall u \in \mathcal{U} \tag{20}$$

$$\sum_{u \in \mathcal{U}} \hat{g}_u^{\text{CU}} \cdot x_{u,h}^{\text{CU}} + \hat{g}_u^{\text{DU}} \cdot x_{u,h}^{\text{DU}} \leq G_h, \forall h \in \mathcal{H} \tag{21}$$

$$(9), (13), (14), (16)$$

We remark that the coherence constraints (8) must be converted to inequality constraints (20), given that not all associated users will have a feasible CU-DU placement. Moreover, computational cost functions (3) are now constant values, i.e., $g_u^{\text{CU}}(\hat{\mathbf{x}}^{\text{RU}}) = \hat{g}_u^{\text{CU}}$ and $g_u^{\text{DU}}(\hat{\mathbf{x}}^{\text{RU}}) = \hat{g}_u^{\text{DU}}$. Therefore, we can replace original constraints (12) with new constraints (21). Constraints (9), (13), (14), and (16) remain the same. The final DU-CU placement resulting from solving Problem 3 is denoted by $\hat{\mathbf{x}}^{\text{DU}}$ and $\hat{\mathbf{x}}^{\text{CU}}$.

Finally, we propose to tackle Problem (1), by sequentially solving Problem 2 and Problem 3. The resulting solution is formalized in Algorithm 2.

Remark. *This strategic division of the problem aims to strike a balance between computational efficiency and solution optimality. In a real-world scenario where the system is constantly changing (for example, the locations of users), the way we break down the problem tends to favor a sequential solution for the complete admission problem (referred to as ”Problem 1”) in a time-efficient manner. Problem 2 can be solved within a short time frame and update the users association to RUs after each frame. Then, after a larger time interval, Problem 3 addresses O-Cloud placement decisions for users while taking into account the latest association decisions of the short time scale problem and so on.*

C. Linearization

The nonlinear objective function and constraints of our problem, e.g., equations (7), (8) and (12), include a product of two or more binary variables. This can be linearized using (and extending) the bilinear terms' linearization method [14]. Due to space constraints, we discuss in detail only the linearization of constraints (8), which has the product of $x_{u,h}^{DU}$ and $x_{u,h'}^{CU}$. The idea is to introduce a set of auxiliary binary variables that are virtually defined as

$$z_{uhh'} \triangleq x_{u,h}^{\text{DU}} \cdot x_{u,h'}^{\text{CU}}, \forall u \in \mathcal{U}, \forall h, h' \in \mathcal{H},$$

although, in practice, their values' coherence is enforced by imposing the following set of constraints

$$z_{uhh'} \leq (x_{u,h}^{\text{DU}} + x_{u,h'}^{\text{CU}})/2, \quad \forall h, h' \in \mathcal{H}, \forall u \in \mathcal{U}$$

$$z_{uhh'} \geq x_{u,h}^{\text{DU}} + x_{u,h'}^{\text{CU}} - 1, \quad \forall h, h' \in \mathcal{H}, \forall u \in \mathcal{U}.$$

The same technique can be applied to equations (7), (12), (13), (14), (19), and (20). Even though the linear version of Problem (1) has larger space complexity due to the additional variables and constraints, it can be solved using traditional integer programming techniques, such as Branch-and-Bound [15]. We emphasize that solving realistic instances of our proposed ILP model might entail substantial computational requirements and long solution time.

V. SIMULATION FRAMEWORK

We build our simulation setup based on the same network topology proposed in [8]. It consists of $|\mathcal{R}| = 4$ O-RUs, distributed across a squared area of side $L = 1$ km. The UEs are scattered within the defined area uniformly at random. The system employs a 20-MHz bandwidth, resulting in 100 RBs available per TTI at each O-RU. Additional radio parameters include four antennas, two MIMO layers, and 64-QAM modulation. The number of UEs varies from $|\mathcal{U}| = 10$ to 110, belonging to different slices including enhanced Mobile Broadband (eMBB), ultra-Reliable Low Latency Communication (uRLLC), and massive Machine Type Communication (mMTC), following the distribution in [12] for an industrial area where 25% of users are eMBB users, 25% are uRLLC users, and 50% are mMTC users. The eMBB and uRLLC UEs are given higher priority over mMTC UEs ensuring a balanced admission among users of different slice type. For the calculation of the required number of RBs per UE in eq. (1), the required data rate λ_s is set to 20 Mb/s, 5 Mb/s, and

1 Mb/s for eMBB, uRLLC, and mMTC UEs, respectively. The achievable data rate per RB is calculated using Shannon theory as in [9]. We consider a distance-dependent path-loss model with a transmission power of 30 dBm [9].

