Flexible Association and Placement for Open-RAN

Hiba Hojeij, Guilherme Iecker, Mahdi Sharara, Sahar Hoteit, Véronique Vèque, Stefano Secci

Flexible Association and Placement for Open-RAN

Hiba Hojeij, Guilherme Iecker Ricardo†, Mahdi Sharara‡, Sahar Hoteit, Véronique Vèque, Stefano Secci§
Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire des Signaux et Systèmes (L2S), 91190, Gif-sur-Yvette, France
†IRIT/Toulouse INP, University of Toulouse, France, §CEDRIC, CNAM, France
‡Orange, Innovation Division, Paris, France
Emails: firstname.lastname@centralesupelec.fr, guilherme.ricardo@irit.fr, mahdi.sharara@orange.com, stefano.secci@cnam.fr

***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 cloud-enabled RAN infrastructure in order to achieve users' required Quality of Service (QoS). Within this framework, our study is dedicated to maximizing the admission of User Equipments (UEs) into the system while ensuring their specific QoS needs. We focus on two primary tasks: (i) establishing an association between UEs and RUs and (ii) placing CUs and DUs across the network's cloud hosts. We initially address these tasks by formulating the joint association-placement optimization problem, subject to the system's available resources and QoS-related constraints. Although it is an NP-Hard problem, we discuss how we can frame it into an Integer Linear Programming (ILP) model. Then, we propose an approximation algorithm based on the decomposition of the original ILP model. We show through exhaustive simulations that our proposed ILP model provides higher admissibility levels than other baseline models. Moreover, it significantly minimizes the deployment costs and increases the overall fairness. Finally, we remark that our decomposition algorithm presents a short optimality gap in practice, with up to 6% less admissions, while reducing the solution time by up to 98%.**

***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 5G cellular network environment, 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 of open interfaces, 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* for gNB disaggregation: the 3GPP's radio protocol stack in classic gNBs is separately implemented in three different functional units: (i) Open Radio Units (O-RU), implementing low-PHY protocols; (ii) Open Distributed Units (O-DU), handling high-PHY and MAC tasks; and (iii) Open Centralized Units (O-CU) in charge of the upper layers. Additionally, as introduced in [3], O-RAN architectures may integrate cloud systems (the O-Cloud, in the official O-RAN nomenclature) in order to run virtualized instances of each functional unit. These features offer enhanced capabilities and flexible provision of communication services, placing O-RAN architectures as the succeeding iteration of Virtualized Radio Access Networks (vRANs) [1]. The use of the O-Cloud leverages different strategies for the deployment of the disaggregated O-RAN 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 disaggregated functional units to be deployed at O-Cloud hosts located at different proximity levels concerning the cell site, where the RUs are placed. By distributing these functional units across different nodes, network operators gain the flexibility to adapt to specific network demands and to meet minimum Quality of Service (QoS) requirements, such as high throughput and low latency. For instance, they can strategically install functional units closer to end-users, at the cell sites, to reduce end-to-end latency for applications with stringent delay requirements. Alternatively, they may choose to deploy these components in further cloud hosts, where more computing resources might be available to handle larger volumes of transiting packets. Thus, if, on the one hand, the disaggregation and distribution of functional units throughout the O-Cloud network provide flexibility, on the other hand, its deployment must be carefully designed to satisfy challenging constraints.

In this paper, we investigate how to maximize the number of UEs admitted to the system that have their minimum QoS requirements satisfied. To this end, we propose an optimization model that determines the admission decision at every transmission time interval (TTI) considering (i) *CU-DU placement*, which defines end-to-end, user-centric traffic forwarding throughout the O-Cloud network and (ii) *resource allocation*, specifically assigning sufficient resource blocks (RBs) to different users. We initially solve this problem jointly, then we propose a lightweight decomposition solution, where the master problem is split into two simpler sub-problems that, in turn, are solved sequentially. We evaluate our proposed model, and we show its superiority over other baseline models with limited RAN control options. Moreover, we discuss the performance of the proposed decomposition heuristic in terms of computational complexity and execution time.

