Resource Allocation in an Open RAN System using Network Slicing

Mojdeh Karbalaee Motalleb*, Vahid Shah-Mansouri*, and Saeede Parsaeefard†

Email: {mojdeh.karbalaee, vmansouri}@ut.ac.ir, {saeede.parsaeefard}@gmail.com

*School of ECE, University of Tehran, Tehran, Iran, †University of Toronto, Toronto, Canada

Abstract—**Taking advantage of virtual radio access network (v-RAN) and Cloud RAN (C-RAN), Open RAN (O-RAN) is introduced as the next generation of RAN systems. O-RAN leads to increased flexibility, openness, reduces operational costs, and agile deployment. O-RAN separates RAN into three different units, namely Radio Unit (O-RU), Distributed Unit (O-DU), and Central Unit (O-CU). In this paper, we study the problem of service-aware baseband resource allocation and virtual network function (VNF) activation in O-RAN systems using network slicing to isolate different types of services. The limited fronthaul capacity and the restriction of end-to-end delay are considered at the same time. The optimization of baseband resources includes O-RU assignment, physical resource block (PRB), and power allocation. The main problem is a mixed-integer non-linear programming problem that is non-trivial to solve numerically. Nevertheless, we break it down into two different steps where an iterative algorithm 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 achieved. The proposed method is validated via simulations that illustrate 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.**

I. INTRODUCTION

A. State Of The Arts

To achieve the desired level of QoS (quality, delay, power, etc.) and isolate different services, network slicing is the most effective solution for 5G wireless cellular networks. There are several implementations of network slicing, including core slicing, RAN slicing, and both. Recent discussion of 5G wireless systems has been around three services, namely enhanced mobile broadband (eMBB), ultra-reliable low latency communications (URLLC), and massive machine-to-machine communications (mMTC). Depending on QoS, each service requires its own slice of the network. The eMBB service fulfills the demand for high capacity and throughput. The URLLC also offers autonomous vehicles, tactile internet, remote surgery services, and other services of high quality and low latency. mMTC is comprised of many low-power internet of things (IoT) devices that transmit small payloads to a typical receiver, including sensors for sensing, metering, and monitoring [1]–[6].

This paper considers various user equipments (UEs) requesting one of these types of services with different QoS. We want to study RAN slicing technology and the isolation between slices in this system to evaluate the performance of various services.

O-RAN is driven by two pillars of openness and RAN intelligence as the next generation of radio access networks. O-RAN was developed as a result of integrating virtual RAN (vRAN) and cloud RAN (CRAN), and it takes advantage of both. With RAN virtualization, operators can improve flexibility, reduce CAPEX and OPEX, and add new capabilities to the network more quickly. The CRAN divides the radio remote head (RRH) and baseband unit (BBU) into two major parts. Several distributed RRHs can be connected to a centralized BBU, called BBU-pool [7]. Unlike CRAN, O-RAN separates RAN into three different units, namely Radio Unit (O-RU), Distributed Unit (O-DU), and Central Unit (O-CU).

O-RU is a logical node that contains RF and lowers PHY. Moreover, the O-DU expresses another logical node with higher PHY, MAC, and RLC. In addition, the O-CU contains two parts: the O-CU user plane (O-CU-UP) and the O-CU control plane (O-CU-CP). O-CU-UP hosts PDCP-UP and SDAP, and O-CU-CP hosts PDCP-CP and RRC. O-DU and O-CU are connected via an open and well-defined interface F_1. Moreover, O-DU is connected to a radio unit (O-RU) with an open fronthaul interface. The architecture of O-RAN contains other principal logical nodes called Orchestration and Automation, RAN Intelligent Controller (RIC)- Near Real-Time, and O-Cloud. O-RAN offers machine learning techniques based on the requirement for an intelligent wireless network for the next generation of RAN systems. The two logical nodes, RIC-Non Real-Time (which is placed in Orchestration and Automation node) and RIC-Near Real-Time, implement the algorithms for network intelligence [8]–[14].

Network function virtualization (NFV) is referred to the separation of network software and hardware elements such that network function can run on commodity hardware. Virtual network functions (VNF) are system function blocks in NFV systems. The NFV technology offers to execute VNFs as virtual machines or containers on a cloud environment [15], [16]. As a result, some O-RAN components, such as user plane function (UPF), O-CU, O-DU, and RAN Intelligent Controller (RIC)-near real-time, are being virtualized and implemented as VNFs that can be run as virtual machines (VMs) or containers.

In order to achieve the specific QoS for eMBB, URLLC, and mMTC, this study optimizes the allocation of baseband

resources, including power allocation, PRB allocation, O-Rus association, and VNF activation to create an isolated network slicing outline.

B. Related Literature

The problem of resource allocation for network slicing in multi-tenant cellular networks has received attention recently [17]–[19]. In [18], dynamic network slicing in multi-tenant heterogeneous CRAN (H-CRAN) is considered. The network slicing scheme includes a higher level and the lower level. The higher level manages user acceptance control, user communication that provides for radio unit association (RRH association to maximize user rates and allocate baseband resource capacity), the allocation of BBU capacity. Also, the allocation of power and physical resource blocks (PRB) is performed in the lower level. In [20], network slicing in the radio section is considered for the fog RAN (F-RAN) system, and a deep reinforcement learning algorithm is proposed for it. In [21], [22], the implementation of RAN level slicing is discussed in mobile network operator (MNO).

Multiplexing eMBB and URLLC services on the same RAN and sharing the resources of these services is challenging, and many researchers pay attention to this issue. In [2], [23], [24], the problem of resource allocation in the coexistence of URLLC and eMBB services is considered based on their QoS. In [6], the problem of resource allocation for joint eMBB and URLLC services is formulated and solved by deep reinforcement learning. In [25], the authors proposed to allocate RAN resources for the network slicing system in coexistence of eMBB and URLLC services. The system guarantees the latency, the service rate, and the maintenance of reliability.

In [26], [27], the authors address the issue of beamforming and VMs activation (using virtualization) in a C-RAN system with limited fronthaul capacity. This paper aims to minimize the energy cost with the system delay, fronthaul capacity, and rate constraint. Also, transmission and processing delays are modeled based on M/M/1 queueing theory to guarantee UE delays. In [28], [29], 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 [30]–[32], the problem of joint power allocation and RRH association in the H-CRAN system is considered to maximize energy efficiency. In [33], optimum power is obtained in massive MIMO aided C-RAN. Also the problem of RRH to BBU and the RRH to UE association is obtained. Moreover, the author talked about the feasible region to make the initial values possible.

