AUTHOR QUERIES

AUTHOR PLEASE ANSWER ALL QUERIES

PLEASE NOTE: We cannot accept new source files as corrections for your paper. If possible, please annotate the PDF proof we have sent you with your corrections and upload it via the Author Gateway. Alternatively, you may send us your corrections in list format. You may also upload revised graphics via the Author Gateway.

Carefully check the page proofs (and coordinate with all authors); additional changes or updates WILL NOT be accepted after the article is published online/print in its final form. Please check author names and affiliations, funding, as well as the overall article for any errors prior to sending in your author proof corrections. Your article has been peer reviewed, accepted as final, and sent in to IEEE. No text changes have been made to the main part of the article as dictated by the editorial level of service for your publication.

AQ1: According to our records, Vahid Shah-Mansouri and Onel Luis Alcaraz López are listed as a Member, IEEE. Please verify. Yes, I verified for both.

AQ2: According to our records, Saeede Parsaeefard is listed as a Senior Member, IEEE. Please verify. Yes, I verified, by the way the correct name is "Saeedeh Parsaeefard", please correct it.

AQ3: Please confirm or add details for any funding or financial support for the research of this article. I confirmed that we do not have any funding and we use self-fund of authors.

AQ4: Please provide the department name for University of Oulu. Centre for wireless communications (CWC)

AQ5: Note that if you require corrections/changes to tables or figures, you must supply the revised files, as these items are not edited for you. I do not need to correct anything

AQ6: Please provide the issue number for Reference [1]. ISSN: 2644-1330

AQ7: Please provide the report number for Reference [9]. Since the other versions are removed from the O-RAN Alliance website, I converted it to the version 6 (v06.00)

```
@TechReport{ORANArch,
title = {{O-RAN} Architecture Description v06.00},
author = {{O-RAN} Working Group 1},
institution = {{O-RAN}{Alliance}},
month ={Mar.},
year = {2022}
}
```

Additional Requirements:
We want to add the biography to this paper and I attached it in the biography folder

Resource Allocation in an Open RAN System Using Network Slicing

AQ1 I verified it
AQ2 I verified it

Mojdeh Karbalaee Motalleb, Vahid Shah-Mansouri, *Member, IEEE*, Saeede Parsaeefard, *Senior Member, IEEE*, and Onel Luis Alcaraz López, *Member, IEEE*

please correct the name to "Saeedeh"

***Abstract*—The next radio access network (RAN) generation, open RAN (O-RAN), aims to enable more flexibility and openness, including efficient service slicing, and to lower the operational costs in 5G and beyond wireless networks. Nevertheless, strictly satisfying quality-of-service requirements while establishing priorities and promoting balance between the significantly heterogeneous services remains a key research problem. In this paper, we use network slicing to study the service-aware baseband resource allocation and virtual network function (VNF) activation in O-RAN systems. The limited fronthaul capacity and end-to-end delay constraints are simultaneously considered. Optimizing baseband resources includes O-RAN radio unit (O-RU), physical resource block (PRB) assignment, and power allocation. The main problem is a mixed-integer non-linear programming problem that is non-trivial to solve. Consequently, we break it down into two different steps and propose an iterative algorithm that finds a near-optimal solution. In the first step, we reformulate and simplify the problem to find the power allocation, PRB assignment, and the number of VNFs. In the second step, the O-RU association is resolved. The proposed method is validated via simulations, which achieve a higher data rate and lower end-to-end delay than existing methods.**

***Index Terms*—Open radio access network (O-RAN), virtual network function (VNF), network slicing, knapsack problem, greedy algorithm, Karush-Kuhn-Tucker (KKT) conditions.**

I. Introduction

NETWORK slicing is a key technology in 5G wireless systems. Specifically, it isolates network resources into slices, e.g., via core slicing and/or radio access network (RAN) slicing, for serving various services [1], [2], [3].

There are three main service classes in 5G, namely enhanced mobile broadband (eMBB), ultra-reliable low latency communications (URLLC), and massive machine-to-machine communications (mMTC). Each service is assigned to a network slice depending on its corresponding quality of service (QoS) requirements. For instance, the eMBB service demands high capacity and throughput, e.g., 8K video streaming and immersive gaming. Meanwhile, the URLLC service provides ultra-reliable and low-latency connectivity, e.g., for autonomous vehicles, Tactile Internet, and remote surgeries. Finally, mMTC services require connectivity for a large number of Internet of Things (IoT) devices that transmit small payloads [4], [5], [6].

A. Motivation

The optimal resource allocation in 5G systems is crucial for reducing costs and improving the performance experienced by the user equipments (UEs). These systems face significant challenges, including interference alignment, limited capacity of the fronthaul links, energy restrictions on virtual machines (VMs), etc. [2], [7], [8].

Many studies have investigated resource allocation in cloud RAN (C-RAN) by considering a single service's power, data rate, and delay limitations. Unfortunately, the existing radio access networks (RANs) currently lack adequate flexibility and openness to handle these simultaneous service demands. Hence, a new RAN paradigm, called open RAN (O-RAN) architecture, has emerged. Therefore, O-RAN can simultaneously support multiple services at a lower cost by being flexible, layered, and modular. One of the fundamental problems lies in balancing services with different QoS, resource requirements, and priorities in O-RAN architecture [1], [9], [10], [11].

The purpose of this paper is to design a system in the O-RAN architecture to support the three types of 5G services, namely, eMBB, URLLC, and mMTC via network slicing and resource allocation.

B. Main Contributions

This paper studies the resource utilization of a downlink O-RAN system to develop an isolated network slicing outline for the three 5G services. We use mathematical methods to decompose and convexify the problem and solve it using hierarchical algorithms. The main contributions of this paper are summarized as follows:

- We examine the problem of baseband resource allocation, such as power, physical resource blocks (PRBs), O-RUs, and activating VNFs, to maximize the weighted throughput of the O-RAN architecture. The three types of 5G service classes, i.e., eMBB, URLLC, and mMTC,

Manuscript received 24 March 2022; revised 13 July 2022 and 5 September 2022; accepted 5 September 2022. The associate editor coordinating the review of this article and approving it for publication was J.-F. Botero. *(Corresponding author: Mojdeh Karbalaee Motalleb.)*

AQ3

Mojdeh Karbalaee Motalleb and Vahid Shah-Mansouri are with the School of ECE, University of Tehran, Tehran 1439957131, Iran (e-mail: mojdeh.karbalaee@ut.ac.ir; vmansouri@ut.ac.ir@gmail.com).

Saeede Parsaeefard is with the Department of Electrical Engineering, University of Toronto, Toronto, ON M5S 3G4, Canada (e-mail: saeede.parsaeefard@gmail.com).

Onel Luis Alcaraz López is with the University of Oulu, 90570 Oulu, Finland (e-mail: onel.alcarazlopez@oulu.fi).

AQ4

Digital Object Identifier 10.1109/TNSM.2022.3205415

Centre for wireless communication (CWC)

together with their corresponding QoS requirements and service priorities, are considered.

- We propose a two-step resource management algorithm for solving the optimization problem. In the first step, we reformulate and simplify the problem so as to find an upper and lower bound for the number of activated VNFs. Moreover, we use the Lagrangian function and Karush-Kuhn-Tucker (KKT) conditions to obtain the optimal power and PRB allocation. In the second step, the problem of O-RU association is converted to a multiple knapsack problem and solved by a greedy algorithm.
- We analyze the complexity of the proposed algorithms and demonstrate their convergence. Additionally, we analyze the feasibility region of the problem and introduce a fast algorithm to check it numerically.
- We show via numerical results that the proposed algorithm outperforms two baseline schemes in terms of achievable data rate and mean total delay. Remarkably, the proposed algorithm performs close to the optimal solution in low-interference conditions.

C. Organization

This paper is organized as follows. Relevant literature related to our work is discussed in Section II, while Section III briefly overviews the O-RAN architecture. The system model and the problem formulation are described in Section IV and Section V, respectively. The details of our proposed resource management algorithm are introduced in Section VI. In Section VII, numerical results are provided to evaluate the performance of the proposed algorithm. Finally, Section VIII concludes the paper. For clarity, Table I lists the main acronyms used throughout the paper.

TABLE I
LIST OF ACRONYMS

no need to change AQ5

Acronym	Definition
VNF	virtual network function
VM	virtual machine
RAN	radio access network
O-RAN	open RAN
vRAN	virtual RAN
CRAN	cloud RAN
RRH	radio remote head
BBU	baseband unit
QoS	quality of service
MIMO	multiple input multiple output
PRB	physical resource block
eMBB	enhanced mobile broadBand
URLLC	ultra-reliable low latency communication
mMTC	massive machine-to-machine communications
O-RU	O-RAN radio unit
O-DU	O-RAN distributed unit
O-CU	O-RAN central unit
UPF	user plane function
UE	user equipment
SINR	signal-to-noise-plus-interference ratio
CAPEX	capital expenditures
OPEX	operating expenses
KKT	Karush-Kuhn-Tucker

II. Related Literature

The network slicing problem in multi-tenant cellular networks has received significant attention recently, e.g., [7], [12], [13]. Two levels of dynamic network slicing in heterogeneous C-RAN (H-CRAN) are examined in [7]. The higher level manages user acceptance control, RRH association, and the allocation of BBU capacity. Meanwhile, PRB and power are allocated at lower levels. In [14], RAN slicing is considered for the fog RAN (F-RAN) system, and executed using deep reinforcement learning. In [15], [16], the implementation of RAN level slicing is discussed among multiple mobile network operators with the specific physical network resources infrastructure. In [17], to provide 5G slicing services, the authors present a framework called O-RANFed that implements and optimizes federated learning tasks in O-RAN devices. Moreover, The authors of [18] propose a federated deep reinforcement learning algorithm to achieve network slicing in O-RAN.

Recent research has focused on multiplexing eMBB and URLLC services within a RAN. In [2], the sum rate of the eMBB, and URLLC, is optimized by ensuring that each user's traffic load demand is met and the slice isolation is guaranteed, thus promoting an amicable service coexistence. In [19], a RAN slicing is considered in a coordinated multipoint system to guarantee the QoS requirements of eMBB and URLLC services. Moreover, [20] investigates the minimization of the system's power for the RAN slicing of eMBB and URLLC downlink services using non-orthogonal multiple access techniques. In [5], the problem of resource allocation for the coexistence of eMBB and URLLC services is formulated and solved by deep reinforcement learning.

In [8], [21], VMs activation and beamforming allocation are discussed in C-RAN systems. Paper [8] minimizes energy cost with system delay, fronthaul capacity, and rate constraints. To guarantee UE delays, M/M/1 queueing theory is used for transmission and processing delays. In [22], [23], the problem of joint virtual computing resource allocation with beamforming is formulated. Also, the association of RRH to the UE is considered and solved using innovative methods.

In [24], [25], the problem of joint power allocation and RRH association in a H-CRAN system is considered to maximize the energy efficiency. Finally, in [26], the optimum power is obtained in the massive MIMO aided C-RAN system, and the problem of RRH to BBU and RRH to UE association is formulated and solved.

III. Background

O-RAN is an appropriate alternative to the next generation of radio access networks due to its flexibility, openness, low operational costs, and intelligence.

O-RAN was developed to jointly benefit from the advantages of virtual RAN (vRAN) and cloud RAN (C-RAN). By virtualizing RANs, operators can improve flexibility, reduce

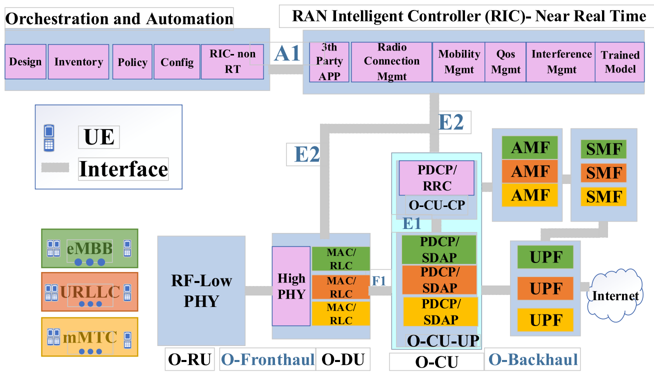

Fig. 1. Network sliced O-RAN system.

CAPEX and OPEX, and add new capabilities to their networks more quickly. The C-RAN architecture divides the RAN into two major parts: the radio remote head (RRH) and the baseband unit (BBU). Several distributed RRHs can be connected to a centralized BBU, called BBU-pool [27]. Unlike C-RAN, O-RAN separates RAN into three different units, namely Radio Unit (O-RU), Distributed Unit (O-DU), and Central Unit (O-CU). Mostly non-real-time baseband processing occurs in the O-CU layer, while real-time baseband processing occurs in the O-DU layer.