The rest of the paper is organized as follows: Section II provides a brief overview of the related work. The system model and our proposed ILP-based solutions are described in Section III and IV, respectively. Section V details the simulation framework. Section VI illustrates the performance evaluation of the proposed algorithms. 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 DU and CU Virtualized/Cloud-native network functions (VNFs/CNFs) and establish optimal user equipment (UE) to 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).

Furthermore, network function placement shares similarities with task distribution in Multi-access Edge Computing (MEC) systems. The authors in [8] reinterpret the problem of maximizing the total throughput, considering task placement and routing under latency constraints as a betweenness-based flow assignment problem, and in [9], they propose a streaming algorithm to approximate the optimal solution with optimality guarantees under specific assumptions.

Additionally, our prior work in [10] 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 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 based on hosts' proximity to the cell site 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 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 [11]. 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 [10], a hybrid deployment scenario between the scenarios B and C that are defined in [12], 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 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 $\mathbf{x}^{\text{CU}} = [x^{\text{CU}}_{u,h} : \forall u \in \mathcal{U}, \forall h \in \mathcal{H}]$. We denote by $\mathbf{x} = [\mathbf{x}^{\text{RU}}, \mathbf{x}^{\text{DU}}, \mathbf{x}^{\text{CU}}]$ the vector of association-placement variables.

[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.

[2]We consider the "Shared-RU" framework introduced in [12, 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.

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., DU and CU, has an associated computational cost [13], given in *Giga Operations Per Second* (GOPS), that is defined as follows

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

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 DUs and CUs, respectively, with respect to their total computational requirements. Specifically, in our system, we adopt the Split-7.2x between RU and DU and the Split-2 between DU and CU, and based on the computational load distribution described in [14], we assign $\alpha_{DU} = 50\%$ and $\alpha_{CU} = 10\%$ of the computational workload to the DU and the CU, respectively (the RU is in charge of the remaining 40%). We formalize the total computational utilization in node h as

$$g_h(\mathbf{x}^{\text{RU}}) \triangleq \sum_{u \in \mathcal{U}} g_u^{\text{CU}}(\mathbf{x}^{\text{RU}}) \cdot x_{u,h}^{\text{CU}} + g_u^{\text{DU}}(\mathbf{x}^{\text{RU}}) \cdot x_{u,h}^{\text{DU}}, \quad (4)$$

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

D. E2E Delay Model

Assuming deterministic delay-bound forwarding is in place in the backhauling segment as commonly done starting from 3G (e.g. using carrier Ethernet, MPLS-TE or DETNET technologies), we consider that the variations in E2E delay experienced by a given UE are primarily affected by the propagation delay between communicating functional units, which has two major components: Midhaul (MH) and Fronthaul (FH) delay. For each UE u, the MH delay is measured between the CU to the DU, is given by:

$$d_u^{\text{MH}}(\mathbf{x}) \triangleq \sum_{h,h' \in \mathcal{H}} \frac{||P_h - P_{h'}||}{v_{\text{Fiber}}} \cdot x_{u,h}^{\text{CU}} \cdot x_{u,h'}^{\text{DU}}, \quad (5)$$

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

$$d_u^{\text{FH}}(\mathbf{x}) \triangleq \sum_{r \in \mathcal{R}} \sum_{h \in \mathcal{H}} \frac{||P_r - P_h||}{v_{\text{Fiber}}} \cdot x_{u,h}^{\text{DU}} \cdot x_{u,r}^{\text{RU}}. \quad (6)$$

IV. Problem Definition

Our goal is to optimize the system's performance by maximizing the admittance of UEs that is conditioned to the system's capability to satisfy their requested services' requirements. We formulate the full association-placement joint optimization problem as follows in Problem 1. Table I summarizes the notations used throughout the paper.

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 bounds [12], 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 regional domains. This limitation is captured by (9).

With (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_u^{MH} and D_u^{FH}, respectively. We refer to the (optimal) solution of Problem 1 as $\mathbf{x}^*$ and $\boldsymbol{\rho}^*$ illustrated in figure 1.