C. Main Contributions

The purpose of this study is to optimize baseband resource allocation, i.e., power allocation, PRB allocation, O-RUs association, and VNF activation, to develop an isolated network slicing outline for different types of services in an O-RAN platform. Unlike other papers, we concentrate more on the multiservice resource management of the RAN slicing in the openness and flexible O-RAN system. The novelty of this paper is to improve the resource utilization of the overall wireless O-RAN system in the presence of the three generic service types introduced in 5G using RAN slicing. In this paper, as depicted in Figure 1, the downlink of the ORAN system is studied. The main contributions of this paper are summarized as follows:

- This paper depicts a network slicing model for three different services introduced in 5G, i.e., eMBB, mMTC, and URLLC. The problem of radio resource allocation and VNF activation is studied in this paper in the O-RAN architecture. We formulate the problem of baseband resource allocation to maximize the weighted throughput of the O-RAN system for different types of services with specific QoS.
- The desired QoS conditions such as delay and throughput are considered for different types of services. We formulate the end-to-end processing delay of the system based on the number of activated VNFs and user throughput. Assuming a limited fronthaul capacity, we calculate the power and capacity of each O-RU. We model the interference of neighboring O-RUs and formulate the actual throughput for different service types. Based on the short packet transmission of URLLC and mMTC, we also model their throughput.
- The main problem is mixed-integer non-linear programming that is extremely difficult to solve. We perform a two-step iterative algorithm to solve it. The two-step iterative algorithm is presented for the resource management framework with the first-step VNF activation, power allocation, PRB association, and the second-step O-RU association.
- We reformulated and simplified the main problem for the first step to find an upper and lower bound for the number of activated VNFs and use lagrangian function and KKT conditions to find optimal power and PRB allocation. For the second step, the problem of O-RU association can be converted to a multiple knapsack problem and solved by the Greedy algorithm.
- We talk about the initial point and the feasible region for the numerical results and introduce a fast algorithm that is less complex than our method to realize the feasible region for our problem.

The rest of this paper is organized as follows. The system model and the problem formulation are described in Section II. The details of our proposed resource management algorithm are introduced in Section III. In Section IV, numerical results are provided to evaluate the performance of the proposed algorithm. Section V concludes the paper.

II. SYSTEM MODEL AND PROBLEM FORMULATION

In this section, we describe the downlink system in O-RAN using RAN slicing as depicted in Figure 1. Firstly, we present the system model. Then, we obtain achievable data rates, power of O-RU, and the fronthaul capacity for the downlink (DL) of the ORAN system. Afterward, we
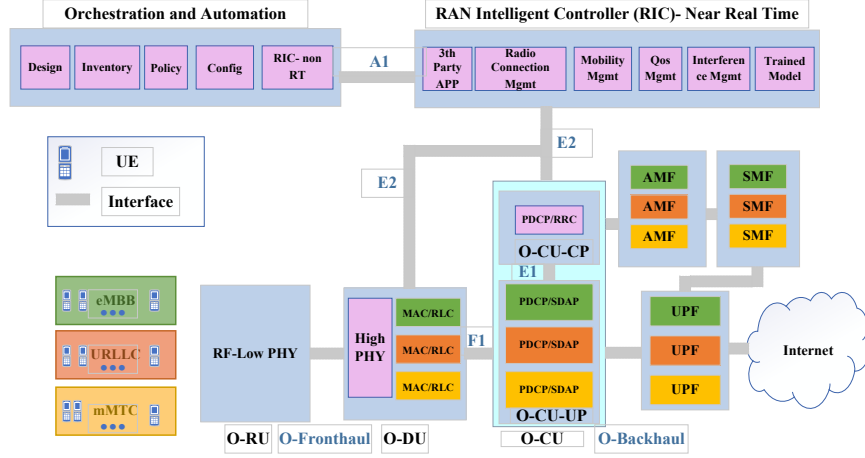
Fig. 1: Network sliced ORAN system

discuss the mean delay and the power of VNFs. Finally, the main problem is expressed.

A. System Model

There are three service types including eMBB, URLLC, and mMTC which support different applications. There are S_1, S_2 and S_3 different applications for the eMBB, URLLC, and mMTC service types, respectively and we have $S = S_1 + S_2 + S_3$. Therefore, there are S pre-allocated slices serving these S services; Each pre-allocated slice contains reserved VNFs for the three logical nodes:

- The MAC/RLC functions in the O-DU logical node
- The PDCP/SDAP functions in the O-CU-UP logical node
- The UPF logical node

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). So, each service request $s \in \{1, \dots, S\}$ is served by its corresponding slice. So we have $\{1, 2, ..., S_1\}$ set of eMBB service instances, $\{1, 2, ..., S_2\}$ set of URLLC service instances, and $\{1, 2, ..., S_3\}$ set of mMTC service instances. Each Service $s_j \in \{1, 2, ..., S_j\}$ consists of U_{s_j} requests from single-antenna UEs which require certain level of QoS. We notice that $j \in \{1, 2, 3\}$ indicates the service type. There are different application requests which fall into one of these service categories. Each application request requires a specific QoS. Based on the application and QoS request, UE may be admitted and allocated to the resources. Each slice $s_j \in \{1, 2, ..., S_j\}$, $j \in \{1, 2, 3\}$ 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, defined as the Service Function Chain (SFC) in the O-RAN system.

Assume we have K physical resource blocks (PRBs) in this system. All K PRBs can be assigned to all UEs in each service. Also, each VNF instance runs on a virtual machine (VM) that uses resources from the data centers. In addition,

there are R multi-antenna O-RUs that are shared between slices. O-RU $r \in \{1, 2, ..., R\}$ has J antennas for transmitting and receiving data. Also $\mathcal{R} = \{r \mid r \in 1, 2, ..., R\}$ depicts the set of O-RUs. Moreover, all O-RUs have access to all PRBs.