In the O-RAN architecture, the PHY is divided into low and high PHY, unlike C-RAN. As shown in Fig. 1, O-RU is a logical node that contains RF and low PHY. The former transmits or receives radio signals, while the latter includes digital beamforming. Typically, the O-DU constitutes a logical node with high PHY, MAC, and RLC. It contains a subfunction of the eNodeB and is deployed near the O-RU. Moreover, O-DU is connected to an O-RU with an open fronthaul interface. In addition to supporting the lower layers of the protocol stack, O-CU also provides support for the higher layers.

The O-CU contains two parts: the O-CU user plane (O-CU-UP) and the O-CU control plane (O-CU-CP). The former hosts the packet data convergence protocol (PDCP)-UP and the service data adaption protocol (SDAP), while the latter hosts PDCP-CP and radio resource control (RRC). O-DU and O-CU are connected via an open and well-defined interface F_1. Moreover, O-CU-UP is connected to user plane function (UPF) via O-backhaul link.

The O-RAN architecture contains other principal logical nodes called Orchestration and Automation, RAN Intelligent Controller (RIC)- Near Real-Time, and O-Cloud. Orchestration and Automation include functions such as RIC Non-Real-Time. RIC is responsible for machine learning methods and making the system more intelligent.

key feature of the O- RAN architecture is that the hardware is disaggregated from the software, leading to network function virtualization (NFV). Additionally, each component is implemented as a virtual network function (VNF), the system function block in NFV, that can be deployed on a virtual machine (VM) or container [28]. As a result, as shown in Fig. 1, O-RAN components, such as UPF, O-CU, O-DU, and RIC-near real-time, are virtualized and implemented as VNFs [9], [10], [11], [29], [30], [31], [32].

IV. System Model

We consider a downlink (DL) system, and an O-RAN architecture using RAN slicing as depicted in Fig. 1. In this section, we present the system and signal model, derive the achievable data rates, power of O-RU, and the fronthaul capacity of the O-RAN system. Moreover, we discuss the mean delay and the power of VNFs.

A. System Model

Assume, there are three service types: eMBB, URLLC, and mMTC, which support different applications. Accordingly, there are S_1 slices for the first service type (eMBB), S_2 slices for the second service type (URLLC), and S_3 slices for the third service type (mMTC). Therefore, there are $S = S_1 + S_2 + S_3$ pre-allocated slices serving these services. Moreover, each service request $s \in \{1, \ldots, S\}$ is served by its corresponding slice. So we have the set $\{1, 2, \ldots, S_1\}$ of eMBB service instances, the set $\{1, 2, \ldots, S_2\}$ of URLLC service instances, and the set $\{1, 2, \ldots, S_3\}$ of mMTC service instances. Each service $s_j \in \{1, 2, \ldots, S_j\}$ consists of U_{s_j} requests from single-antenna UEs requiring certain level of QoS. Notice that $j \in \{1, 2, 3\}$ indicates the service type. Based on the application and QoS request, UE may be admitted and allocated to the resources.

Each pre-allocated slice contains reserved VNFs for the three logical nodes:

- MAC/RLC functions in the O-DU
- PDCP/SDAP functions in the O-CU-UP
- UPF which is a functional layer

Each slice $s \in \{1, 2, \ldots, S\}$, consists of M_s^d VNFs for the processing of O-DU, M_s^c VNFs for the processing of O-CU-UP, and M_s^u VNFs for the processing of UPF. The VNFs of O-DU, O-CU-UP, and UPF are interconnected, which is defined as the service function chain in the O-RAN system. Also, each VNF instance runs on a VM that uses resources from the data centers.

Assume there are K PRBs in this system. Suppose each slice s consists of $\bar{K}_s$ pre-allocated virtual resource blocks that are mapped to PRBs. Therefore, we have $\sum_s \bar{K}_s \leq K$. In addition, there are R multi-antenna O-RUs that are shared between the slices. Specifically, the O-RU $r \in \mathcal{R} = \{1, 2, \ldots, R\}$ has J antennas for transmitting and receiving data. Moreover, all O-RUs have access to all PRBs.

B. Signal Model

Let $y_{u(s,i)}$ be the received signal of UE i in the s^{th} service such that

$$y_{u(s,i)} = \sum_{r=1}^{R} \sum_{k=1}^{K_s} \boldsymbol{h}^{H\,k}_{r,u(s,i)} g^{r}_{u(s,i)} e^{k}_{r,u(s,i)} x_{Q}{}^{k}_{r,u(s,i)} + z_{u(s,i)}, \tag{1}$$

where $x_{Q}{}^{k}_{r,u(s,i)} = x_{P}{}^{k}_{r,u(s,i)} + \boldsymbol{q}_r$, $x_{P}{}^{k}_{r,u(s,i)} = \boldsymbol{w}^{k}_{r,u(s,i)} \sqrt{p^{k}_{r,u(s,i)}}\, x_{u(s,i)}$, $x_{u(s,i)}$ depicts the transmitted symbol vector, $z_{u(s,i)} \sim \mathcal{CN}(0, BN_0)$ is the receive additive Gaussian noise, and BN_0 is the noise power in a given bandwidth B. Here, x_P denotes the precoded message before compression, while x_Q illustrates the precoded message after compression. In addition, $\boldsymbol{q}_r \sim \mathcal{CN}(0, {\sigma_q}^2 \boldsymbol{I_R})$ indicates the quantization Gaussian noise which comes from the signal compression in O-DU. Furthermore, $g^r_{u(s,i)} \in \{0,1\}$ is a binary variable that illustrates whether O-RU r serves the i^{th} UE that is allocated to the s^{th} slice or not. Furthermore, $p^k_{r,u(s,i)}$ represents the transmission power of the O-RU r serve the i^{th} UE in slice s and PRB k, while $\mathbf{h}^k_{r,u(s,i)} \in \mathbb{C}^J$ corresponding channel vector. In addition, $\mathbf{w}^k_{r,u(s,i)} \in \mathbb{C}^J$ depicts the associated transmit beamforming vector. Therefore, the SINR of the i^{th} UE served at slice s on PRB k is given by

$$\rho^k_{r,u(s,i)} = \frac{p^k_{r,u(s,i)} |\mathbf{h}^{k\,H}_{r,u(s,i)} \mathbf{w}^k_{r,u(s,i)}|^2}{BN_0 + I^k_{r,u(s,i)}}, \tag{2}$$

A UE in an O-RU r using PRB k receives interference from other O-RUs in the set $\mathcal{R} \backslash r$ that are using the same PRB k. Two types of interference occur between UEs in each slice: i) inter-slice interference between signals transmitted over different slices, and ii) intra-slice interference between signals transmitted over the same slice.

Network slicing techniques significantly reduce inter-service (inter-slice) interference. One way to leverage a two-time scale PRB scheduling is to isolate PRBs in slices (in the first time scale) and schedule the PRBs to the UEs of the slices (in the second time scale). Another method consists of allocating part of the PRBs of eMBB services to URLLC and mMTC [2], [5], [33]. In this paper, we assume that the PRB scheduling is performed. Also, in Section V-A, we briefly study the PRB scheduling between slices. Since there are limited resources, inter-service interference cannot be eliminated entirely. Nevertheless, isolating the slices reduces inter-service interference considerably and allows us to ignore it mathematically.

Back to (2), $I^k_{r,u(s,i)}$ is the sum of the power of interfering signals and quantization noise, and can be represented as

$$I^k_{r,u(s,i)} = \underbrace{\sum_{j=1}^{R} {\sigma_q}^2 |\boldsymbol{h}^k_{r,u(s,i)}|^2}_{\text{(quantization noise)}} + \underbrace{\sum_{\substack{l=1\\ l\neq i}}^{U_s} e^k_{u(s,i)} e^k_{u(s,l)} p^k_{u(s,l)} \sum_{\substack{r'=1\\ r'\neq r}}^{R} |\mathbf{h}^{k\,H}_{r',u(s,i)} \mathbf{w}^k_{r',u(s,l)} g^{r'}_{u(s,l)}|^2}_{\text{(intra-slice interference)}}, \tag{3}$$

where $e^k_{u(s,i)}$ is a binary variable that indicates whether the k^{th} PRB is allocated to the UE i in slice s, assigned to r^{th} O-RU, or not. Furthermore, there is no inter-slice interference, only intra-slice interference, since slices are assumed to be isolated.

Herein, we consider a zero forcing beamforming vector, which minimizes the experienced intra-slice interference, and is given by [34]

$$\mathbf{w}^k_{r,u(s,i)} = \hat{\mathbf{h}}^k_{r,u(s,i)} (\hat{\mathbf{h}}^{k\,H}_{r,u(s,i)} \hat{\mathbf{h}}^k_{r,u(s,i)})^{-1}. \tag{4}$$

where $\mathbf{h}^k_{r,u(s,i)}$ is the channel estimate, which is assumed imperfect. Mathematically, $\hat{\boldsymbol{h}}_{r,u(s,i)} = \boldsymbol{h}_{r,u(s,i)} + \Delta \boldsymbol{h}_{r,u(s,i)}$, where $\Delta \boldsymbol{h}_{r,u(s,i)} \sim \mathcal{N}(0, \boldsymbol{\phi}^2_{r,u(s,i)})$ indicates the estimating error vector with a Gaussian distribution and $\boldsymbol{\phi}_{r,u(s,i)} = \operatorname{diag}(\phi_{r,u(s,i)}, \ldots, \phi_{r,u(s,i)})$.

C. Achievable Data Rate

The achievable data rate for the i^{th} UE request in the s_1^{th} application of service type 1 (eMBB) can be written as

$$\mathcal{R}_{u(s_1,i)} = \sum_{r=1}^{R} \mathcal{R}_{r,u(s_1,i)} g^r_{u(s_1,i)}, \tag{5}$$

where

$$\mathcal{R}_{r,u(s_1,i)} = \sum_{k=1}^{K} \mathcal{R}^k_{r,u(s_1,i)} e^k_{r,u(s_1,i)} \tag{6}$$

is the achievable data rate of RU r to UE i in slice s_1, which depends on the achievable data rate per PRB, i.e.,

$$\mathcal{R}^k_{r,u(s_1,i)} = B \log_2(1 + \rho^k_{r,u(s_1,i)}), \tag{7}$$

Since the blocklength in URLLC and mMTC is finite, the achievable data rate for the i^{th} UE request in the application of service type 2 (URLLC) and 3 (mMTC) is not achieved from the Shannon capacity formula. Instead, in a short packet transmission, the achievable data rate is approximated as [2]

$$\mathcal{R}_{u(s_j,i)} = \sum_{r=1}^{R} \mathcal{R}^r_{u(s_j,i)} g^r_{u(s_j,i)}, \tag{8}$$

where

$$\mathcal{R}_{r,u(s_j,i)} = \mathcal{R}^k_{r,u(s_j,i)} e^k_{u(s_j,i)}, \tag{9}$$

is the achievable data rate of RU r to UE i in slice s_1, which depends on the achievable data rate per PRB, i.e.,

$$\mathcal{R}^k_{r,u(s_j,i)} = B(\log_2(1 + \rho^k_{r,u(s_j,i)}) - \zeta^k_{u(s_j,i)}) e^k_{u(s_j,i)}, \tag{10}$$

where

$$\zeta^k_{u(s_j,i)} = \log_2(e) Q^{-1}(\epsilon) \sqrt{\mathfrak{C}^k_{u(s_j,i)} / N^k_{u(s_j,i)}}. \tag{11}$$

Here, ϵ is the transmission error probability, Q^{-1} is the inverse of the Q function, $\mathfrak{C}^k_{u(s_j,i)} = 1 - \frac{1}{(1+\rho^k_{u(s_j,i)})^2}$ depicts the channel dispersion of UE i at slice s_j and PRB k, while $N^k_{u(s_j,i)}$ represents the corresponding transmit blocklength. $\mathcal{R}_{r,u(s_j,i)}$ is the achievable data rate that is transmitted by O-RU r to UE i requesting service s_j.

If we replace $p^k_{u(s,l)}$ and $p^k_{u(n,l)}$ in (3) by $P_s^{\max}$, an upper bound $\bar{I}^k_{r,u(s,i)}$ is obtained for $I^k_{r,u(s,i)}$. Therefore, $\bar{\mathcal{R}}_{u_{(s,i)}} \forall s, i$ is derived by using $\bar{I}^k_{r,u(s,i)}$ instead of $I^k_{r,u(s,i)}$ in (8) and (5).