TABLE I: System notations

Notations	*Definition*
$\mathcal{R}, \mathcal{U}, \mathcal{H}, \mathcal{S}$	Sets of RUs, UEs, Hosts, and Slices, respectively
M_r	Parameter for the number of RBs available in RU r
a_u	Binary variable indicating UE u admittance
$x_{u,r}^{RU}$	Binary variable indicating UE u association to RU r
$x_{u,h}^{CU}$	Binary variable indicating if node h hosts UE u's CU
$x_{u,h}^{DU}$	Binary variable indicating if node h hosts UE u's DU
$\rho_{r,s}$	Integer variable for the RBs in RU r allocated to slice s
$\eta_{u,r}$	Transmission rate per PRB defined by Shannon's theorem
λ_s	Required data rate by slice s
d_u^{FH}	Maximum allowed fronthaul delay for user u
d_u^{MH}	Maximum allowed midhaul delay for user u
ϵ_s	Priority value for slice s

Problem 1 (Full joint Problem)**.**

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

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

$$\sum_{h \in \mathcal{H}_R} x_{u,h}^{\text{DU}} = 0, \forall u \in \mathcal{U} \quad (9)$$

$$\sum_{s \in \mathcal{S}} \rho_{r,s} \leq M_r, \forall r \in \mathcal{R} \quad (10)$$

$$RB_u(\mathbf{x}^{\text{RU}}) \leq \rho_{r,s}, \forall r \in \mathcal{R}, \forall s \in \mathcal{S} \quad (11)$$

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

$$d_u^{MH}(\mathbf{x}) \leq D_u^{MH}, \forall u \in \mathcal{U} \quad (13)$$

$$d_u^{FH}(\mathbf{x}) \leq D_u^{FH}, \forall u \in \mathcal{U} \quad (14)$$

$$x_{u,r}^{RU} \in \{0,1\}, \forall u \in \mathcal{U}, \forall r \in \mathcal{R} \quad (15)$$

$$x_{u,h}^{\text{CU}}, x_{u,h}^{DU} \in \{0,1\}, \forall u \in \mathcal{U}, \forall h \in \mathcal{H} \quad (16)$$

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

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

The intuition behind Proposition IV.1 is that a reduced version of Problem 1 has knapsack-like constraints [15]. Then, it is at least as hard as Knapsack problems, which is proven to be NP-Hard. Therefore, Problem 1 is NP-Hard as well.

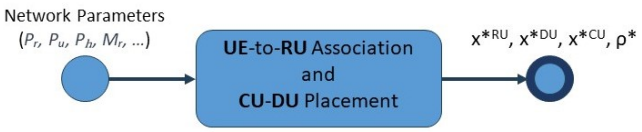

Figure 1: Full joint solution

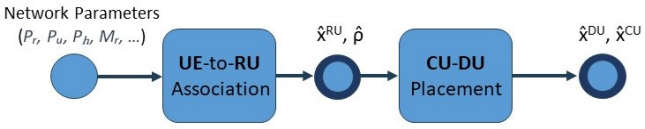

Figure 2: Decomposed sequential solution

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

A. Two-stage Decomposition

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 different 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)**.**

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

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

We define the second optimization sub-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)**.**

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

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

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

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

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., $g^{\text{CU}}_u(\hat{\mathbf{x}}^{\text{RU}}) = \hat{g}^{\text{CU}}_u$ and $g^{\text{DU}}_u(\hat{\mathbf{x}}^{\text{RU}}) = \hat{g}^{\text{DU}}_u$. 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 illustrated in figure 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.*

B. 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 [16]. Due to space constraints, we discuss in detail only the linearization of constraints (8), which has the product of $x^{DU}_{u,h}$ and $x^{CU}_{u,h'}$. The idea is to introduce a set of auxiliary binary variables that are virtually defined as

$$z_{uhh'} \triangleq x^{\text{DU}}_{u,h} \cdot x^{\text{CU}}_{u,h'}, \forall u \in \mathcal{U}, \forall h, h' \in \mathcal{H}, \tag{22}$$