B. The Achievable Rate

The SNR of the ith UE served at slice s on PRB k is obtained from

$$\rho_{r,u(s,i)}^k = \frac{|p_{r,u(s,i)}^k \mathbf{h}_{r,u(s,i)}^{H\,k} \mathbf{w}_{r,u(s,i)}^k|^2}{BN_0 + I_{r,u(s,i)}^k}, \qquad (1)$$

where $p_{r,u(s,i)}^k$ represents the transmission power from O-RU r to the ith UE served at slice s on PRB k. $\mathbf{h}_{r,u(s,i)}^k \in \mathbb{C}^J$ is the vector of channel gain of a wireless link from r^{th} O-RU to the ith UE in s^{th} slice. In addition, $\mathbf{w}_{r,u(s,i)}^k \in \mathbb{C}^J$ depicts the transmit beamforming vector from r^{th} O-RU to the ith UE in s^{th} slice that is the zero forcing beamforming vector to minimize the interference which is indicated as below

$$\mathbf{w}_{r,u(s,i)}^k = \hat{\mathbf{h}}_{r,u(s,i)}^k (\hat{\mathbf{h}}_{r,u(s,i)}^{H\,k} \hat{\mathbf{h}}_{r,u(s,i)}^k)^{-1} \qquad (2)$$

The channel state information is imperfect. So, the channel gain is known with errors; the imperfection of channel estimation is shown as $\hat{\boldsymbol{h}}_{r,u(s,i)} = \boldsymbol{h}_{r,u(s,i)} + \Delta\boldsymbol{h}_{r,u(s,i)}$. $\Delta\boldsymbol{h}_{r,u(s,i)}$ indicates the estimating error vector with a Gaussian distribution of $\Delta\boldsymbol{h}_{r,u(s,i)} \backsim \mathcal{N}(0, \boldsymbol{\phi}_{r,u(s,i)}^2)$. where $\boldsymbol{\phi}_{r,u(s,i)} = \mathrm{diag}(\phi_{r,u(s,i)}, \dots, \phi_{r,u(s,i)})$.

Moreover, $g_{u(s,i)}^r \in \{0, 1\}$ is the binary variable that illustrates whether O-RU r served the ith UE that is allocated to s^{th} slice or not. Also, BN_0 denotes the power of Gaussian additive noise. UEs in different O-RUs can be assigned to different PRBs. As a result, a UE in O-RU r using PRB k may experience interference from other O-RUs in the set of $r' \in R - r$. Network Slicing methods significantly reduce inter-service interference but do not eliminate it at all. There are some techniques to minimize inter-slice interference. One of these techniques is to have two-time scale scheduling. The PRB scheduling to the slices
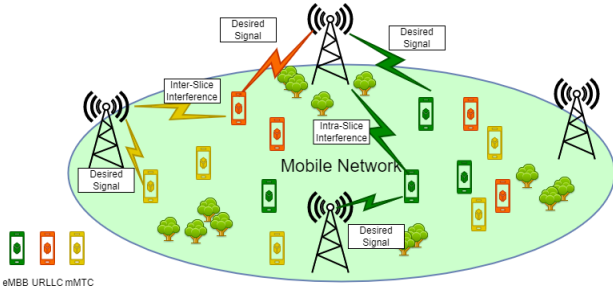
Fig. 2: Type of Interference Signal

Fig. 3: Precoding and Quantization of Signal

is performed on the first time scale, and in the second time scale, the PRB scheduling to the UEs of slices is carried out. Since there are limited resources, inter-service interference cannot be eliminated entirely. The other method is to allocate part of the RB of eMBB services to URLLC and mMTC [2], [6], [34]. We have two types of interference, the first one is between slices (inter-slice) and the second one is between UEs in each slices or services (intra-slice) that is shown in figure 2. $I_{r,u(s,i)}^k$ is the sum of the power of interfering signals and quantization noise represented as follow

$$I_{r,u(s,i)}^k = \tag{3a}$$

$$\underbrace{\sum_{\substack{l=1 \\ l \neq i}}^{U_s} e_{u(s,i)}^k e_{u(s,l)}^k p_{u(s,l)}^k \sum_{\substack{r'=1 \\ r' \neq r}}^{R} |\mathbf{h}_{r',u(s,i)}^{H\,k} \mathbf{w}_{r',u(s,l)}^k g_{u(s,l)}^{r'}|^2 +}_{\text{(intra-slice interference)}} \tag{3b}$$

$$\underbrace{\sum_{\substack{n=1 \\ n \neq s}}^{S} \sum_{l=1}^{U_s} e_{u(s,i)}^k e_{u(n,l)}^k p_{u(n,l)}^k \sum_{\substack{r'=1 \\ r' \neq r}}^{R} |\mathbf{h}_{r',u(s,i)}^{H\,k} \mathbf{w}_{r',u(n,l)}^k g_{u(n,l)}^{r'}|^2}_{\text{(inter-slice interference)}} \tag{3c}$$

$$\underbrace{+ \sum_{j=1}^{R} \sigma_q^2 |\boldsymbol{h}_{r,u(s,i)}^k|^2,}_{\text{(quantization noise)}} \tag{3d}$$

where $e_{u(s,i)}^k$ is the binary variable to show whether the kth PRB is allocated to the UE i in slice s, assigned to r^{th} O-RU. Here we have two Gaussian noise types: additive Gaussian noise and the other is Gaussian quantization noise. The second noise is added to the interfering signal and shown with $I_{r,u(s,i)}^k$ and it is different with interference. To obtain SNR as formulated in (1), let $y_{u(s,i)}$ be the received signal of UE i in s^{th} service formulated as

$$y_{u(s,i)} = \sum_{r=1}^{R} \sum_{k=1}^{K_s} \boldsymbol{h}_{r,u(s,i)}^{H\,k} g_{u(s,i)}^r e_{r,u(s,i)}^k x_{Q\,r,u(s,i)}^k + z_{u(s,i)}, \tag{4}$$

where $x_{Q\,r,u(s,i)}^k = x_{P\,r,u(s,i)}^k + \boldsymbol{q}_r$. Also, $x_{P\,r,u(s,i)}^k = \boldsymbol{w}_{r,u(s,i)}^k p_{r,u(s,i)}^{k\,\frac{1}{2}} x_{u(s,i)}$, and $x_{u(s,i)}$ depicts the transmitted symbol vector of UE i in s^{th} set of service, $z_{u(s,i)}$ is the additive Gaussian noise $z_{u(s,i)} \backsim \mathcal{N}(0, N_0)$ and N_0
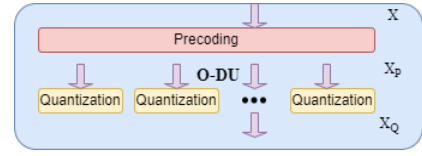
is the noise power. Moreover, x_P denotes the precoded message before compression, and x_Q illustrates the pre-coded message after compression that is shown in figure 3. In addition, $\boldsymbol{q}_r \in \mathbb{C}^J$ indicates the quantization Gaussian noise ($\boldsymbol{q}_r \backsim \mathcal{N}(0, \sigma_q^2 \boldsymbol{I_R})$), which is made from signal compression in O-DU. The achievable data rate for the ith UE request in the s_1^{th} application of service type 1 (eMBB) can be written as $\mathcal{R}_{u(s_1,i)}$ that is formulated as

$$\mathcal{R}_{r,u(s_1,i)}^k = B \log_2(1 + \rho_{r,u(s_1,i)}^k),$$

$$\mathcal{R}_{u(s_1,i)}^r = \sum_{k=1}^{K} B \log_2(1 + \rho_{r,u(s_1,i)}^k e_{r,u(s_1,i)}^k),$$