D. Power of the O-RU and the Fronthaul Capacity

Let P_r denote the power of the transmitted signal from the r^{th} O-RU to all the UEs served by it. From (1), the power of each O-RU r is obtained as follows,

$$P_r = \sum_{s=1}^{S} \sum_{k=1}^{K_s} \sum_{i=1}^{U_s} |\mathbf{w}^k_{r,u(s,i)}|^2 \alpha^k_{r,u(s,i)} + \sigma_q^2, \tag{12}$$

where $\alpha^k_{r,u(s,i)} = p^k_{r,u(s,i)} g^r_{u(s,i)} e^k_{r,u(s,i)}$. Since we have a fiber link between O-RU and O-DU, the rate of users on the fronthual link between O-DU and the r^{th} O-RU is formulated as

$$C_r = \log_2\left(1 + \sum_{s=1}^{S}\sum_{k=1}^{K_s}\sum_{i=1}^{U_s} |\mathbf{w}^k_{r,u(s,i)}|^2 \alpha^k_{r,u(s,i)}/\sigma_q^2\right)$$
$$= \log_2\left(P_r/\sigma_q^2\right). \tag{13}$$

E. Mean Delay

In this part, the end-to-end mean delay for each service is obtained. The total delay (T^{tot}), is the sum of the processing delay (T^{proc}), the transmission delay (T^{tr}), and the total propagation delay (T^{pro}).

$$T^{\mathrm{tot}} = T^{\mathrm{proc}} + T^{tr} + T^{\mathrm{pro}}, \tag{14a}$$
$$T^{\mathrm{proc}} = T^{\mathrm{RU}} + T^{\mathrm{DU}} + T^{\mathrm{CU}} + T^{\mathrm{UPF}}, \tag{14b}$$
$$T^{\mathrm{tr}} = T^{\mathrm{fr},t} + T^{\mathrm{mid},t} + T^{\mathrm{b},t}, \tag{14c}$$
$$T^{\mathrm{pro}} = T^{\mathrm{fr},p} + T^{\mathrm{mid},p} + T^{b,p}. \tag{14d}$$

Mathematically, the total propagation delay (T^{pro}) is the sum of the propagation delay in the fronthaul link $T^{fr,p}$, the midhaul link $T^{\mathrm{mid},p}$, and the backhaul link $T^{b,p}$. In each link, the propagation delay is the time a signal to reach its destination. It is obtained based on the length of the fiber link and the capacity of the link (as T = L / c, where L is the length of the link and c is the propagation speed of the medium). Meanwhile, the total transmission delay (T^{tr}) is the sum of the transmission delay in the fronthaul $T^{fr,t}$, the midhaul $T^{\mathrm{mid},t}$, and the backhaul $T^{b,t}$. In each link, the transmission delay is the amount of time required to push all the packets into the transmission medium, and can formulated as $T = \frac{\alpha}{R}$, where *R* is the data-rate of the packet and α is the mean packet size. Notice that taking the propagation and transmission delays into account in the formulation is straightforward, but we have avoided it for the sake of succinctness and simplicity. Therefore, the propagation delay is fixed and does not affect the optimization problem.

Next, we present a brief calculation of propagation delay. Assume a distance between the O-RU and O-DU around 10 km, the distance between O-DU and O-CU around 80 km, not greater than the distance from O-CU to the network around 200 km [35]. Then, assuming the fronthaul, midhaul and backhaul are connected with fiber optics and *c* is the speed of like, the propagation delay is about $T^{\mathrm{pro}} = (10 + 80 + 200) \times 10^3/(3 \times 10^8) < 1$ ms.

The following is a brief calculation of the transmission delay to show that its contribution to the total delay is negligible and does not affect the optimization. In URLLC and mMTC, the mean packet size may be between 20 to 32 bytes, while the minimum data rate is assumed to be $1\ bps/Hz \times BW(180\ KHz)$. Thus, the transmission delay from O-RU to O-DU is about $T^{fr,t} = \frac{20\times 8}{1\times 180\times 10^3} < 0.1\ ms$, and $T^{fr,t} \approx T^{mid,t} \approx T^{b,t}$. For eMBB, the packet size may be 100 times larger and the delay does not exceed 0.6 ms. Therefore, in the following, we assume that the total delay is approximate to the processing delay ($T^{\mathrm{tot}} \approx T^{\mathrm{proc}}$).

1) Processing Delay: Assume the packet arrival of UEs follows a Poisson process with arrival rate $\lambda_{u(s,i)}$ for the i^{th} UE of the s^{th} service (or slice). Therefore, the mean arrival data rate of the s^{th} slice in the UPF layer is $\alpha^U_s = \sum_{u=1}^{U_s} \lambda_{u(s,i)}$. Assume the mean arrival data rate of the UPF layer for slice *s* (α^U_s) is approximately equal to the mean arrival data rate of the O-CU-UP layer (α^C_s) and the O-DU (α^D_s), i.e., $\alpha_s = \alpha^U_s \approx \alpha^C_s \approx \alpha^D_s$. This is because the amount of data transferred along the route (regardless of frame changes) is constant. In fact, according to Burke's theorem, the mean arrival data rate of the second and third layers, which are processed in the first layer, is still Poisson with rate α_s. It is assumed that there are load balancers in each layer for each service to divide equally the incoming traffic to VNFs. Suppose the baseband processing of each VNF is modeled by an M/M/1 processing queue. Each packet is processed by one of the VNFs of the corresponding slice. Therefor, the mean delay for the s^{th} slice in the O-DU, the O-CU, and the UPF is modeled as M/M/1 queue, and can be respectively [8], [22], [23],

$$T^{\mathrm{DU}}_s = 1/(\mu^d_s - \alpha_s/M^d_s), \tag{15}$$
$$T^{\mathrm{CU}}_s = 1/(\mu^c_s - \alpha_s/M^c_s), \tag{16}$$
$$T^{\mathrm{UPF}}_s = 1/(\mu^u_s - \alpha_s/M^u_s), \tag{17}$$

where M^d_s, M^c_s and M^u_s represent the number of VNFs in O-DU, O-CU-UP and UPF, respectively. Moreover, $1/\mu^d_s$, $1/\mu^c_s$, and $1/\mu^u_s$ are the mean service times of the O-DU, O-CU, and the UPF layers, respectively. The arrival rate of each VNF for each slice *s* is $\alpha_s/M^i_s\ i \in \{d, c, u\}$.

On the other hand, arrival data rate of wireless link for each UE *i* of service *s* is $\lambda_{u(s,i)}$, thus $\sum_{i=1}^{U_s} \lambda_{u(s,i)} = \alpha_s$. Moreover, the service time of transmission queue for UE *i* requesting service *s* has an exponential distribution with mean $1/R_{u(s,i)}$ and can be modeled as a M/M/1 queue [8], [22], [23]. Therefore, the mean delay of the transmission layer for UE *i* in slice *s* is

$$T^{\mathrm{RU}}_{u(s,i)} = 1/(R_{u(s,i)} - \lambda_{u(s,i)}). \tag{18}$$

we assume $T^{\mathrm{tot}}_{u(s,i)} \approx T^{\mathrm{proc}}_{u(s,i)}$.

F. VNF Power

Assume the power consumption of each VNF in each logial node (O-DU, O-CU, and UPF) in the slice s, is represented by ϕ^d_s, ϕ^c_s, and ϕ^u_s, respectively. Then, the system's total cost of energy of all the slices can be represented as $\phi_{\mathrm{tot}} = \sum_{s=1}^{S} \phi_s$.

A significant issue facing the industry is reducing energy consumption. Data centers are one of the most energy-consuming. As a result, restrictions are placed on data centers' energy, including VMs. So, one of our goals is to limit the energy consumption of total VNFs that can be run as VM on data centers. So, by applying a custom policy on total power consumption, we can control data centers' power consumption ($\phi^{\mathrm{tot}} \leq \phi^{\mathrm{max}}$).

V. Problem Statement

Suppose the slice *s* (which is assigned to service s) has a priority factor δ_s (based on the priority of its hosting service) where $\sum_{s=1}^{S} \delta_s = 1$. The priority factor of each slice is obtained according to the service level agreement to promote a fairness in the system. This paper aims to maximize the sum-rate of all UEs subject to QoS constraints as follows

$$\max_{\boldsymbol{P},\boldsymbol{E},\boldsymbol{M},\boldsymbol{G}} \sum_{s=1}^{S}\sum_{i=1}^{U_s} \delta_s \bar{\mathcal{R}}_{u_{(s,i)}} \tag{19a}$$

$$\text{subject to } P_r \le P_r^{\max} \quad \forall r, \tag{19b}$$

$$p_{r,u(s,i)}^{k} \ge 0 \quad \forall i,r,s,k, \tag{19c}$$

$$p_{r,u(s,i)}^{k} \le P_s^{\max} \quad \forall i,r,s,k, \tag{19d}$$

$$\bar{\mathcal{R}}_{u_{(s,i)}} \ge \mathcal{R}_s^{\min} \quad \forall s, \tag{19e}$$

$$C_r \le C_r^{\max} \quad \forall r, \tag{19f}$$

$$T_{u(s,i)}^{\mathrm{tot}} \le T_s^{\max} \quad \forall i,s, \tag{19g}$$

$$\mu_s \ge \alpha_s / M_s \quad \forall s, \tag{19h}$$

$$\bar{\mathcal{R}}_{u_{(s,i)}} \ge \lambda_{u_{(s,i)}} \quad \forall i,s, \tag{19i}$$

$$0 \le M_s \le M_s^{\max} \quad \forall s, \tag{19j}$$

$$\phi^{\mathrm{tot}} \le \phi^{\max}, \tag{19k}$$

$$\sum\nolimits_{\forall r} g_{u(s,i)}^{r} = 1 \quad \forall s,i, \tag{19l}$$

$$\sum_{k=1}^{K_s} g_{u(s,i)}^{r} e_{r,u(s,i)}^{k} \ge 1 \quad \forall s,i,r, \tag{19m}$$

$$\sum_{s=1}^{S}\sum_{i=1}^{U_s} g_{u(s,i)}^{r} e_{r,u(s,i)}^{k} \le 1 \quad \forall s,i,r, \tag{19n}$$

$$g_{u(s,i)}^{r} \in \{0,1\} \quad \forall s,i, \tag{19o}$$

$$e_{r,u(s,i)}^{k} \in \{0,1\} \quad \forall s,i. \tag{19p}$$

Here, $\bar{\mathcal{R}}_{u_{(s,i)}}$ is derived by using $\bar{I}_{r,u(s,i)}^{k}$ instead of $I_{r,u(s,i)}^{k}$ in (8) and (5). In addition, $\boldsymbol{P} = [p_{r,u(s,i)}^{k}],\ \forall s,i,r,k$, is the four-dimensional (4D) matrix of power for UEs, $\boldsymbol{E} = [e_{r,u(s,i)}^{k}],\ \forall s,i,r,k$ indicates the binary 4D matrix for the PRB association. Moreover, $\boldsymbol{G} = [g_{u(s,i)}^{r}],\ \forall s,i,r$ is a binary three dimensional (3D) matrix for the O-RU association. Furthermore, $M = [M_s^d, M_s^c, M_s^u],\ \forall s$ is a matrix containing the number of VNFs in each layer of slice. Notice that (19b), (19c) and (19d) limit the power of each O-RU and UE. Also, (19e) constrains the rate of each UE requesting each type of service, i.e., eMBB, mMTC, and URLLC, to be greater than a threshold. Meanwhile, (19f) and (19g) represent the limited fronthaul capacity and the limited end-to-end delay of the received signal, respectively. (19h) and (19i) are related to the stability of the M/M/1 queue, (19j) restrictes the number of VNFs in each slice due to the limited resources, while (19l) and (19m) guarantee that the O-RU and PRB are associated with the UE, respectively. Also, (19n) ensures that each PRB can not be assigned to more than one UE associated with the same O-RU, (19k) indicates that the fixed cost of energy of VNFs in each slice does not exceed the threshold, while (19o) and (19p) constrain $\boldsymbol{E}$ and $\boldsymbol{G}$ to be binary matrices.