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

$$z_{uhh'} \leq (x^{\text{DU}}_{u,h} + x^{\text{CU}}_{u,h'})/2, \quad \forall h, h' \in \mathcal{H}, \forall u \in \mathcal{U} \tag{23}$$

$$z_{uhh'} \geq x^{\text{DU}}_{u,h} + x^{\text{CU}}_{u,h'} - 1, \quad \forall h, h' \in \mathcal{H}, \forall u \in \mathcal{U}. \tag{24}$$

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 [17]. 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 [10]. It consists of $|\mathcal{R}| = 4$ 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 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, which in turn varies the load in the system from underloaded to overloaded systembased on the selected number of O-RUs. Users belong to different slices, including enhanced Mobile Broadband (eMBB), ultra-Reliable Low Latency Communication (uRLLC), and massive Machine Type Communication (mMTC), following the distribution in [14] 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 [11]. We consider a distance-dependent path-loss model with a transmission power of 30 dBm [11].

We consider $|\mathcal{H}_E| = 3$ edge-cloud nodes, such that the distance between any pair of edge-cloud nodes and 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 [18]. 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 [13]. 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 [14]. Notably, our ILP-based problem is solved using IBM CPLEX software [19], a mathematical optimization solver, running on a computer equipped with an 11th generation Intel® Core™ i9-11950H Processor and 16 GB RAM.

VI. Numerical 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 proposed joint and sequential solutions 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. CUs and 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*: CUs are all statically placed on regional servers, while DUs are exclusively deployed on edge servers. UEs are also dynamically associated to the RUs as explained in Section III-A.
- *Placement-Only Model*: This model was proposed in [10], in which CUs and DUs are dynamically placed across edge and regional clouds, but UEs are associated to the closest RU.

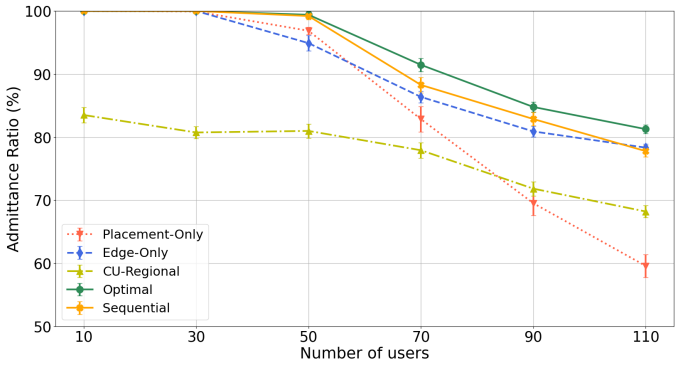

Figure 3: Average admittance rate as a function of the nb. of users.

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%.

In our experiment, we employ the following performance metrics:

- Average Admittance Ratio: it calculates the average number of admitted users at each Transmission Time Interval (TTI) among all users in the network.
- Deployment Cost: it computes the average cost associated with deploying 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 [14], 1 GOPS costs 1.59$ at an edge server whereas 0.5$ per GOPS at a regional cloud.
- Fairness Index: it assesses the fairness of user admission across the three service types (eMBB, uRLLC, mMTC), Jain's fairness index is used [20]. 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 3, 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* behavior 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. Secondly, 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 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 highlight the reach that our approach has, as it encompasses UE to RU association and CU/DU placement; our models 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 4 illustrates the 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* and *Sequential* solutions demonstrate lower deployment costs because CUs can now be strategically placed in regional clouds. Additionally, the *Placement-Only* and *CU-Regional* models exhibit the lowest costs, primarily because they accommodate fewer users compared to the other scenarios. Figure