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

where B is the bandwidth of system. $\mathcal{R}_{u(s_1,i)}^r$ is the achievable rate of RU r to UE i in slice s_1. Since the blocklength in URLLC and mMTC is finite, the achievable data rate for the ith UE request in the s_j^{th} ($j \in \{2,3\}$) application of service type 2 (URLLC) and 3 (mMTC) is not achieved from Shannon Capacity formula. So, for the short packet transmission, the achievable data rate is approximated from following [2],

$$\mathcal{R}_{r,u(s_j,i)}^k = B \log_2(1 + \rho_{r,u(s_j,i)}^k - \zeta_{u(s_j,i)}^k) e_{u(s_j,i)}^k, \tag{6a}$$

$$\mathcal{R}_{u(s_j,i)}^r = \sum_{k=1}^{K} B(\log_2(1 + \rho_{u(s_2,i)}^k) - \zeta_{u(s_j,i)}^k) e_{u(s_j,i)}^k, \tag{6b}$$

$$\mathcal{R}_{u(s_j,i)} = \sum_{r=1}^{R} \mathcal{R}_{u(s_j,i)}^r g_{u(s_j,i)}^r, \tag{6c}$$

where

$$\zeta_{u(s_j,i)}^k = \log_2(e) Q^{-1}(\epsilon) \sqrt{\frac{\mathfrak{C}_{u(s_j,i)}^k}{N_{u(s_j,i)}^k}}, \tag{7}$$

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

If we replace $p_{u(s,l)}^k$ and $p_{u(n,l)}^k$ in (3) by P_s^{max}, an upper bound $\bar{I}_{r,u(s,i)}^k$ is obtained for $I_{r,u(s,i)}^k$. Therefore, $\bar{\mathcal{R}}_{u(s,i)} \forall s, i$ is derived by using $\bar{I}_{r,u(s,i)}^k$ instead of $I_{r,u(s,i)}^k$ in (6) and (5).

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

Let P_r denote the power of the transmitted signal from the rth O-RU to all the UEs served by it. From (4), the power of each O-RU r is obtained as follows,

$$P_r = \sum_{s=1}^{S} \sum_{k=1}^{K_s} \sum_{i=1}^{U_s} |\mathbf{w}_{r,u(s,i)}^k|^2 p_{r,u(s,i)}^k g_{u(s,i)}^r e_{r,u(s,i)}^k + \sigma_q^2. \tag{8}$$

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 rth O-RU is formulated as

$$C_r = \log\left(1 + \frac{\sum_{s=1}^{S} \sum_{k=1}^{K_s} \sum_{i=1}^{U_s} |\mathbf{w}_{r,u(s,i)}^k|^2 \alpha_{r,u(s,i)}^k}{\sigma_q^2}\right), \tag{9}$$

where $\alpha_{r,u(s,i)}^k = p_{r,u(s,i)}^k g_{u(s,i)}^r e_{r,u(s,i)}^k$ and σ_q^2 is the power of quantization noise.

D. Mean Delay

In this part, the end-to-end mean delay for each service is obtained. Suppose the mean total delay is depicted as T_{tot},

$$T^{\text{tot}} = T^{\text{proc}} + T^{tr} + T^{\text{pro}}, \tag{10a}$$

$$T^{\text{proc}} = T^{RU} + T^{DU} + T^{CU} + T^{UPF}, \tag{10b}$$

$$T^{tr} = T^{fr,t} + T^{\text{mid},t} + T^{b,t}, \tag{10c}$$

$$T^{\text{pro}} = T^{fr,p} + T^{\text{mid},p} + T^{b,p}. \tag{10d}$$

The total delay (T^{tot}), is the sum of the processing delay (T^{proc}), the transmission delay (T^{tr}), and the propagation delay (T^{pro}). The propagation delay is the time takes for a signal to reach its destination. It obtains based on the length of the fiber link and the capacity of the link ($T = L/c$, where L is the length of the link and c is the propagation speed of the link). The total propagation delay (T^{pro}), is the sum of the propagation delay in the fronthaul link $T^{fr,p}$, the midhaul link $T^{\text{mid},p}$, and the backhaul link $T^{b,p}$. Also, the transmission delay is the amount of time required to push all the packets into the fiber link. Moreover, it can formulated as $T = \frac{\alpha}{R}$, where R is the data-rate of the packet transmission in each link and α is the mean packet size transmitting in each link. So the total transmission delay (T^{tr}) is the sum of the transmission delay in the fronthaul $T^{fr,t}$, the midhaul $T^{\text{mid},t}$, and the backhaul $T^{b,t}$.

Considering propagation and transmission delays is straightforward to add to the system, but we did not consider them for better presentation and focus on obtaining the optimal number of VNF. In this paper, we focus only on the processing delays to find the optimal number of VNFs, and we consider the other two delays are fixed for simplification; The propagation delay is considered when each O-RU can connect to the number of O-DUs and can select to connect to which O-DU based on the route, capacity, and priority (Also, the connections of O-DUs to O-CUs is the same as O-RU to O-DU). Therefore, different routes are added to the system to solve the routing problem, but we do not consider it since it changes the model and does not fit in this model. So, the connection between O-RUs to O-DUs

is fixed and transparent. As a result, the propagation delay is fixed and does not affect the optimization problem. The following is a brief calculation of propagation delay. Since the distance between the O-RU to O-DU is about 10 km and also the distance between O-DU and O-CU is about 80 km. Moreover, the distance from O-CU to the network should not exceed 200 km [35]. so, the propagation delay is about $T^{\text{pro}} = (10 + 80 + 200) \times 10^3/(3 \times 10^8) < 1ms$. Since fronhaul, midhaul, and backhaul are fiber optics, c is the speed of light. Also, due to the edge technique in O-DU or O-CU for users with low latency, this amount of latency is greatly reduced. But we also do not focus on edge processing in this paper. The following is a brief calculation of transmission delay to show that it does not affect the optimization since it has such a small amount. In URLLC and mMTC, the mean packet size can be between 20 to 32 byte; Also, the minimum data rate is assume to be $46 bits/sec/Hz \times BW(180KHz)$. So the transmission delay from O-RU to O-DU is about $T^{fr,t} = \frac{20 \times 8}{46 \times 180 \times 10^3} < 2us$. As a result, the $T^{fr,t} \approx T^{\text{mid},t} \approx T^{b,t}$. for eMBB, the packet size can be 100 times larger and the delay is not exceed the 0.6ms.