A. PRB Scheduling

In this section, we provide a brief study on the problem of PRB scheduling which can be completed in two steps to eliminate the inter-slice interference and guarantee the isolation of slices [36]. Firstly, we should assign the PRBs to the slices. Secondly, we assign PRBs of slices to UEs, find the optimal number of VNFs for each slice, allocate power of UEs, and assign O-RU to UEs, which uses the proposed Algorithm VI. Suppose, $\mathcal{R}_s^{\min}$, and $\mathcal{R}_s^{\max}$ are the minimum data rate and maximum data rate of each UE in slice s, respectively. Firstly, we need to find the average PRB number used by the UEs in each service. Since mMTC and URLLC require usually short packet transmissions, each UE in mMTC and URLLC requires 1 PRB. So if slice s serves mMTC or URLLC services, with U_s UEs, it requires $K_s = U_s \times 1$ PRBs. For eMBB, assume the average rate of each UE in slice s serving eMBB UEs is $\bar{R}_s = B\log_2(1+\bar{\rho_s})$, where, $\bar{\rho_s}$ is the average SINR of UEs in slice s. Therefore, the minimum number of PRBs that slice s with U_s UEs requires is $K_s^{\min} = \lceil U_s \times \frac{\bar{R}_s}{\mathcal{R}_s^{\max}} \rceil$. Moreover, the maximum number of PRB that the slice s with U_s UEs requires is $K_s^{\max} = \lceil U_s \times \frac{\bar{R}_s}{\mathcal{R}_s^{\min}} \rceil$. Also, $K_s = (K_s^{\min} + K_s^{\max})/2$ is the average number of required PRBs in slice s. We aim to obtain the number of PRBs assigned to each slice s ($\bar{K}_s$) by solving

$$\max_{\bar{\boldsymbol{K}}_s} \sum_{s=1}^{S} \delta_s K_s \ln(\bar{K}_s) \tag{20a}$$

$$\text{subject to } \sum_{s} \bar{K}_s \le K, \tag{20b}$$

$$K_s^{\min} \le \bar{K}_s \le K_s^{\max} \quad \forall s \in S_1, \tag{20c}$$

$$\bar{K}_s \le K_s \quad \forall s \in S_2, S_3. \tag{20d}$$

We use logarithms to assign PRBs to all slices to make them equally fair, since proportional fairness is achieved by maximizing the log utility function [36]. Equation (20b) illustrates that the sum of PRBs of slices can not exceed the maximum number of PRBs (K). Equation (20c) restricts the number of PRBs of eMBB slices and (20d) limits the number of the PRBs of URLLC and mMTC slices. By relaxing $\bar{K}_s$, the objective function and constraints become convex and can be solved using the Lagrangian function.

B. Slice Management

In this section, we will look at the life cycle of network slicing on a practical level. The goal is to examine slice management, which includes creating, managing, and deleting slices. Network slices generally have four life cycle stages [37]:

- Preparation phase: the operator plans to create a network slice instance (NSI) by designing the its template, onboarding users, and preparing the environment. Also, the evaluation of requirements is performed in this step.
- Commissioning phase: the NSI is created, and the requirements are considered and allocated to the slice.
- Operation phase: the NSIs are activated, managed, monitored (e.g., KPIs), modified, and deactivated. As the slice enters the activated phase, it is ready to support services, and as the slice exits the de-activated phase, the slice is inactive, and communication services are stopped.
- Decommissioning phase: an NSI that is decommissioned no longer exists after this phase.

Since the requirements evaluation is considered in the preparation phase, we need an algorithm to estimate the UE traffic in the system at different times. Moreover, based on this estimation, we need to evaluate resources, including the optimal number of VNFs, PRB assignment of UEs for each slice, and the total power requirements. In this phase, we use our algorithm to calculate resources after estimating the system's traffic. As shown in Fig. 1, we have three different slices for eMBB, URLLC, and mMTC. The system must prepare VNFs for MAC/RLC functions in O-DU, PDCP/SDAP functions in O-CU, UPF, SMF, and AMF functionality layers for each slice. Moreover, O-RU, high PHY in O-DU, and O-CU-CP are shared between slices. Thus, we do not require evaluating and preparing for the share environments and platforms in the network slicing cycles. Moreover, the estimation of PRB and power is needed based on the proposed algorithm. After evaluating, assessing, and preparing the resources and environments for each slice, the commissioning phase is started. In this phase, the slices are created based on the previous phase estimation. These created slices are activated in the operation phase, and the actual resources are assigned based on the proposed algorithm. It is possible to modify the slice's resources even when the evaluation changes during the operation phase. If we need to remove a slice or any service not used in a zone, the unshared resources are released in the decommissioning phase.

VI. Proposed Algorithm

In this section, we first apply some simplifications to the system; Solving the problem (19) is complicated since this is non-convex mixed-integer non-linear problem (MINLP) with a binary variable and an integer variable. We applied some simplifications and use an iterative heuristic algorithm to solve the problem. We solve this problem in two levels, iteratively, until it converges [25].

At the first level, the main purpose is to assign appropriate PRBs and power to the UEs. Furthermore, sufficient activated VNFs are assigned to each slice. Hence, at this level, we would like to obtain the variables $\boldsymbol{P}$, $\boldsymbol{E}$, and $\boldsymbol{M}$. Despite the simplification of the problem (19), it is still NP-hard and challenging to solve. Therefore, we relax the variable $\boldsymbol{E}$ [7], [25] and reformulating the constraint (19g), to turn them into a jointly-convex problem; Afterward, we solve this problem using a conventional dual Lagrangian method. In the second level, finding the optimal O-RU association, $\boldsymbol{G}$, is concerned with the fixed parameter of power, PRB allocation, and the number of activated VNFs. We repeat this procedure until the algorithm converges.

A. Sub-Problem 1

Suppose that $\boldsymbol{G}$ is fixed, we want to obtain $\boldsymbol{P}$, $\boldsymbol{E}$ and $\boldsymbol{M}$. Here, we first simplify and relax the parameters to convexify the problem. As we mentioned before, by replacing $p^k_{u(s,l)}$ and $p^k_{u(n,l)}$ in (3) with $P^{\max}_s$, an upper bound $\bar{I}^k_{r,u(s,i)}$ is obtained for $I^k_{r,u(s,i)}$, and also the lower bound $\bar{\rho}^k_{u(s,i)}$ is achieved for $\rho^k_{u(s,i)}$. Moreover, the lower bound $\bar{\mathcal{R}}_{u_{(s,i)}}, \forall s, \forall i$ for $\mathcal{R}_{u_{(s,i)}}$ is obtained by replacing $I^k_{r,u(s,i)}$ with $\bar{I}^k_{r,u(s,i)}$ in (8) and (5) and make these equations become concave functions.

Suppose $\hat{\rho}^k_{r,u(s,i)} = P^{\max}_s |\mathbf{h}^{H\,k}_{r,u(s,i)} \mathbf{w}^k_{r,u(s,i)} g^r_{u(s,i)}|^2/(BN_0)$. We replace $\rho^k_{r,u(s,i)}$ with $\hat{\rho}^k_{r,u(s,i)}$ in (11), to convexify the (8) for the URLLC and mMTC services that have the short packet transmission. So, a lower bound for (8) is given that is a concave function.

$$\bar{\mathcal{R}}^r_{u(s_j,i)} = B \sum_{k=1}^{K_{s_j}} e^k_{u(s_j,i)} (\log_2(1+\bar{\rho}^k_{u(s_2,i)}) - \hat{\zeta}^k_{u(s_j,i)}), \tag{21a}$$

$$\bar{\mathcal{R}}_{u(s_j,i)} = \sum_{r=1}^{R} \bar{\mathcal{R}}^r_{u(s_j,i)}, \tag{21b}$$

$$\hat{\zeta}^k_{u(s_j,i)} = \log_2(e) Q^{-1}(\epsilon) \sqrt{\hat{\mathfrak{C}}^k_{u(s_j,i)}/N^k_{u(s_j,i)}}, \tag{21c}$$

$$\hat{\mathfrak{C}}^k_{u(s_j,i)} = 1 - 1/(1+\hat{\rho}^k_{u(s_j,i)})^2. \tag{21d}$$

Without loss of generality, assume that UPF, O-CU and O-DU use the processors with the same processing capability. We notice that it makes the formulation simpler. However, loosing this assumption does not change the formulation significantly and the problem can be solved in the same manner. Therefore, we have $\mu_s = \mu^u_s \approx \mu^c_s \approx \mu^d_s$. Moreover, as mentioned before, the mean arrival data rate of the UPF layer for a service s (α^U_s) is equal to the mean arrival data rate of the O-CU-UP layer (α^C_s) and O-DU (α^D_s). So $\alpha_s = \alpha^U_s \approx \alpha^C_s \approx \alpha^D_s$. Again, this assumption only simplifies the notations and loosing it does not make the solution inefficient. These assumptions lead to having the same processing power for each layer $\phi^u_s = \phi^c_s = \phi^d_s$. As a result, we have $M_s = M^u_s = M^c_s = M^d_s$. Using the above assumption, we have $T^{\mathrm{DU}}_s = T^{\mathrm{CU}}_s = T^{\mathrm{UPF}}_s$ and we have $T^{\mathrm{proc}}_s = T^{\mathrm{RU}}_s + T^{\mathrm{DU}}_s + T^{\mathrm{CU}}_s + T^{\mathrm{UPF}}_s$. So, $T^{\mathrm{proc}}_s = T^{\mathrm{RU}}_s + 3 \times T^{\mathrm{DU}}_s$.

The problem (19) is mixed-integer nonlinear programming with two integer variables, the PRB assignment, e, and the number of VNFs in slice s, M_s, and by relaxing the variables, the problem is also non-convex; therefore, this problem is NP-hard. Solving the problem is not trivial. To solve the problem by inspiring Stackelberg, we reformulate the equation in (19g) to reduce one of the variables (i.e., M_s) that can be solved after obtaining the rate of UEs. We notice that M_s is similar to the followers in Stackelberg Competition, and power and PRB assignment are identical to the leader. So, the new problem has two variables: power and PRB assignment. This new problem is convex by relaxing the binary variable, the PRB assignment, and estimating the lower bounds (21). The objective function and constraints of the problem are convex and can be solved by the Lagrangian function. After obtaining the power of UEs and PRB assignment, we can obtain the achievable rate of each UE so we can find the optimal number of VNFs in each slice (M_s).

In the following, we define a lemma to find the upper and lower bounds for the optimal number of VNFs based on the achievable rates. Afterward, we obtain the formula to attain the optimal number of VNFs.

Lemma 1: The optimal number of VNFs in each slice s can be achieved by the $M_s = \max\{\mathtt{M}_{u(s,i)} | i \in 1, 2, \ldots, U_s\} \quad \forall s$.

where, $\mathtt{M}_{u(s,i)} = \frac{\alpha_s(T_s^{\max} R_{u(s,i)} - T_s^{\max}\lambda_{u(s,i)} - 1)}{(T_s^{\max}\mu_s - 3)(R_{u(s,i)} - \lambda_{u(s,i)}) - \mu_s}$ for each UE i in slice s.

Proof: In problem (19), the constraint (19g) can be reformulated as

$$T_s^{\max} \geq \frac{1}{R_{u(s,i)} - \lambda_{u(s,i)}} + \frac{3}{\mu_s - \alpha_s/M_s}, \tag{22a}$$

$$M_s \geq \frac{\alpha_s(T_s^{\max} R_{u(s,i)} - T_s^{\max}\lambda_{u(s,i)} - 1)}{(T_s^{\max}\mu_s - 3)(R_{u(s,i)} - \lambda_{u(s,i)}) - \mu_s}. \tag{22b}$$

Also from equations in (19k), (19h) and (19j), we have

$$\alpha_s/\mu_s \leq M_s \leq \min\{M^{\max}, \phi_{\max}/3\phi_s\}. \tag{23}$$

We denote $\mathfrak{M}_s = \min\{M^{\max}, \phi_{\max}/3\phi_s\}$. Thus, if we restrict constraint (19g) to equality, constraint (19g) is still valid. Also, we have the following inequality

$$\alpha_s/\mu_s \leq \mathtt{M}_{u(s,i)} \leq \mathfrak{M}_s, \tag{24}$$

where $\mathtt{M}_{u(s,i)} = \frac{\alpha_s(T_s^{\max} R_{u(s,i)} - T_s^{\max}\lambda_{u(s,i)} - 1)}{(T_s^{\max}\mu_s - 3)(R_{u(s,i)} - \lambda_{u(s,i)}) - \mu_s} \geq 0$ since the numerator and the denominator both have the same sign. In the numerator, according to (19i), $R_{u(s,i)} - \lambda_{u(s,i)} \geq 0$, and as we know that $\alpha_s \geq 0$, we have $\alpha_s(R_{u(s,i)} - \lambda_{u(s,i)}) \geq 0$. If we assume that the $(R_{u(s,i)} - \lambda_{u(s,i)})T_s^{\max} \geq 1$, the numerator will be positive. $(R_{u(s,i)} - \lambda_{u(s,i)})T_s^{\max} \geq 1$ since the order of $T_s^{\max}$ is about milli second and the difference between achievable rate and packet rate can be more than $1/T_s^{\max}$. Therefore, to ensure that this constraint will be valid, we restrict constraint (19i) to $R_{u(s,i)} \geq \lambda_{u(s,i)} + 1/T_s^{\max}$. So the numerator will be positive. In the denominator, we can say that $(T_s^{\max}\mu_s)(R_{u(s,i)} - \lambda_{u(s,i)}) - \mu_s \geq 0$, since, $\mu_s \geq 0$ and $(R_{u(s,i)} - \lambda_{u(s,i)}) \geq 1/T_s^{\max}$ as mentioned above. The left side of the equation (24), leads to $R_{u(s,i)} \geq \lambda_{u(s,i)}$ that is the constraint (19i). For the right side, by reformulating the equation (24), we have a new constraint $\forall i, \forall s$ given by

$$\begin{aligned}\mathcal{R}_{u(s,i)} \geq \lambda_{u(s,i)} &+ 3/(T_s^{\max}\mu_s - \alpha_s T_s^{\max}/\mathfrak{M}_s - 3) \\ &+ 1/T_s^{\max} = \varpi_{u(s,i)}.\end{aligned} \tag{25}$$

Thus, to obtain the optimal number of activated VNF in each slice, we need to find the maximum of the $\mathtt{M}_{u(s,i)}$ in each slice as $M_s = \max\{\mathtt{M}_{u(s,i)} | i \in 1, 2, \ldots, U_s\} \quad \forall s$. ■

Despite simplifying the problem in (19), it is still non-convex and hard to solve. Therefore, the conventional approach to solve the problem of the PRB and the power allocation is to relax the variable $\mathbf{E}$ into continuous value $e^k_{r,u(s,i)} \in [0,1]\,\forall s, \forall i, \forall r, \forall k$ [7], [25]. Furthermore, the problem can be solved using the Lagrangian function and iterative algorithm.