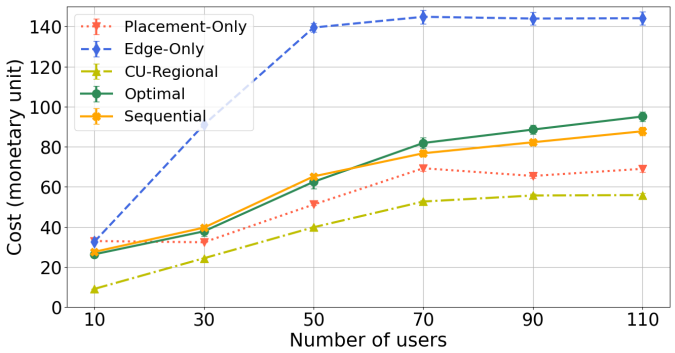

Figure 4: CU deployment costs for each scenario

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

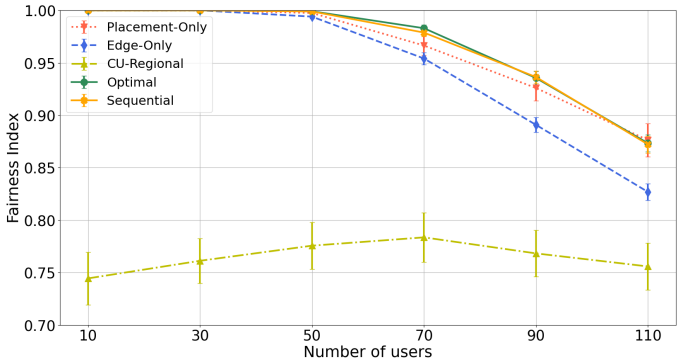

Figure 5: User fairness as a function of the number of users.

Finally, we evaluate the execution time of our two proposed solutions. Figure 6 reports the reduction in execution time of the *Sequential* solution in comparison with the *Optimal* one. The execution time required by the former is reduced by up to 98% compared to that of the latter. Overall, our simulation results demonstrate that both *Optimal* and *Sequential* solutions outperform static CU-DU placement and static UE to RU association in terms of user admission and computational deployment cost. Nevertheless, our two proposed solutions showcase a trade-off between user admission and computational efficiency, as reflected in the reduction of execution time analysis.

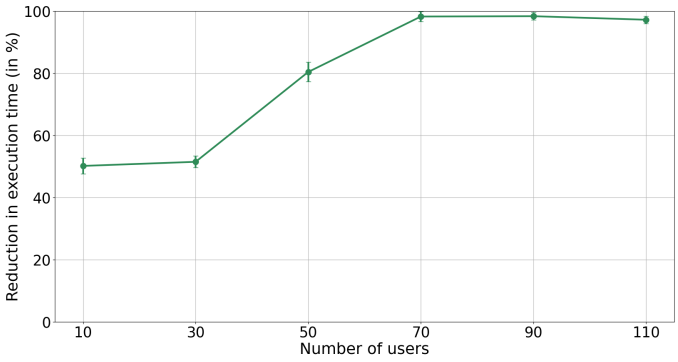

Figure 6: Reduction of execution time (in %) of the *Sequential* model compared to the *Optimal* one as a function of the number of users

VII. Conclusion

The transition to the Open RAN architecture leads to a transformative architectural shift in access networks, characterized by increased openness, flexibility, and intelligence. In this paper, we address the challenge of optimizing the placement of CU and 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 QoS requirements by efficiently utilizing these resources within the O-RAN framework. A comprehensive performance analysis of our approach with respect to baselines from state-of-the-art shows an enhanced user admission and that it can ease offloading O-RAN network functions 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 user association and functionality placement problem in a mobile scenario.

VIII. Acknoledgement

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

References

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

[2] M. Polese, L. Bonati, S. D'Oro, S. Basagni, and T. Melodia, "Understanding o-ran: Architecture, interfaces, algorithms, security, and research challenges," *IEEE 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, 2020.

[8] G. Iecker Ricardo, A. Benhamiche, N. Perrot, and Y. Carlinet, "Latency-constrained task distribution in multi-access edge computing systems," in *2022 IEEE 11th Intern. Conf. on Cloud Networking (CloudNet)*, 2022.

[9] G. I. Ricardo, A. Benhamiche, N. Perrot, and Y. Carlinet, "Heuristic distribution of latency-sensitive tasks in multi-access edge computing systems," in *2022 IEEE Globecom Workshops (GC Wkshps)*, 2022.

[10] 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.

[11] 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.

[12] "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

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

[14] 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.

[15] 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.

[16] 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.

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

[18] 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.

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

[20] 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.