So, we assume that the total delay is approximate to the processing delay ($T^{\text{tot}} \approx T^{\text{proc}}$).

1) Processing Delay: Assume the packet arrival of UEs follows a Poisson process with arrival rate $\lambda_{u(s,i)}$ for the ith UE of the sth service (or slice). Therefore, the mean arrival data rate of the sth slice in the UPF layer is $\alpha_s^U = \sum_{u=1}^{U_s} \lambda_{u(s,i)}$. Assume the mean arrival data rate of the UPF layer for slice s (α_s^U) is approximately equal to the mean arrival data rate of the O-CU-UP layer (α_s^C) and the O-DU (α_s^D). so $\alpha_s = \alpha_s^U \approx \alpha_s^C \approx \alpha_s^D$, Because the amount of data traffic transferred along the route (regardless of frame changes) is constant. Since, by using Burkes 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 the incoming traffic to VNFs equally. Suppose the baseband processing of each VNF is depicted as M/M/1 processing queue. Each packet is processed by one of the VNFs of a slice. So, the mean delay for the sth slice in the O-DU, the O-CU, and the UPF is modeled as M/M/1 queue, is formulated as follows, respectively [26], [28], [29],

$$T_s^{DU} = \frac{1}{\mu_s^d - \alpha_s/M_s^d},$$

$$T_s^{CU} = \frac{1}{\mu_s^c - \alpha_s/M_s^c}, \tag{11}$$

$$T_s^{UPF} = \frac{1}{\mu_s^u - \alpha_s/M_s^u},$$

where M_s^d, M_s^c and M_s^u are the variables that depict the number of VNFs in O-DU, O-CU-UP and UPF, respectively. Moreover, $1/\mu_s^d$, $1/\mu_s^c$, and $1/\mu_s^u$ are the mean service time of the O-DU, O-CU, and the UPF layers, respectively. Besides, α_s is the arrival rate which is divided by load balancer before arriving to the VNFs. The arrival

rate of each VNF in each layer for each slice s is α_s/M_s^i $i \in \{d, c, u\}$.

$T_{u(s,i)}^{RU}$ is the mean transmission delay of the ith UE of the sth service on the wireless link. The arrival data rate of wireless link for each UE i of service s is $\lambda_{u(s,i)}$ As a result, we have $\sum_{i=1}^{U_s} \lambda_{u(s,i)} = \alpha_s$. 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 [26], [28], [29].

Therefore, the mean delay of the transmission layer for UE i in slice s is

$$T_{u(s,i)}^{RU} = \frac{1}{R_{u(s,i)} - \lambda_{u(s,i)}}. \quad (12)$$

So, the mean processing delay for UE i in slice s is $T_{u(s,i)}^{\text{proc}} = T_{u(s,i)}^{RU} + T_s^{DU} + T_s^{CU} + T_s^{UPF}$. Hence, for the simplification and focusing on the processing delay, we assume $T_{u(s,i)}^{\text{tot}} \approx T_{u(s,i)}^{\text{proc}}$.

The processing delay can be considered an M/M/1; since the number of arrival packets in the system is enormous, and the packets are independent. Moreover, the impact of a single packet on the system's performance is minimal. Also, the queue discipline will be first-in, first-out (FIFO), and the arrival packet is assumed to follow a Poisson process. The system's clock is constant, and the size of the tasks is not fixed. So, we suppose that service times are exponentially distributed [36]. In addition, we assume that the arrival packets of each service have the same priority, and we consider priority between services, not between UEs of one service. Because we assumed that the UEs of a service have the same priority, their sent packets also have an equal priority. Also, the services are isolated, and this priority does not invalidate the queue assumption. Furthermore, one service's priority over another can be higher, affecting the whole optimization and not queue formulation since the UEs in each service have the same priority, and each service has its processing delay independent of other services.

E. VNF Power

Assume the power consumption of each VNF in each logical node (the O-DU, the O-CU, and the UPF) in the slice s, is depicted as ϕ_s^d, ϕ_s^c, and ϕ_s^u, respectively.

So the system's total cost of energy of all the slices can be represented as $\phi_{\text{tot}} = \sum_{s=1}^{S} \phi_s$. where ϕ_s is obtained from $\phi_s = M_s^u \phi_s^u + M_s^c \phi_s^c + M_s^d \phi_s^d$. Moreover, ϕ_s^u, ϕ_s^c, and ϕ_s^d are the fixed cost of energy in UPF, O-CU, and O-DU, respectively.

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 virtual machines (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 ($\phi^{\text{tot}} \leq \phi^{max}$).

F. Problem Statement

Suppose slice s (which is assigned to service s) has the 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 (SLA) of that service to have a fairness in the system. This paper aims to maximize the sum-rate of all UEs with the presence of constraints as follows.