We consider $|\mathcal{H}_E| = 3$ edge-cloud nodes, such that the distance between any pair of edge-cloud nodes and O-RUs is between $[L, L'] = [5, 10]$ km. Moreover, we consider $|\mathcal{H}_R| = 1$ regional-cloud node randomly located within $[L'', L'''] = [40, 80]$ km away from the edge-cloud nodes. The O-Cloud setup is inline with the specification in [16]. The computational capacity G_h follows a uniform random distribution ranging from 100 to 200 GOPS for edge-cloud servers, and 1000 to 2000 GOPS for the regional-cloud nodes, which is in line with the findings in [11]. Regarding MH delay bounds D_u^{MH}, we consider values taken uniformly at random from the interval $[100, 300]$ µsec for uRLLC users, exactly 500 µsec for eMBB users, and 1000 µsec for mMTC users. The FH delay bounds D_u^{DU} are set to 100 µsec for all service types. These values are consistent with the considerations in [12]. Notably, our ILP-based problem is solved using IBM CPLEX software [17], a mathematical optimization solver, running on a computer equipped with an 11th generation Intel® Core™ i9-11950H Processor and 16 GB RAM.

VI. Performance Evaluation

In this section, we investigate the performance of the proposed models for different users densities. We base our simulation setup on the framework described in Section V. We refer to the solutions of the proposed joint and sequential models (Alg. 1 and 2), as *Optimal* and *Sequential*, respectively. We compare them with other baseline models:

- *Edge-Only Model*: Only the edge-cloud servers are available in this scenario. O-CUs and O-DUs are always deployed on the edge-cloud servers. The UEs are dynamically associated to the RUs following the description in section III-A.
- *CU-Regional Model*: O-CUs are all statically placed on regional servers, while O-DUs are exclusively deployed on edge servers. UEs are also dynamically associated to the RUs as explained in III-A.
- *Placement-Only Model*: This model was proposed in [8], in which O-CUs and O-DUs are placed across edge and regional clouds, but UEs are associated to the closest RU.

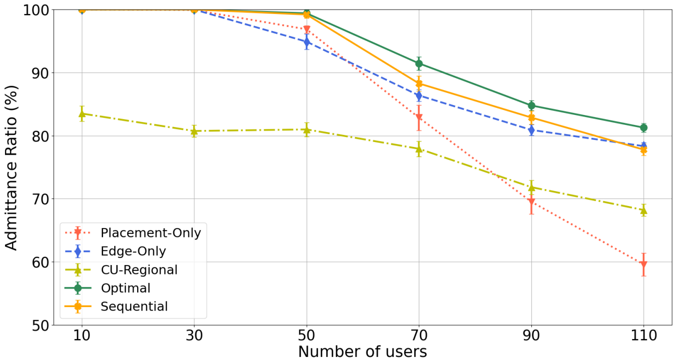

Figure 1: Average admittance ratio as a function of the number of users in the system

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

The paper employs the following performance metrics:

- Average Admittance Ratio: This metric calculates the average number of admitted users at each Transmission Time Interval (TTI) among all users in the network.
- Deployment Cost: This metric computes the average cost associated with deploying O-CUs in terms of running computational operations on a server (measured in GOPS). The cost of running functions on regional servers is lower than on edge servers, as regional servers offer greater processing capacity and consume less energy [6]. As reported in [12], 1 GOPS costs 1.59$ at an edge server whereas 0.5$ per GOPS at a regional cloud.
- Fairness Index: To assess the fairness of user admission across the three service types (eMBB, uRLLC, mMTC), Jain's fairness index is used. It is represented by $\zeta = (\sum_{j=1}^{N} AAR_j)^2/(N \cdot \sum_{j=1}^{N} AAR_j^2)$, where $N = 3$ is the number of distinct service types, and AAR_j is the average admittance ratio for users of service type j.