In order to transform (19) into a convex optimization problem in standard form, it is required to change the variable of equations (13) to $P_r = \sigma^2_{q_r} \times 2^{C^r}$ so the constraint (19f) is changed to $P_r \leq \sigma^2_{q_r} \times 2^{C^r_{\max}}$. The combination of equations (19b) and (19f) leads to the following equation

$$P_r \leq \zeta_r = \min\{P_{\max}, \sigma^2_{q_r} \times 2^{C^r_{\max}}\}. \tag{26}$$

Moreover, the combination of equations in (19e), (19i) and (25) leads to the following equation

$$\bar{R}_{u_{(s,i)}} \geq \eta_{u_{(s,i)}} = \max\{\mathcal{R}^{\min}_{u_{(s,i)}}, \lambda_{u_{(s,i)}} + \frac{1}{T^s_{\max}}, \varpi_{u(s,i)}\}. \tag{27}$$

Assume $\boldsymbol{\mathfrak{v}}$, $\boldsymbol{\mathfrak{m}}$, $\boldsymbol{\mathfrak{h}}$, $\boldsymbol{\xi}$, $\boldsymbol{\chi}$, $\boldsymbol{\mathfrak{q}}$ and $\boldsymbol{\kappa}$ are the matrix of Lagrangian multipliers that have non-zero positive elements. The Lagrangian function is written as

$$\begin{aligned}\mathcal{L}(\boldsymbol{P}, \boldsymbol{E}; \boldsymbol{\mathfrak{v}}, \boldsymbol{\chi}, \boldsymbol{\mathfrak{h}}, \boldsymbol{\xi}, \boldsymbol{\kappa}, \boldsymbol{\mathfrak{m}}) &= \sum_{s=1}^{S}\sum_{i=1}^{U_s} \delta_s \bar{R}_{u_{(s,i)}} \\ &+ \sum_{s=1}^{S}\sum_{i=1}^{U_s} \mathfrak{h}_{u_{(s,i)}}(\bar{R}_{u_{(s,i)}} - \eta_{u_{(s,i)}}) - \sum_{r=1}^{R} \mathfrak{m}_r(P_r - \zeta_r) \\ &+ \sum_{s=1}^{S}\sum_{i=1}^{U_s}\sum_{k=1}^{K}\sum_{r=1}^{R}\Big(\kappa^k_{r,u(s,i)} p^k_{r,u(s,i)} - \mathfrak{v}^k_{r,u(s,i)}(e^k_{r,u(s,i)} - 1) \\ &+ \xi^k_{r,u(s,i)} e^k_{r,u(s,i)} + \mathfrak{q}^k_{r,u(s,i)}(P_s^{\max} - p^k_{r,u(s,i)})\Big) \\ &+ \sum_{r=1}^{R}\sum_{s=1}^{S}\sum_{i=1}^{U_s} \chi_{r,u(s,i)}\left(\sum_{k=1}^{K_s} e^k_{r,u(s,i)} - 1\right).\end{aligned} \tag{28}$$

Lemma 2: The derivatives of the Lagrangian function (28) with respect to the $\boldsymbol{P}$ and $\boldsymbol{E}$ give the KKT conditions to obtain the optimal value of these two variables [7], [25].

Proof: Assume UE i in slice s, associated with O-RU r, is allocated to PRB k (i.e., $e^k_{r,u(s,i)} = 1$). Therefore, we have the following KKT condition

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial p^k_{r,u(s,i)}} &= (\delta_s + \mathfrak{h}_{u_{(s,i)}})\mathfrak{B}^k_{r,u(s,i)} \\ &+ (\mathfrak{s}^k_{r,u(s,i)} - \mathfrak{D}^k_{r,u(s,i)}) = 0,\end{aligned} \tag{29}$$

where $\mathfrak{s}^k_{r,u(s,i)} = \kappa^k_{r,u(s,i)} - \mathfrak{q}^k_{r,u(s,i)}$ and other parameters are as follows

$$\mathfrak{D}^k_{r,u(s,i)} = \mathfrak{m}_r |\mathbf{w}^k_{r,u(s,i)}|^2 g^r_{u(s,i)} e^k_{r,u(s,i)}, \tag{30a}$$

$$\mathfrak{B}^k_{r,u(s,i)} = B|\mathbf{h}^{H\ k}_{r,u(s,i)} \mathbf{w}^k_{r,u(s,i)}|^2 g^r_{u(s,i)} e^k_{r,u(s,i)} \mathfrak{S}^k_{r,u(s,i)}/\ln(2), \tag{30b}$$

$$\mathfrak{S}^k_{r,u(s,i)} = (|\mathbf{h}^{H\ k}_{r,u(s,i)} \mathbf{w}^k_{r,u(s,i)}|^2 \mathfrak{k}^k_{r,u(s,i)} + BN_0 + I^k_{r,u(s,i)})^{-1}. \tag{30c}$$

Also, $\mathfrak{k}^k_{r,u(s,i)} = g^r_{u(s,i)} e^k_{r,u(s,i)} p^k_{r,u(s,i)}$. Thus, from equation (29), optimal power is obtained and power is allocated. We denote $\mathfrak{j}^k_{r,u(s,i)} = g^r_{u(s,i)} e^k_{r,u(s,i)}$. The optimal power is as follow.

$$\begin{aligned}&p^k_{r,u(s,i)} \\ &= \left[\frac{(\delta_s + \mathfrak{h}_{u_{(s,i)}})B\mathfrak{j}^k_{r,u(s,i)}}{\ln(2)(\mathfrak{D}^k_{r,u(s,i)} - \mathfrak{s}^k_{r,u(s,i)})} - \frac{BN_0 + I^k_{r,u(s,i)}}{|\mathbf{h}^{H\ k}_{r,u(s,i)} \mathbf{w}^k_{r,u(s,i)}|^2 \mathfrak{j}^k_{r,u(s,i)}}\right]^+.\end{aligned} \tag{31}$$

Also $[a]^+ = \max(0, a)$. In addition, PRB assignment can be achieved from the derivatives of the Lagrangian function (28) with respect to the $\boldsymbol{E}$ as follow.

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial e^k_{r,u(s,i)}} &= \bar{R}^k_{r,u(s,i)}(\delta_s + \mathfrak{h}_{u_{(s,i)}}) - \mathfrak{m}_r |\mathbf{w}^k_{r,u(s,i)}|^2 p^k_{r,u(s,i)} g^r_{u(s,i)} \\ &+ (\xi^k_{r,u(s,i)} - \mathfrak{v}^k_{r,u(s,i)} + \chi_{r,u(s,i)}) = 0.\end{aligned} \tag{32}$$

So, the optimal $\boldsymbol{E}$ is obtained using the KKT conditions, which require solving

$$e^k_{r,u(s,i)}(\mathfrak{F}^k_{r,u(s,i)} - \mathfrak{v}^k_{r,u(s,i)} - \mathfrak{m}_r |\mathbf{w}^k_{r,u(s,i)}|^2 p^k_{r,u(s,i)} g^r_{u(s,i)}) = 0, \tag{33}$$

Algorithm 1 Greedy Algorithm for Assignment of O-RU to UEs (GAA)

```
Set $g^r_{u(s,i)} = 0$, $\mathfrak{C}_r = \mathfrak{t}_r$, and $\mathfrak{B}^{rem}_{u_{(s,i)}} = \mathcal{R}\ \forall s, \forall i, \forall r$.
Sort slices according to their $\delta_s$ in descending order
for $s \leftarrow 1$ to $S$ do
    for $i \leftarrow 1$ to $U_s$ do
        $RU = 0$
        for $r \leftarrow 1$ to $R$ do
            Acquire $\mathfrak{G}^r_{u_{(s,i)}} = \bar{\mathcal{R}}^r_{u_{(s,i)}}$
        end for
        Obtain $r^* = \text{argmax}_{r \in \mathfrak{B}^{rem}_{u_{(s,i)}}} \mathfrak{G}^r_{u_{(s,i)}}$
        while $RU == 0$ do
            if $\mathfrak{C}_{r^*} \geq \psi_{r^*,u(s,i)}$ then
                Set $g^{r^*}_{u(s,i)} = 1$
                Set $\mathfrak{C}_{r^*} = \mathfrak{C}_{r^*} - \psi_{r^*,u(s,i)}$
                Set $RU = 1$
            else: $\mathfrak{B}^{rem}_{u_{(s,i)}} = \mathfrak{B}^{rem}_{u_{(s,i)}} \setminus \{r^*\}$
            end if
        end while
    end for
end for
```

where $\mathfrak{F}^k_{r,u(s,i)} = \bar{R}^k_{r,u(s,i)}(\delta_s + \mathfrak{h}_{u_{(s,i)}}) + (\xi^k_{r,u(s,i)} + \chi_{r,u(s,i)})$. Hence, from equation (32) and (33), PRB assignment is performed as follows

$$e^k_{r,u(s,i)} = \begin{cases} 1 & u(s,i) = \text{argmax}\, \mathfrak{Z}^k_{r,u(s,i)} \forall r, k \in K, s \in S, \\ 0 & \text{otherwise,} \end{cases} \tag{34}$$

where $\mathfrak{Z}^k_{r,u(s,i)} = (\mathfrak{F}^k_{r,u(s,i)} - \mathfrak{v}^k_{r,u(s,i)} - \mathfrak{m}_r |\mathbf{w}^k_{r,u(s,i)}|^2 p^k_{r,u(s,i)} g^r_{u(s,i)})$. ■

Thus, the user in slice *s* that has the most considerable value of $\mathfrak{F}^k_{r,u(s,i)}$, should be allocated to PRB *k*. Since just one PRB can be allocated to a UE between those UEs (regardless of the services), that is associated to the same O-RU. The number of UEs are $\mathfrak{N} = \sum_{s=1}^{S} \sum_{i=1}^{U_S} 1$. Also, assume that the algorithm converges after T_{conv} times. The complexity order of this problem is about $O(T_{conv} \times \mathfrak{N} \times K)$.

B. Sub-Problem 2

After power allocation and PRB assignment, the remaining problem is to assign O-RU to each UE in each service. For this, assume $\boldsymbol{P}$ and $\boldsymbol{E}$ are fixed, so we want to find $\boldsymbol{G}$. Next, we introduce a greedy algorithm that assigns an O-RU to each UE.