$$\max_{\boldsymbol{P},\boldsymbol{E},\boldsymbol{M},\boldsymbol{G}} \quad \sum_{s=1}^{S}\sum_{i=1}^{U_s} \delta_s \bar{\mathcal{R}}_{u(s,i)} \quad (13a)$$

subject to
$$P_r \leq P_r^{max} \quad \forall r \quad (13b)$$
$$p_{r,u(s,i)}^k \geq 0 \quad \forall i, \forall r, \forall s, \forall k, \quad (13c)$$
$$p_{r,u(s,i)}^k \leq P_s^{max} \quad \forall i, \forall r, \forall s, \forall k, \quad (13d)$$
$$\bar{\mathcal{R}}_{u(s,i)} \geq \mathcal{R}_s^{min} \quad \forall s, \quad (13e)$$
$$C_r \leq C_r^{max} \quad \forall r, \quad (13f)$$
$$T_{u(s,i)}^{tot} \leq T_s^{max} \quad \forall i, \forall s, \quad (13g)$$
$$\mu_s \geq \alpha_s/M_s \quad \forall s, \quad (13h)$$
$$\bar{\mathcal{R}}_{u(s,i)} \geq \lambda_{u(s,i)} \quad \forall i, \forall s, \quad (13i)$$
$$0 \leq M_s \leq M_s^{max} \quad \forall s, \quad (13j)$$
$$\phi^{\text{tot}} \leq \phi^{max}, \quad (13k)$$
$$\sum_r g_{u(s,i)}^r = 1 \quad \forall s, \forall i, \quad (13l)$$
$$\sum_{k=1}^{K_s} g_{u(s,i)}^r e_{r,u(s,i)}^k \geq 1 \quad \forall s, \forall i, \forall r \quad (13m)$$
$$\sum_{s=1}^{S}\sum_{i=1}^{U_s} g_{u(s,i)}^r e_{r,u(s,i)}^k \leq 1 \quad \forall s, \forall i, \forall r \quad (13n)$$
$$g_{u(s,i)}^r \in \{0,1\} \quad \forall s, \forall i, \quad (13o)$$
$$e_{r,u(s,i)}^k \in \{0,1\} \quad \forall s, \forall i, \quad (13p)$$

where $\bar{\mathcal{R}}_{u(s,i)}$, $\forall s, \forall i$ is derived by using $\bar{I}_{r,u(s,i)}^k$ instead of $I_{r,u(s,i)}^k$ in (6) and (5). In addition, $\boldsymbol{P} = [p_{r,u(s,i)}^k]$, $\forall s, \forall i, \forall r, \forall k$, is the matrix of power for UEs, $\boldsymbol{E} = [e_{r,u(s,i)}^k]$, $\forall s, \forall i, \forall r, \forall k$ indicate the binary variable for PRB association. Moreover, $\boldsymbol{G} = [g_{u(s,i)}^r]$, $\forall s, \forall i, \forall r$ is a binary variable for O-RU association. Furthermore, $M = [M_s^d, M_s^c, M_s^u]$, $\forall s$ is the matrix that shows the number of VNFs in each layer of slice. (13b), (13c) and (13d) indicate that the power of each O-RU does not exceed the maximum power, the power of each UE is a positive integer value, and the power of each UE in each service does not exceed the maximum power of each service, respectively. Also, (13e) shows that the rate of each UE requesting each type of service, i.e., eMBB, mMTC, and URLLC, is more than a threshold, respectively. (13f) and (13g) expressed the limited fronthaul capacity and the limited end-to-end delay of the received signal, respectively. (13h) and (13i) denoted the stability of the M/M/1 queue model. (13j) restricted the number of VNF in each slice due to the limited resources. (13l) and (13m) guarantee that O-RU and PRB are associated with the UE, respectively. Also, (13n) ensures that each PRB can not be assigned to more

than one UE associated with the same O-RU. In addition, (13k) indicates that the fixed cost of energy of VNFs in each slice does not exceed the threshold. Moreover, (13o) and (13p) depict that $\boldsymbol{E}$ and $\boldsymbol{G}$ are matrix of binary variables.

1) PRB Scheduling: In this section, we study the problem of PRB scheduling to illuminate the inter-slice interference and guarantee the isolation of slices [37]. We need to have an algorithm to illuminate the inter-slice interference before solving the problem 13. Firstly, we should assign PRBs to slices, and in the second step, the assignment of PRBs of each slice to each UEs of a specific slice is performed. So, the assignment of PRB can be completed in two steps to remove inter-slice interference and isolate the slices. Firstly we assign PRBs to slices. Secondly, we allocate power of UEs, assign PRBs of slices to UEs, find the optimal number of VNFs for each slice and assign O-RU to UEs, which uses the proposed algorithm III. Firstly, we need to find the average PRB number used by UEs in each service. Since mMTC and URLLC transmit a short packet, 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 SNR of UEs in slice s. Also, $\mathcal{R}_s^{min}$ is the minimum data rate of each UE in slice s (eMBB slice). So, the average number of PRB that slice s with U_s UEs requires is $K_s = U_s \times \frac{\bar{R}_s}{\mathcal{R}_s^{min}}$. K_s is the average number of PRBs needed for slice s, and K is the total number of PRBs in the system. Our goal is to obtain the number of PRBs assigned to each slice s ($\bar{K}_s$). The problem can be written as follow

$$\max_{\bar{K}_s} \quad \sum_{s=1}^{S} \delta_s \bar{\mathcal{R}}_s^{min} \log(\frac{\bar{K}_s}{K_s}) \tag{14a}$$

$$\text{subject to} \quad \sum_{s} \bar{K}_s \leq K \tag{14b}$$

$$\bar{K}_s \leq K_s \quad \forall s, , \tag{14c}$$

III. PROPOSED ALGORITHM

In this section, we first apply some simplifications to the system; Solving the problem (13) 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 [32].

In the first level, the main purpose is to allocate adequate PRBs and the power to the UEs and allocate enough activated VNFs to each slice. So, at this level, we would like to obtain the variables $\boldsymbol{P}, \boldsymbol{E}$, and $\boldsymbol{M}$. Despite simplification of the problem (13), it is still NP-hard and challenging to solve. Therefore, we relax the variable $\boldsymbol{E}$ [18], [32] and reformulating the constraint (13g), 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 $p_{u(s,l)}^k$ and $p_{u(n,l)}^k$ in (3) with P_s^{max}, an upper bound $\bar{I}_{r,u(s,i)}^k$ is obtained for $I_{r,u(s,i)}^k$, and also the lower bound $\bar{\rho}_{u(s,i)}^k$ is achieved for $\rho_{u(s,i)}^k$. Moreover, the lower bound $\bar{\mathcal{R}}_{u(s,i)}, \forall s, \forall i$ for $\mathcal{R}_{u(s,i)}$ is obtained by replacing $I_{r,u(s,i)}^k$ with $\bar{I}_{r,u(s,i)}^k$ in (6) and (5) and make these equations become concave functions.