In Figure 1, we analyze the average admittance ratio versus the number of users for each of the considered models. As expected the *Optimal* and the *Sequential* solutions outperform the other models. A comparative analysis between the *Optimal* and *Sequential* models shows that dividing the problem into separate parts barely diminishes performance and moves it slightly further away from optimality, with the *Sequential* model exhibiting an admittance ratio order of 6% lower than that of the *Optimal* model. Primarily, the *Edge-only* model's solution shows a similar trend. However, it exhibits a slightly lower average admittance ratio. This decrease can be attributed to the limitation of computational resources within edge clouds, which makes it challenging to meet the diverse requirements of users, particularly those of the eMBB users who demand higher computational capacity. The *CU-Regional* model displays poorer average admittance ratios. This observation is referred to the fact that uRLLC users have stringent low-latency demands. Placing O-CUs in regional clouds introduces latency in the communication links. Lastly, the *Placement-Only* model also exhibits poor performance compared to the optimal approach. Overall, these results show the significance of our proposed models that encompasses UE to RU association and CU/DU placement, as they outperform scenarios solely relying on the dynamic placement of CUs and DUs. We remark that both *CU-Regional* and *Placement-Only* scenarios perform worse when the number of users is high, meaning that the static placement of the CU-Regional model diminishes the flexibility gained at the RU level, making it comparable in performance to the *Placement-Only* model. Figure 2 illustrates the O-CUs' deployment costs when adopting different scenarios. The *Edge-Only* model incurs the highest expenses due to the relatively higher cost of edge clouds compared to regional clouds. In contrast, the *Optimal* solution demonstrates lower deployment costs because CUs can now be strategically placed in regional clouds. An interesting observation is a peculiar cost reduction when the user count ranges between 50 and 70. This can be attributed to the model's behavior. When fewer users are present in the system, the model tends to favor edge clouds for CUs, as the edge resources are enough. However, as the number of users surpasses 50 UEs, competition for edge resources increases. Consequently, the *Optimal* model opts for regional clouds to deploy CUs while adhering to constraints, thereby freeing up edge cloud capacity for other users' DUs. Cost of the *Sequential* model: to be checked Additionally, the *Placement-Only* and *CU-Regional* models exhibit the lowest costs, primarily because they accommodate fewer users compared to the other scenarios.

Figure 3 evaluates the Jain's fairness index [18] of the

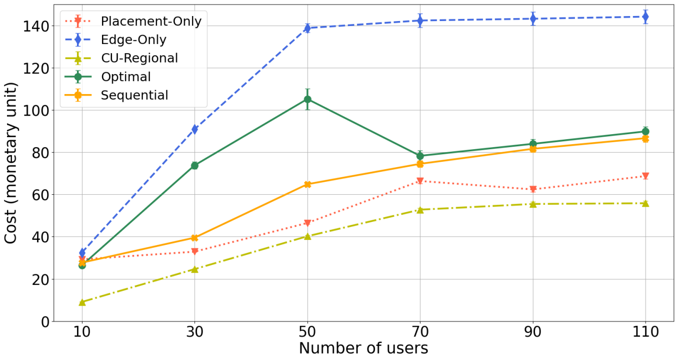

Figure 2: O-CU deployment costs for each scenario

admittance ratio among the three service types. Notably, the *Optimal* and *Sequential* models present a higher fairness index among users than the other baselines. On the other hand, the *CU-Regional* scenario shows the worst performance.

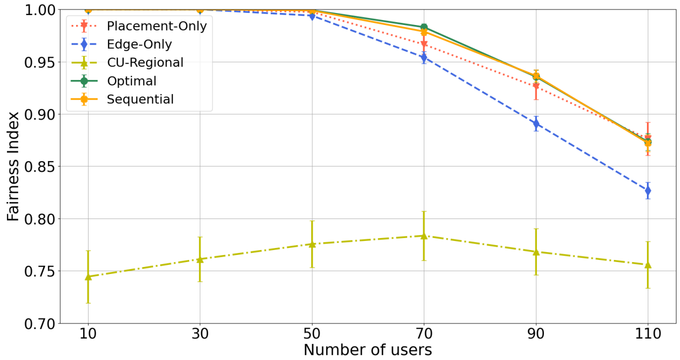

Figure 3: Fairness among all users as a function of the number of users

Finally, we evaluate the execution time of our two proposed models. Figure 4 reports the reduction in execution time of the *Sequential* model in comparison with the *Optimal* model. The execution time required by the former model is up to 80% less than that of the latter, offering a trade-off between performance and computational efficiency.