Greedy Algorithm Assignment (GAA): The following is a reformulated version of the problem.

$$\max_{\boldsymbol{G}} \sum_{s=1}^{S} \sum_{i=1}^{U_s} \sum_{r=1}^{R} \delta_s g^r_{u(s,i)} \bar{\mathcal{R}}^r_{u_{(s,k)}} \tag{35a}$$

$$\text{subjectto} \sum_{s=1}^{S} \sum_{i=1}^{U_s} g^r_{u(s,i)} \psi_{r,u(s,i)} \leq \mathfrak{t}_r \quad \forall r, \tag{35b}$$

$$\sum_{r} g^r_{u(s,i)} = 1 \quad \forall s, i, \tag{35c}$$

$$g^r_{u(s,i)} \in \{0, 1\} \quad \forall s, i, \tag{35d}$$

where $\psi_{r,u(s,i)} = \sum_{k=1}^{K_s} |\mathbf{w}^k_{r,u(s,i)}|^2 p^k_{r,u(s,i)} e^k_{r,u(s,i)}$ and $\mathfrak{t}_r = \zeta_r - \sigma_r$ because of the equations (26) and (12). Since we obtained (27) in Section VI-A, we can ignore this constraint in (35). The problem (35) is an NP-complete 0-1 multiple knapsack problem. We solve this problem using the heuristic GAA Algorithm 1, which is a greedy algorithm [7], [38]. Firstly, we set all the variables to zero ($g^r_{u(s,i)} = 0, \quad \forall s, i, r$). Then we define the parameter $\mathfrak{B}^{rem}_{u_{(s,i)}}$. This parameter is used as a set of O-RUs that can be assigned to the UE *i* in slice *s*, which initially includes all the O-RUs ($\mathfrak{B}^{rem}_{u_{(s,i)}} = \mathcal{R}, \forall s, i$). Also we introduce another parameter $\mathfrak{C}_r = \mathfrak{t}_r, \forall r$ which is the knapsack capacity of each O-RU. Next, we sort all the slices based on their priority. Afterward, based on the sorting of the UEs, we assign the O-RU that provides the highest achievable data rate for each UE on the condition that the value of the desired UE ($\psi_{r,u(s,i)}$) does not exceed the knapsack capacity of each O-RU ($\mathfrak{C}_r$). If it exceeds the capacity of the desired O-RU, we remove the specific O-RU from the set of O-RUs that can be assigned to that UE ($\mathfrak{B}^{rem}_{u_{(s,i)}} = \mathfrak{B}^{rem}_{u_{(s,i)}} \setminus \{r^*\}$). Then, the O-RU with the highest achievable data rate from the new set of O-RUs $\mathfrak{B}^{rem}_{u_{(s,i)}}$ is selected. The complexity of sorting S slices based on their priority is $O(S \log(S))$. Depict $\mathfrak{N} = \sum_{s=1}^{S} \sum_{i=1}^{U_s} 1$ as the whole number of UEs in the system. The complexity order of this algorithm is about $O(S \log(S)) + O(R \times \mathfrak{N})$.

Algorithm 2 Iterative Algorithm for the Baseband Resource Allocation and VNF Activation (IABV)

```
Set the maximum num. of iter. $\mathrm{I}_{\max}$, convergence condition $\epsilon > 0$
Assign Users to O-RU randomly (Initialize $\boldsymbol{G}$)
for $i \leftarrow 1$ to $\mathrm{I}_{\max}$ do
    Acquire $\boldsymbol{P}^{(i)}$, $\boldsymbol{E}^{(i)}$ and $\boldsymbol{M}^{(i)}$ using Lagrangian function and sub-gradient method based on (VI-A)
    Update $\boldsymbol{G}^{(i)}$ based on algorithm GAAOU (1) in (VI-B)
    if the algorithm converged with the tolerence of $\epsilon$ then
        Break
    else: Continue the algorithm
    end if
end for
```

C. Iterative Proposed Algorithm

In Sections VI-A and VI-B, the details of solving each sub-problem are depicted. Here, the iterative algorithm for the whole problem is demonstrated. Firstly, we fixed $\boldsymbol{G}$ to achieve $\boldsymbol{P}$ and $\boldsymbol{E}$, using the Lagrangian method and the KKT conditions. Afterward, $\boldsymbol{G}$ is updated using the GAA algorithm. This process is repeated until it converges. The whole algorithm (IABV method) is depicted as follows (Algorithm 2).

1) Complexity Order: The number of UEs are $\mathfrak{N} = \sum_{s=1}^{S} U_S$. Also, assume that the algorithm converges after T_{conv} times. As we mentioned before, the complexity order of the first sub-problem is about $O(T_{conv} \times \mathfrak{N} \times K)$ and the complexity order of the second sub-problem is about $O(S \log(S)) + O(R \times \mathfrak{N})$. So the complexity of the main problem (19) is $O(T_{conv} \times \mathfrak{N} \times K \times (S \log(S) + R\mathfrak{N}))$.

2) Convergence Analysis: Due to limited system resources, we have limits on VNFs' power, UE or O-RU power, fronthaul capacity, etc. As a result, the objective function, which is the aggregate throughput, cannot exceed its optimal value and become infinite. Therefore, if the aggregate throughput is infinite and increases without limit, the resources must also be unlimited. Hence, the system has an optimal solution: its maximum aggregate throughput in the feasible region.

Consequently, we can guarantee the convergence of the iterative algorithm if the objective function is the ascending function concerning the number of iterations and has an upper bound. Thus, it will converge to its optimum value if it is a strictly ascending function and to its local optimum if it is a non-monotonically ascending function.

Consider the aggregate throughput as $\mathcal{T}(\mathbf{P}, \mathbf{E}, \mathbf{G}) = \sum_{s=1}^{S} \sum_{i=1}^{U_s} \delta_s \bar{\mathcal{R}}_{u_{(s,i)}}$. In the first step of the iteration i of the Algorithm 2 (IABV), we have $\mathcal{T}(\mathbf{P}^i, \mathbf{E}^i, \mathbf{G}^{i-1})$. In this step, optimal power and PRB allocation are obtained for the fixed O-RU association, so we have $\mathcal{T}(\mathbf{P}^i, \mathbf{E}^i, \mathbf{G}^{i-1}) \geq \mathcal{T}(\mathbf{P}^{i-1}, \mathbf{E}^{i-1}, \mathbf{G}^{i-1})$. In the second step of the iteration i, the optimal O-RU association is achieved to maximize the aggregate throughput. So we have this inequality $\mathcal{T}(\mathbf{P}^i, \mathbf{E}^i, \mathbf{G}^i) \geq \mathcal{T}(\mathbf{P}^i, \mathbf{E}^i, \mathbf{G}^{i-1})$. As a result, we have $\mathcal{T}(\mathbf{P}^i, \mathbf{E}^i, \mathbf{G}^i) \geq \mathcal{T}(\mathbf{P}^{i-1}, \mathbf{E}^{i-1}, \mathbf{G}^{i-1})$. Hence, in each step of the iteration, the aggregate throughput increased. Note that $\mathcal{T}^*(\mathbf{P}^*, \mathbf{E}^*, \mathbf{G}^*)$ is the achieved aggregate throughput for all the feasible resource allocation solutions of $\{\mathbf{P}, \mathbf{E}, \mathbf{G}\}$. So, $\mathcal{T}^*(\mathbf{P}^*, \mathbf{E}^*, \mathbf{G}^*) \geq \mathcal{T}(\mathbf{P}^i, \mathbf{E}^i, \mathbf{G}^i)$ and thus in each iteration, the aggregate throughput can not be larger than the optimal solution.

In addition, if we assume that the interference is set to be zero $I^k_{r,u_{(s,i)}} = 0$, and we suppose that each UE has the maximum power $p^k_{r,u_{(s,i)}} = P^{max}_s$, and we consider that all PRB is assigned to all UE $e^k_{r,u_{(s,i)}} = 1 \forall s, \forall i$ and each UE is assigned to the nearest O-RU with the best channel quality. So, the solution of this allocation, is the upper bound for the aggregate throughput. Thus, we can guarantee the local convergence of our iterative algorithm since the objective function $\mathcal{T}$ is the ascending function concerning the number of iterations and it has the upper bound which is not infinite.

In addition, to extend our solution to the global optimum, we must prove that the algorithm monotonically increases in the non-optimal set of solutions and is Lipschitz monotone contraction mapping. Here, we briefly discussed our algorithm's global convergence for a low interference system.

In the low interference system, the PRB and VNF assignment is obtained straightforwardly, and the first step for solving the problem is power allocation. Given fixed O-RU association, we have $\mathcal{T}(\mathbf{P}^i, \mathbf{E}^i, \mathbf{G}^{i-1}) > \mathcal{T}(\mathbf{P}^{i-1}, \mathbf{E}^{i-1}, \mathbf{G}^{i-1})$ (strictly increase). Notice that for the power allocation problem, the objective function is convex, and the convex functions are Lipschitz monotone contractions. However, in the second step of the algorithm, we can show as before that the algorithm is increased, but we can not talk about monotonically increases. Hence we have $\mathcal{T}(\mathbf{P}^i, \mathbf{E}^i, \mathbf{G}^i) \geq \mathcal{T}(\mathbf{P}^i, \mathbf{E}^i, \mathbf{G}^{i-1})$. Nevertheless, the objective function is still Lipschitz monotone mapping. Accordingly, we have $\mathcal{T}(\mathbf{P}^i, \mathbf{E}^i, \mathbf{G}^i) > \mathcal{T}(\mathbf{P}^{i-1}, \mathbf{E}^{i-1}, \mathbf{G}^{i-1})$. Therefore the objective function is the strictly ascending function concerning the number of iterations and has an upper bound. Consequently, the algorithm converges to the global optimum solution in low interference.

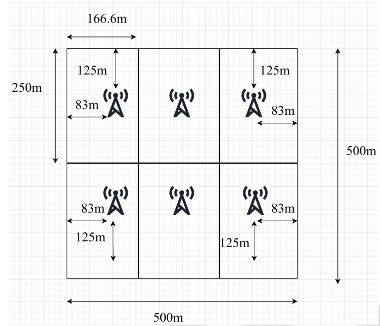

Fig. 2. O-RU placement in a cell.

VII. Numerical Results

In this section, firstly, we describe the initial points and the comparison algorithms. Then, we talk about the feasible region of our system model. Afterward, we illustrate the numerical results.

A. The Initial Points and the Comparison Algorithms

In this part, numerical results for the main problem are depicted to evaluate the performance of the algorithms using the Monte-Carlo method. We consider three network slices for eMBB, URLLC, and mMTC services. Assume we have six 4-antenna O-RU (MISO) located in a place with a diameter of 500 meters as shown in Fig. 2. In addition, we consider the users placed randomly in this area.

Here, the channel vector from the O-RU r to the UE i in service s is set as $\mathbf{h}^k_{r,u(s,i)} = d^{-\mathcal{L}}_{r,u(s,i)} \Omega^k_{r,u(s,i)}$, where $d^{-\mathcal{L}}_{r,u(s,i)}$ is the distance between the O-RU r and UE i in service s and $\mathcal{L} = 3.8$ is the path-loss exponent. Also, $\Omega^k_{r,u(s,i)}$ is the random variable that is generated by the Rayleigh distribution and it is the Rayleigh fading channel between the UE and O-RU. We consider 25 PRBs in the network. The packet size for mMTC is equal to 20 bytes, and for URLLC is equal to 32 bytes [39]. The maximum number of VNF for each slice is 25 and the mean arrival data rate for the eMBB service is $\lambda = 3Mbps$ and for the mMTC service and the URLLC service is $\lambda = 0.2Mbps$. Also, the quantization noise is assumed to be 10^{-13}. Moreover, we set $\mathfrak{h}_{u_{(s,i)}} = \eta_{u_{(s,i)}}/200$, $\mathfrak{m}_r = \zeta_r/10$ and $\mathfrak{q}^k_{r,u(s,i)} = P^{max}_s/100$. The other parameters of these simulations are depicted in Table II [39], [40], [41], [42].

Finding a feasible initial value is almost tricky. We use a fast method discussed in Section VII-B to overcome this challenge. Two different methods are used to compare the performance of the proposed method (IABV) and show the optimality of our approach. The first one is a baseline scheme, which uses random PRB allocation. Therefore, the allocation of PRB to each UE is random when we have low interference, but in figures with high interference, we randomly assign just one RB to each UE. Also, the association of O-RU is carried out based on distance. It means that each UE is assigned to the nearest O-RU. The optimal power is obtained using the CVX of MATLAB, which uses the successive convex approximation (SCA) method since the problem is convex. After

TABLE II
SIMULATION PARAMETERS

Parameter	Value
noise power	-174 dBm
bandwidth	180 KHz
maximum transmit power of each O-RU	40 dBm
maximum delay for eMBB	4 msec
maximum delay for URLLC	1 msec
maximum delay for mMTC	5 msec
maximum fronthaul capacity	46 bps
minimum data rate for eMBB	20 bps
minimum data rate for URLLC and mMTC	2 bps
maximum received power for mMTC	20 dBm
maximum received power for eMBB and URLLC	33 dBm

Algorithm 3 Fast Algorithm (FA) to Check Convergence

Set count = 0
Set $p^{k}_{r,u(s,i)} = 0$, $e^{k}_{r,u(s,i)} = 0$ and $g^{r}_{u(s,i)} = 0\ \forall r, k, s, i$
for $s \leftarrow 1$ to S **do**
  for $i \leftarrow 1$ to U_s **do**
    count = count +1
    $r^{*} = \arg\min_{r} d_{r,u(s,i)} \quad \forall r$
    $g^{r^{*}}_{u(s,i)} = 1$
    temp =mod(count, K)
    if temp=0 **then**
      $e^{K}_{r^{*},u(s,i)} = 1$
      Set $p^{K}_{r^{*},u(s,i)} = \min\{P^{\max}_{s}, P^{\max}_{r}/\mathfrak{N}\}$
    else: $e^{\text{temp}}_{r^{*},u(s,i)} = 1 \quad \forall r$
      Set $p^{\text{temp}}_{r^{*},u(s,i)} = \min\{P^{\max}_{s}, P^{\max}_{r}/\mathfrak{N}\}$
    end if
  end for
end for

achieving power and other parameters, the achievable rate will be obtained, and the optimal number of VNF is achieved from Lemma (1).