Suppose $\hat{\rho}_{r,u(s,i)}^k = \frac{|P_s^{max}\mathbf{h}_{r,u(s,i)}^{H\ k}\mathbf{w}_{r,u(s,i)}^k g_{u(s,i)}^r|^2}{BN_0}$. we replace $\rho_{r,u(s,i)}^k$ with $\hat{\rho}_{r,u(s,i)}^k$ in (7), to convexify the (6) for the URLLC and mMTC services that have the short packet transmission. So, a lower bound for (6) is given that is a concave function.

$$\bar{\mathcal{R}}_{u(s_j,i)}^r = \sum_{k=1}^{K_{s_j}} B(\log_2(1 + \bar{\rho}_{u(s_2,i)}^k) - \hat{\zeta}_{u(s_j,i)}^k)e_{u(s_j,i)}^k \tag{15a}$$

$$\bar{\mathcal{R}}_{u(s_j,i)} = \sum_{r=1}^{R} \bar{\mathcal{R}}_{u(s_j,i)}^r \tag{15b}$$

$$\hat{\zeta}_{u(s_j,i)}^k = log_2(e)Q^{-1}(\epsilon)\sqrt{\frac{\hat{\mathfrak{C}}_{u(s_j,i)}^k}{N_{u(s_j,i)}^k}} \tag{15c}$$

$$\hat{\mathfrak{C}}_{u(s_j,i)}^k = 1 - \frac{1}{(1 + \hat{\rho}_{u(s_j,i)}^k)^2}. \tag{15d}$$

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 $\mu_s = \mu_s^u \approx \mu_s^c \approx \mu_s^d$. Moreover, as mentioned before, the mean arrival data rate of the UPF layer for a service s (α_s^U) is equal to the mean arrival data rate of the O-CU-UP layer (α_s^C) and O-DU (α_s^D). So $\alpha_s = \alpha_s^U \approx \alpha_s^C \approx \alpha_s^D$. 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 $\phi_s^u = \phi_s^c = \phi_s^d$. As a result, we have $M_s = M_s^u = M_s^c = M_s^d$. Using the above assumption, we have $T_s^{DU} = T_s^{CU} = T_s^{UPF}$ and we have $T_s^{\text{proc}} = T_s^{RU} + T_s^{DU} + T_s^{CU} + T_s^{UPF}$. So, $T_s^{\text{proc}} = T_s^{RU} + 3 \times T_s^{DU}$.

Although the problem is feasible and it has feasibility points that are discussed in subsection IV-B, it is not convex and difficult to solve. Since problem (13) is mixed-integer nonlinear programming with two integer constraints which are a binary variable (e shows the PRB assignment) and an integer variable (M_s indicate the number of VNFs in

slice s), the problem is NP-hard. Solving the problem is complex. To solve the problem by inspiring Stackelberg, we reformulate the equation (13g) to reduce one of the variables (M_s) that can be solved after obtaining the rate of UEs (M_s is similar to followers in Stackelberg Competition and power and PRB assignment is similar to leader). So, the new problem has two variables of power and PRB assignment. This new problem is convex by relaxing the binary variable (PRB assignment) and estimating the lower bounds in equation (15) because 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 the following, we define a lemma to find the upper and lower bound 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 $M_s = \max\{\mathtt{M}_{u(s,i)} | i \in 1, 2, ..., U_s\}$ $\forall s$. 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 (13), the constraint (13g) can be reformulated as $\forall i, \forall s$

$$T_s^{\max} \geq \frac{1}{R_{u(s,i)} - \lambda_{u(s,i)}} + \frac{3}{\mu_s - \alpha_s/M_s}$$
$$M_s \geq \frac{\alpha_s(T_s^{\max}R_{u(s,i)} - T_s^{\max}\lambda_{u(s,i)} - 1)}{(T_s^{\max}\mu_s - 3)(R_{u(s,i)} - \lambda_{u(s,i)}) - \mu_s} \quad (16)$$

Also from equations in (13k), (13h) and (13j), we have

$$\alpha_s/\mu_s \leq M_s \leq \min\{M^{\max}, \phi_{\max}/3\phi_s\} \quad (17)$$

We denote $\mathfrak{M}_s = \min\{M^{max}, \phi_{max}/3\phi_s\}$. Thus, if we restrict constraint (13g) to equality, constraint (13g) is still valid. Also, we have the following inequality.

$$\alpha_s/\mu_s \leq \frac{\alpha_s(T_s^{max}R_{u(s,i)} - T_s^{max}\lambda_{u(s,i)} - 1)}{(T_s^{max}\mu_s - 3)(R_{u(s,i)} - \lambda_{u(s,i)}) - \mu_s} \leq \mathfrak{M}_s \quad (18)$$

In equation (18), $0 \leq \frac{\alpha_s(T_s^{max}R_{u(s,i)} - T_s^{max}\lambda_{u(s,i)} - 1)}{(T_s^{max}\mu_s - 3)(R_{u(s,i)} - \lambda_{u(s,i)}) - \mu_s}$ is established due to the fact that the numerator and the denominator will both have the same sign. In the numerator, according to the (13i), $R_{u(s,i)} - \lambda_{u(s,i)} \geq 0$, 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 (13i) 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 (18), leads to $R_{u(s,i)} \geq \lambda_{u(s,i)}$ that is the constraint (13i). For the right side, by reformulating

the equation (18), we have a new constraint $\forall i, \forall s$ as below,

$$\mathcal{R}_{u(s,i)} \geq \varpi_{u(s,i)}. \quad (19a)$$

$$\varpi_{u(s,i)} = \lambda_{u(s,i)} + \frac{1}{T_s^{max}} \quad (19b)$$

$$+ \frac{3}{T_s^{max}\mu_s - \alpha_s T_s^{max}/\mathfrak{M}_s - 3} \quad (19c)$$

In addition, we denote $\mathtt{M}_{u(s,i)} = \frac{\alpha_s(T_s^{max}R_{u(s,i)} - T_s^{max}\lambda_{u(s,i)} - 1)}{(T_s^{max}\mu_s - 3)(R_{u(s,i)} - \lambda_{u(s,i)}) - \mu_s}$ for each UE i in slice s. So 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 $M_s = \max\{\mathtt{M}_{u(s,i)} | i \in 1, 2, ..., U_s\}$ $\forall s$. $\square$

Despite simplifying the problem in (13), 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 $e_{r,u(s,i)}^k \in [0,1]$ $\forall s, \forall i, \forall r, \forall k$ [18], [32]. Furthermore, the problem can be solved using the Lagrangian function and iterative algorithm.