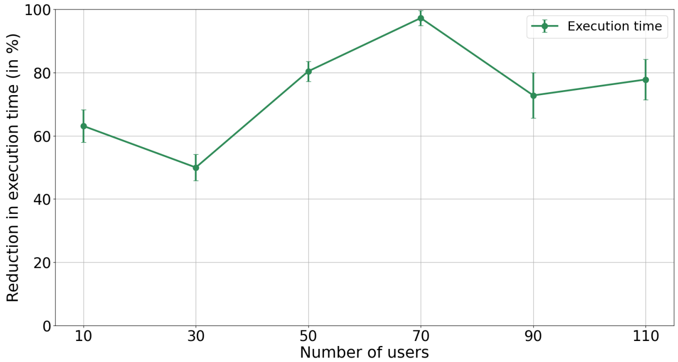

Figure 4: Reduction of execution time (in %) of the *Sequential* model compared to the *Optimal* one as a function of the number of users

VII. Conclusion

The transition to the Open RAN architecture signifies a transformative shift in access networks, characterized by increased openness, flexibility, and intelligence. In this paper, we address the challenge of optimizing the placement of O-CU and O-DU O-RAN components across edge and regional clouds while simultaneously considering users-to-RU associations. Our approach involves formulating two mathematical optimization models aimed at efficiently allocating available system resources, encompassing radio and computing resources, both jointly and sequentially. The primary objective was to maximize the number of users in the system while meeting their Quality of Service (QoS) requirements by efficiently utilizing these resources within the O-RAN framework. A comprehensive performance analysis of our models with respect to baselines from state-of-the-art shows an enhanced user admission and incentivized the offloading of O-RAN network functionalities to regional clouds, thereby reducing costs. Furthermore, we study the advantage of deploying a sequential optimization model instead of a joint one in terms of reduced execution time. As future work, we plan to leverage this decomposed optimization model to develop a two-time-scale solution, incorporating a temporal dimension for addressing the users association and functionalities placement problem dynamically.

References

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

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

[3] O-RAN Working Group 6, "O-ran cloud architecture and deployment scenarios for o-ran virtualized ran 4.0," O-RAN Alliance, Tech. Rep. O-RAN.WG6.CADS-v04.00, October 2022. [Online]. Available: https://orandownloadsweb.azurewebsites.net/specifications

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

[5] 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.

[6] 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 Network and Service Mang.*, 2022.

[7] M. Klinkowski, "Latency-aware du/cu placement in convergent packet-based 5g fronthaul transport networks," *Applied Sciences*, vol. 10, no. 21, 2020.

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

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

[10] "O-ran cloud architecture and deployment scenarios for o-ran virtualized ran 2.02," O-RAN Alliance, Tech. Rep., Feb 2021. [Online]. Available: https://orandownloadsweb.azurewebsites.net/specifications

[11] 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.

[12] 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 Management*, vol. 19, no. 3, 2022.

[13] A. Filali, Z. Mlika, S. Cherkaoui, and A. Kobbane, "Dynamic sdn-based radio access network slicing with deep reinforcement learning for urllc and embb services," *IEEE Trans. on Network Science and Eng.*, 2022.

[14] R. Fortet, "Applications de l'algèbre de boole en recherche opérationnelle," *Revue Française d'Automatique, d'Informatique et de Recherche Opérationnelle*, vol. 4, pp. 5–36, 1959.

[15] E. L. Lawler and D. E. Wood, “Branch-and-bound methods: A survey,” *Operations research*, vol. 14, no. 4, pp. 699–719, 1966.

[16] 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.

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

[18] R. Jain, D. Chiu, and W. Hawe, *A Quantitative Measure of Fairness and Discrimination for Resource Allocation in Shared Computer Systems*. DEC Research Report TR-301, Sep 1984.