For the second one, we use the idea of the fixed BBU capacity and dynamic resource allocation (FBDR) algorithm proposed in [7] and named it the dynamic resource allocation scheme (DR scheme). We have services with different QoS in this work, similar to tenants with different QoS introduced in [7]. Therefore, we can use the DR scheme similar to the FBDR method adapted to our conditions for comparison. Instead of BBU in C-RAN, we have O-DU and O-CU in O-RAN. Since we do not talk about O-DU and O-CU capacity, we use the dynamic resource allocation scheme (DR scheme) algorithm and do not consider BBU capacity. In the DR scheme, PRB and power are dynamically allocated. The number of VNFs is obtained from the simulation. The UEs are associated with the O-RU based on the quality of their channels and the channel distance instead of using the greedy Algorithm 1 (GAA algorithm) for O-RU assignment. The figures in [7] show that dynamic BBU capacity and dynamic resource allocation (DBDR) perform better than FBDR for the same priority area. Notice that our proposed algorithm performs better than the DR scheme.

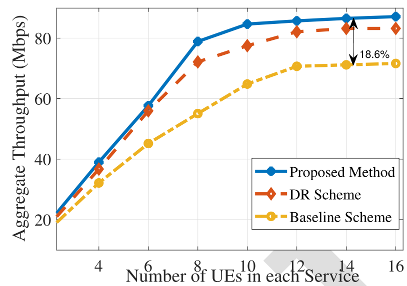

Fig. 3. Aggr. throughput vs. number of UEs in each service.

B. Feasible Region

Applying the correct initial point is crucial in making the system feasible and convergent. Therefore, we investigated the non-converging and converging simulation for models with fixed initial parameters and UE random channel gains. We experimentally found that in non-converged simulations, there are UEs at the edge of the boundaries or far away from the O-RU and have a weak channel gain. One solution is to eliminate UEs who undermine system convergence. For a large number of UEs with a fixed number of PRBs, the probability of having an infeasible solution increases due to a large number of UE interference. Another solution is to remove the simulations in the Monte-Carlo that do not converge using the fast algorithm (FA) to check the convergence before the proposed algorithm (IABV). Therefore, if more than half of the iterations have a feasible solution for the initial condition, the simulation can be displayed as a feasible model. If the conditions in (27), (26), (19d) and (19c) are met in the fast algorithm (FA), the given algorithm will converge. Assume, the number of UEs is $\mathfrak{N} = \sum_{s=1}^{S}\sum_{i=1}^{U_s} 1$, the number of PRBs is K, and the distance between the r^{th} O-RU to the UE i in slice s is $d_{r,u(s,i)}$. The FA algorithm is represented in Algorithm 3. The complexity order of this algorithm is $O(R \times \mathfrak{N})$ which is remarkably lower than the complexity order of the IABV method. In the FA algorithm, the O-RU association is based on the distance of the UE to the O-RU. Each UE is associated with the nearest O-RU. Also, the power of each UE is set to be the minimum of the maximum power of each UE and the maximum power of each O-RU divided by the total number of UEs ($\min\{P^{\max}_{s}, P^{\max}_{r}/\mathfrak{N}\}$). Moreover, the allocation of PRBs to UEs is based on dividing the number of UEs by the total number of PRBs.

C. Performance Results

In Fig. 3, the aggregate throughput is plotted versus the number of UEs in each service for these three methods. Suppose we have one service instance for each type of service, so we have three various services in this figure. Also, we have between 6 to 48 UEs in the system. Here, we did not consider the priority. The figure presented that the proposed method, IABV, is 18.6% higher throughput than the baseline scheme. As the number of UEs increases in each service, the aggregated throughput initially increases. Still, due to the interference and the power constraint, it will be saturated from 12 UEs in each service.

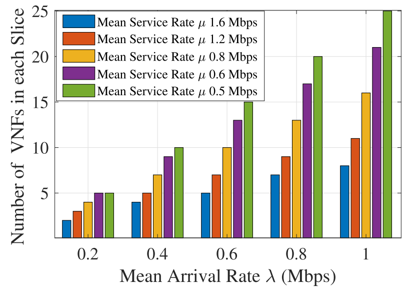

Fig. 4. Number of VNFs in each service vs. arrival rate.

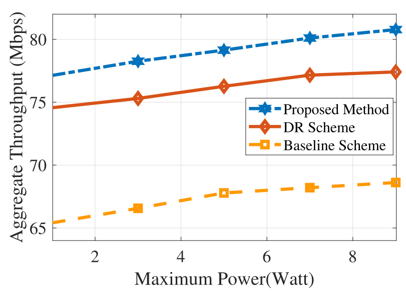

Fig. 5. Aggr. throughput for eMBB vs. maximum power.

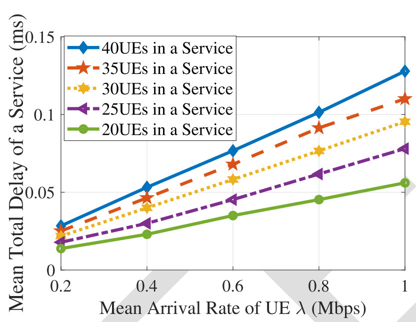

Fig. 6. Total Delay of a URLLC vs. the arrival rate of a UE.

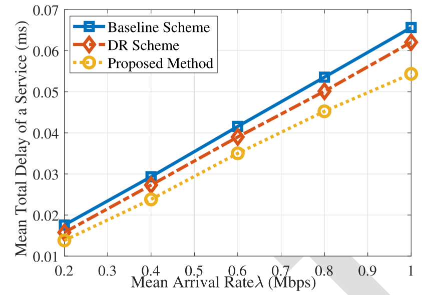

Fig. 7. Total Delay of a URLLC vs. arrival rate of a UE.

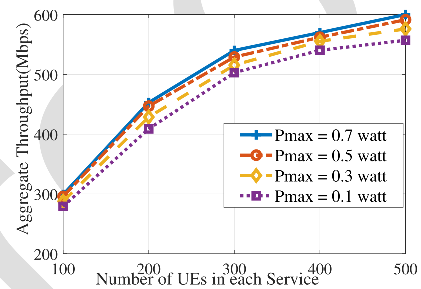

Fig. 8. Aggr. throughput vs. the number of mMTC UEs.

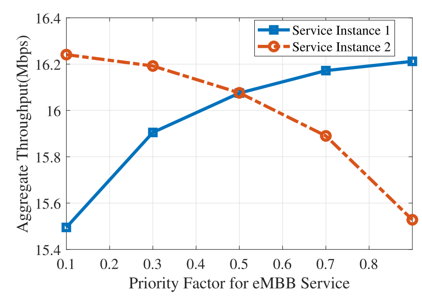

Fig. 9. Throughput of eMBB services vs. priority of the first one.

Fig. 4 depicts the number of activated VNFs for five different mean service times of one URLLC service vs. the mean arrival time for 12 UEs. This figure presents that as the mean arrival rate increases, the number of activated VNF increases. Moreover, the number of activated VNFs decreases when the mean service rate increases.

In Fig. 5, the aggregate throughput is depicted vs. the maximum power of UE for three different instances of eMBB service using proposed method (IABV), DR scheme and the baseline scheme. Here, we suppose that we have 12 UEs in each service. We assume that these three services require 5bits/sec/Hz, 10bits/sec/Hz, and 15bits/sec/Hz. As you can see in the figure, increasing the maximum power increases the aggregate throughput. Moreover, the proposed method (IABV), gives higher aggregate rates in compared to the DR scheme and the baseline scheme.

Fig. 6 illustrates the mean total delay of a UE in a URLLC service regarding the mean arrival rate of the UE and the number of UEs in the service for the proposed method (IABV). It is shown that the delay is an ascending function of the mean arrival rate (when the mean service time is fixed) and the number of UEs in the service. Moreover, we can see that the mean delay of a URLLC service does not reach the maximum threshold of the delay.

Fig. 7 is the same as Fig. 6 that presented the mean total delay of a UE in a URLLC service regarding the mean arrival rate of the UE for 20 UEs using three different methods. As you can see, the proposed method (IABV) outperforms the other scenarios.

Fig. 8, represents the aggregate throughput concerning the number of UEs in each service and the maximum power for three different mMTC service instances. Assume each UE in each mMTC service instance requires 0.1 bits/sec/Hz data rate and is not sensitive to the end-to-end delay. There is no restriction on fronthaul link capacity and the number of VNFs. The figure depicts that by increasing the number of UEs in each instance of the service, or by increasing the maximum power of each UE in each instance of mMTC service, the aggregate throughput increases.

Assume we have two types of eMBB service instances. In Fig. 9, the aggregate throughput (by considering the priority factor δ_s) is depicted for two eMBB service instances. Here we consider 4 UEs in each service. The Fig. 9 presented that by increasing the priority factor for one service instance,

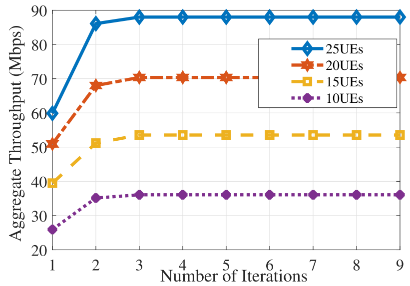

Fig. 10. Aggregate throughput vs. number of iterations.

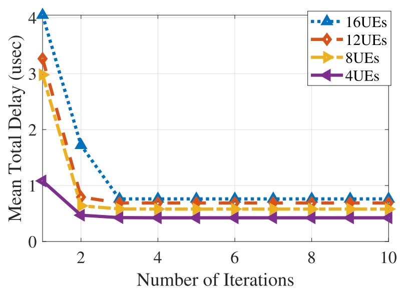

Fig. 11. Mean Delay time vs. number of iterations.

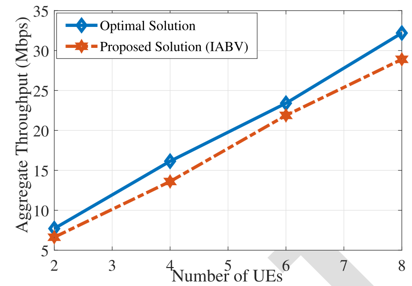

Fig. 12. Aggregate throughput vs. number of UEs.

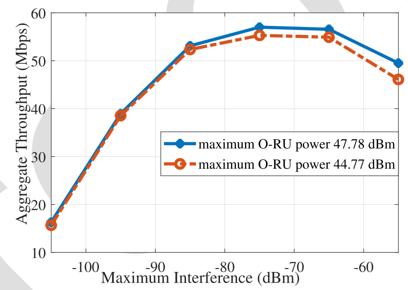

Fig. 13. Aggregate throughput vs. maximum interference.

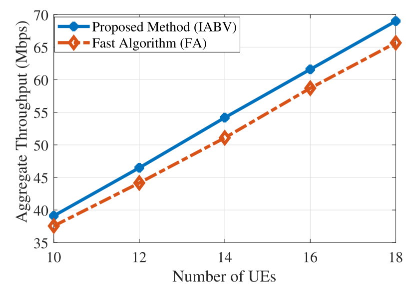

Fig. 14. Aggregate throughput vs. number of UEs.

more resources are allocated to this service instance, and the aggregate throughput of this service is increased and vice versa. Also, we can realize from this figure that the aggregate throughput has the most significant value at the same priority.

In Fig. 10, the aggregate throughput is shown according to the number of iteration (outer loop) of the proposed algorithm (IABV) for different numbers of UEs for one service. In this figure, the convergence of the IABV method is illustrated. The minimum data rate for each UE is assumed to be 2 Mbps. After four iterations, IABV converges to the fixed value.

In Fig. 11, the mean total delay of URLLC service is indicated according to the number of iterations of the proposed algorithm (IABV) for different numbers of UEs for one URLLC service. This figure shows that the algorithm converged to the fixed value after four iterations.

In Fig. 12, the aggregate throughput is shown according to the number of UEs for two different methods, namely the proposed algorithm (IABV) and the optimal method for URLLC service for the low interference. The minimum data rate is 5bits/sec/Hz for each UE and the maximum delay is 0.1ms. Also the mean arrival rate is set to be 0.2Mbps and the mean service rate is 0.5Mbps. The optimal approach is obtained from the two-step joint exhaustive search and using CVX. In each iteration in the first step, the PRB allocation and O-RU association are obtained from brute force, and in the second step, we use CVX to get optimal power. Our solution is close to the optimal value in a small number of UEs. Therefore, this figure numerically demonstrates that the proposed algorithm converges to the global optimal in low interference.

In Fig. 13, the aggregate throughput is depicted vs. the maximum interference for two different maximum power thresholds of O-RU. Here we assume that with the increase of every ten dBm of interference power, it is assumed that ten users have been added to the system. In -105 dBm, we have 5 UEs, and at the end, we have 55 UEs in the system. Since the amount of interference in the system is entered as a fixed value, the allocation of PRBs is not considered. The higher maximum power threshold leads to a greater aggregate throughput. The aggregate throughput first increases with the number of UEs and at the same time the amount of the maximum interference, then it becomes almost fixed and finally decreases so much. When the aggregate throughput decreases, the maximum interference is so high that it takes the system out of feasibility.

In Fig. 14, the aggregate throughput is shown versus the number of UEs for an eMBB service with low interference for the IABV and FA methods in the feasible region. The minimum data rate for each UE is 1Mb/s/Hz. The maximum power for each O-RU is 34dBm, and the maximum power for each UE is 30dBm. We assume that the system is not sensitive to fronhaul capacity and end-to-end delay and has enough VNF resources. By increasing the number of UEs, the aggregate throughput raises. And we can see that the IABV method is better than the FA method.

Table III shows the execution time given a number of UEs for one service for the three methods. We run our simulation on the system with configuration (RAM = 8 GB, CPU = Core

TABLE III
EXECUTION TIME VS. NUMBER OF UEs

Number of UEs	Execution Time (usec)		
	Proposed method	DR scheme	Baseline scheme
5	12.156	8.9546	6.6436
10	19.156	12.3112	8.7870
15	29.140	15.4778	9.5648
20	44.573	21.5342	14.8334
25	67.912	32.7926	21.5510

i5, SSD Hard Disk). As the number of UEs in the system increases, the execution time increases polynomially for all three algorithms. Since the baseline scheme is a simpler algorithm, with random PRB allocation and O-RU association based on distance, the execution time is less than the two other algorithms. Power and PRB are allocated in the DR scheme, but O-RUs are associated based on distance. Therefore the execution time is less than the proposed algorithm.

VIII. Conclusion

In this paper, we modeled the downlink of the O-RAN system using isolated network slicing for different 5G services, i.e., eMBB, mMTC, and URLLC. Our goal is to maximize aggregate throughput by activating VNFs in each service, associating RUs, allocating power, and PRBs. The limited fronthaul capacity and the mean delay for each service are considered. The problem is mixed-integer non-linear programming that is solved by a two-step iterative algorithm. In the first step, we reformulated the problem to achieve the number of activated VNFs as a function of data rate. Then, we obtained PRB association and power allocation using the Lagrangian method. In the second step, the O-RU association is carried out. The performance of our proposed method (i.e., IABV) is compared with the baseline scheme and DR scheme in [7]. In addition, the feasible region is discussed, and the FA algorithm is introduced to check the feasibility of the initial values. Also, we assume distinct scenarios for each service, i.e., eMBB, mMTC, and URLLC, based on their requirement QoS. Simulation results show that the proposed method (i.e., IABV) achieves 18.6% higher data rate than the baseline scheme. Moreover, simulation results illustrate more minor delays for the proposed method (IABV) than DR scheme and the baseline scheme.

References

AQ6 [1] X. Shen *et al.*, "AI-assisted network-slicing based next-generation wireless networks," *IEEE Open J. Veh. Technol.*, vol. 1, pp. 45–66, Jan. 2020. ISSN: 2644-1330

[2] M. Setayesh, S. Bahrami, and V. W. S. Wong, "Joint PRB and power allocation for slicing eMBB and URLLC services in 5G C-RAN," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Taipei, Taiwan, Dec. 2020, pp. 1–6.

[3] P. Popovski, K. F. Trillingsgaard, O. Simeone, and G. Durisi, "5G wireless network slicing for eMBB, URLLC, and mMTC: A communication-theoretic view," *IEEE Access*, vol. 6, pp. 55765–55779, 2018.

[4] A. Dogra, R. K. Jha, and S. Jain, "A survey on beyond 5G network with the advent of 6G: Architecture and emerging technologies," *IEEE Access*, vol. 9, pp. 67512–67547, 2021.

[5] M. Alsenwi, N. H. Tran, M. Bennis, S. R. Pandey, A. K. Bairagi, and C. S. Hong, "Intelligent resource slicing for eMBB and URLLC coexistence in 5G and beyond: A deep reinforcement learning based approach," *IEEE Trans. Wireless Commun.*, vol. 20, no. 7, pp. 4585–4600, Jul. 2021.

[6] S. Riolo, D. Panno, and L. Miuccio, "Modeling and analysis of tagged preamble transmissions in random access procedure for mMTC scenarios," *IEEE Trans. Wireless Commun.*, vol. 20, no. 7, pp. 4296–4312, Jul. 2021.

[7] Y. L. Lee, J. Loo, T. C. Chuah, and L.-C. Wang, "Dynamic network slicing for multitenant heterogeneous cloud radio access networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 4, pp. 2146–2161, Apr. 2018.

[8] J. Tang, W. P. Tay, T. Q. S. Quek, and B. Liang, "System cost minimization in cloud RAN with limited fronthaul capacity," *IEEE Trans. Wireless Commun.*, vol. 16, no. 5, pp. 3371–3384, May 2017.

[9] "O-RAN architecture description," O-RANAlliance, Alfter, Germany, Rep., 2020. 2022 v06.00 AQ7

[10] L. Gavrilovska, V. Rakovic, and D. Denkovski, "From cloud RAN to open RAN," *Wireless Pers. Commun.*, vol. 113, no. 3, pp. 1523–1539, Mar. 2020.

[11] N. Kazemifard and V.Shah-Mansouri, "Minimum delay function placement and resource allocation for Open RAN (O-RAN) 5G networks," *Comput. Netw.*, vol. 188, Apr. 2021, Art. no. 107809.

[12] L. Feng, Y. Zi, W. Li, F. Zhou, P. Yu, and M. Kadoch, "Dynamic resource allocation with RAN slicing and scheduling for uRLLC and eMBB hybrid services," *IEEE Access*, vol. 8, pp. 34538–34551, 2020.

[13] Y. Loong Lee, J. Loo, and T. C. Chuah, "A new network slicing framework for multi-tenant heterogeneous cloud radio access networks," in *Proc. Int. Conf. Adv. Electr. Electron. Syst. Eng. (ICAEES)*, Nov. 2016, pp. 414–420.

[14] H. Xiang, S. Yan, and M. Peng, "A realization of fog-RAN slicing via deep reinforcement learning," *IEEE Trans. Wireless Commun.*, vol. 19, no. 4, pp. 2515–2527, Apr. 2020.

[15] S. E. Elayoubi, S. B. Jemaa, Z. Altman, and A. Galindo-Serrano, "5G RAN slicing for verticals: Enablers and challenges," *IEEE Commun. Mag.*, vol. 57, no. 1, pp. 28–34, Jan. 2019.

[16] S. D'Oro, F. Restuccia, and T. Melodia, "Toward operator-to-waveform 5G radio access network slicing," *IEEE Commun. Mag.*, vol. 58, no. 4, pp. 18–23, Apr. 2020.

[17] A. K. Singh and K. K. Nguyen, "Joint selection of local trainers and resource allocation for federated learning in open RAN intelligent controllers," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, 2022, pp. 1874–1879.

[18] H. Zhang, H. Zhou, and M. Erol-Kantarci, "Federated deep reinforcement learning for resource allocation in O-RAN slicing," 2022, *arXiv:2208.01736*.

[19] P. Yang, X. Xi, T. Q. S. Quek, J. Chen, X. Cao, and D. Wu, "How should I orchestrate resources of my slices for bursty URLLC service provision?" *IEEE Trans. Commun.*, vol. 69, no. 2, pp. 1134–1146, Feb. 2021.

[20] F. Saggese, M. Moretti, P. Popovski, "Power minimization of downlink spectrum slicing for eMBB and URLLC users," Jun. 2021, *arXiv:2106.08847*.

[21] K. Guo, M. Sheng, J. Tang, T. Q. S. Quek, and Z. Qiu, "Exploiting hybrid clustering and computation provisioning for green C-RAN," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 12, pp. 4063–4076, Dec. 2016.

[22] P. Luong, F. Gagnon, C. Despins, and L.-N. Tran, "Joint virtual computing and radio resource allocation in limited fronthaul green C-RANs," *IEEE Trans. Wireless Commun.*, vol. 17, no. 4, pp. 2602–2617, Apr. 2018.

[23] P. Luong, F. Gagnon, C. Despins, and L.-N. Tran, "A novel energy-efficient resource allocation approach in limited fronthaul virtualized C-RANs," in *Proc. IEEE 87th Veh. Technol. Conf. (VTC Spring)*, Jun. 2018, pp. 1–6.

[24] S. Ali, A. Khan, and A. Ahmad, "Energy-efficient resource allocation and RRH association in multitier 5G H-CRANs," *Trans. Emerg. Telecommun. Technol.*, vol. 30, no. 1, Jan. 2019, Art. no. e3521.

[25] S. Ali, A. Khan, and A. Ahmad, "Joint RRH-association, sub-channel assignment and power allocation in multi-tier 5G C-RANs," *IEEE Access*, vol. 6, pp. 34393–34402, 2018.

[26] N. Amani, S. Parsaeefard, H. Taheri, and H. Pedram, "Power-efficient resource allocation in massive MIMO aided cloud RANs," Aug. 2019, *arXiv:1908.07568*.

[27] B. Han *et al.*, "Research on resource migration based on novel RRH-BBU mapping in cloud radio access network for HSR scenarios," *IEEE Access*, vol. 7, pp. 108542–108550, 2019.

[28] R. Mijumbi, J. Serrat, J.-L. Gorricho, N. Bouten, F. De Turck, and R. Boutaba, "Network function virtualization: State-of-the-art and research challenges," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 1, pp. 236–262, 1st Quart., 2016.

[29] S. Niknam *et al.*, "Intelligent O-RAN for beyond 5G and 6G wireless networks," May 2020, *arXiv:2005.08374*.

[30] C. B. Both *et al.*, "System intelligence for UAV-based mission critical with challenging 5G/b5G connectivity," Feb. 2021, *arXiv:2102.02318*.

[31] "O-RAN Working Group 2: AI/ML workflow description and requirements," O-RAN Alliance, Alfter, Germany, Rep. O-RAN.WG2.AIML-v01.02.02, Mar. 2020.

[32] B. S. Lin, "Toward an AI-Enabled O-RAN-based and SDN/NFV-driven 5G and IoT network era," *Netw. Commun. Technol.*, vol. 6, no. 1, pp. 6–15, Jun. 2021.

[33] J. Mei, X. Wang, K. Zheng, G. Boudreau, A. B. Sediq, and H. Abou-Zeid, "Intelligent radio access network slicing for service provisioning in 6G: A hierarchical deep reinforcement learning approach," *IEEE Trans. Commun.*, vol. 69, no. 9, pp. 6063–6078, Sep. 2021.

[34] S. Huang, H. Yin, J. Wu, and V. C. M. Leung, "User selection for multiuser MIMO downlink with zero-forcing beamforming," *IEEE Trans. Veh. Technol.*, vol. 62, no. 7, pp. 3084–3097, Sep. 2013.

[35] J. Cavazos. "5G testing: What is O-RAN? Part 2." 2020. [Online]. Available: https://blogs.keysight.com/blogs/inds.entry.html

[36] D. Marabissi and R.Fantacci, "Highly flexible RAN slicing approach to manage isolation, priority, efficiency," *IEEE Access*, vol. 7, pp. 97130–97142, 2019.

[37] *5G Management and Orchestration; Concepts, Use Cases and Requirements (3GPP TS 28.530 Version 15.0.0, Release 15)*, ETSI Standard TS-128-530-V15.0.0, Oct. 2018.

[38] Y. Akçay, H. Li, and S. H. Xu, "Greedy algorithm for the general multidimensional knapsack problem," *Ann. Oper. Res.*, vol. 150, no. 1, pp. 17–29, Dec. 2007.

[39] "5G; study on scenarios and requirements for next generation access technologies (3GPP TR 38.913 version 14.3.0 release 14)," ETSI, Sophia Antipolis, France, Rep. TR-138-913-V14.3.0, Oct. 2017.

[40] *Evolved Universal Terrestrial Radio Access (E-UTRA); Base Station (BS) Radio Transmission and Reception (Release=13)*, 3GPP-TS-36.104-V13.3.0, Mar. 2016.

[41] *Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Frequency (RF) Requirements for LTE Pico Node B (Release 13)*, 3GPP, Sophia Antipolis, France, Rep. TR-36.931-V13.0.0, Jan. 2016.

[42] E. Mohyeldin, "Minimum technical performance requirements for IMT radio interface(s)," ITU, Geneva, Switzerland, Rep. ITU-R M.2410-0, 2020.