In order to make (13) as a standard form of a convex optimization problem, it is required to change the variable of equations (9) to $P_r = \sigma_{q_r}^2 \times 2^{C^r}$ so the constraint (13f) is changed to $P_r \leq \sigma_{q_r}^2 \times 2^{C_{max}^r}$. The combination of equations (13b) and (13f) leads to the following equation

$$\zeta_r = min\{P_{max}, \sigma_{q_r}^2 \times 2^{C_{max}^r}\},$$
$$P_r \leq \zeta_r. \quad (20)$$

Moreover, the combination of equations in (13e), (13i) and (19) leads to the following equation

$$\eta_{u(s,i)} = \max\{\mathcal{R}_{u(s,i)}^{\min}, \lambda_{u(s,i)} + 1/T_{\max}^s, \varpi_{u(s,i)}\},$$
$$\bar{\mathcal{R}}_{u(s,i)} \geq \eta_{u(s,i)}. \quad (21)$$

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

$$\mathcal{L}(\boldsymbol{P}, \boldsymbol{E}; \mathfrak{v}, \chi, \mathfrak{h}, \xi, \kappa, \mathfrak{m}) = \sum_{s=1}^{S}\sum_{i=1}^{U_s}\delta_s\bar{\mathcal{R}}_{u(s,i)} \quad (22a)$$

$$+ \sum_{s=1}^{S}\sum_{i=1}^{U_s}\mathfrak{h}_{u(s,i)}(\bar{\mathcal{R}}_{u(s,i)} - \eta_{u(s,i)}) \quad (22b)$$

$$- \sum_{r=1}^{R}\mathfrak{m}_r(P_r - \zeta_r) \quad (22c)$$

$$+ \sum_{s=1}^{S}\sum_{i=1}^{U_s}\sum_{k=1}^{K}\sum_{r=1}^{R}\kappa_{r,u(s,i)}^k p_{r,u(s,i)}^k \quad (22d)$$

$$+ \sum_{s=1}^{S}\sum_{i=1}^{U_s}\sum_{k=1}^{K}\sum_{r=1}^{R}\mathfrak{q}_{r,u(s,i)}^k(P_s^{max} - p_{r,u(s,i)}^k) \quad (22e)$$

$$+ \sum_{r=1}^{R}\sum_{s=1}^{S}\sum_{i=1}^{U_s}\chi_{r,u(s,i)}(\sum_{k=1}^{K_s}e_{r,u(s,i)}^k - 1) \quad (22f)$$

$$- \sum_{s=1}^{S}\sum_{i=1}^{U_s}\sum_{k=1}^{K}\sum_{r=1}^{R}\mathfrak{v}_{r,u(s,i)}^k(e_{r,u(s,i)}^k - 1) \quad (22g)$$

$$+ \sum_{s=1}^{S} \sum_{i=1}^{U_s} \sum_{k=1}^{K} \sum_{r=1}^{R} \xi_{r,u(s,i)}^{k} e_{r,u(s,i)}^{k}. \tag{22h}$$

Lemma 2. *The derivatives of the Lagrangian function (22) with respect to the P and E give the Karush-Kuhn-Tucker (KKT) conditions to obtain the optimal value of these two variables [18], [32].*

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

$$\frac{\partial \mathcal{L}}{\partial p_{r,u(s,i)}^{k}} = (\delta_s + \mathfrak{h}_{u_{(s,i)}}) \mathfrak{B}_{r,u(s,i)}^{k} \tag{23}$$
$$+ (\mathfrak{s}_{r,u(s,i)}^{k} - \mathfrak{D}_{r,u(s,i)}^{k}) = 0,$$

where $\mathfrak{s}_{r,u(s,i)}^{k} = \kappa_{r,u(s,i)}^{k} - \mathfrak{q}_{r,u(s,i)}^{k}$ and other parameters are as follows:

$$\mathfrak{D}_{r,u(s,i)}^{k} = \mathfrak{m}_r |\mathbf{w}_{r,u(s,i)}^{k}|^2 g_{u(s,i)}^{r} e_{r,u(s,i)}^{k}, \tag{24a}$$

$$\mathfrak{B}_{r,u(s,i)}^{k} = \frac{B|\mathbf{h}_{r,u(s,i)}^{H\,k} \mathbf{w}_{r,u(s,i)}^{k}|^2 g_{u(s,i)}^{r} e_{r,u(s,i)}^{k}}{\ln(2)} \mathfrak{S}_{r,u(s,i)}^{k}, \tag{24b}$$

$$\mathfrak{S}_{r,u(s,i)}^{k} = \frac{1}{|\mathbf{h}_{r,u(s,i)}^{H\,k} \mathbf{w}_{r,u(s,i)}^{k}|^2 \mathfrak{k}_{r,u(s,i)}^{k} + BN_0 + I_{r,u(s,i)}^{k}}. \tag{24c}$$

Also, $\mathfrak{k}_{r,u(s,i)}^{k} = g_{u(s,i)}^{r} e_{r,u(s,i)}^{k} p_{r,u(s,i)}^{k}$. Thus, from equation (23), optimal power is obtained and power is allocated. We denote $\mathfrak{j}_{r,u(s,i)}^{k} = g_{u(s,i)}^{r} e_{r,u(s,i)}^{k}$. The optimal power is as follow.

$$p_{r,u(s,i)}^{k} = \left[\frac{(\delta_s + \mathfrak{h}_{u_{(s,i)}}) B \mathfrak{j}_{r,u(s,i)}^{k}}{Ln2 \times (-\mathfrak{s}_{r,u(s,i)}^{k} + \mathfrak{D}_{r,u(s,i)}^{k})} \tag{25a}$$

$$- \frac{BN_0 + I_{r,u(s,i)}^{k}}{|\mathbf{h}_{r,u(s,i)}^{H\,k} \mathbf{w}_{r,u(s,i)}^{k}|^2 \mathfrak{j}_{r,u(s,i)}^{k}} \right]^{+}. \tag{25b}$$

Also $[a]^{+} = \max(0, a)$. In addition, PRB assignment can be achieved from the derivatives of the Lagrangian function (22) with respect to the E as follow.

$$\frac{\partial \mathcal{L}}{\partial e_{r,u(s,i)}^{k}} = \bar{\mathcal{R}}_{r,u(s,i)}^{k} (\delta_s + \mathfrak{h}_{u_{(s,i)}}) \tag{26a}$$

$$- \mathfrak{m}_r |\mathbf{w}_{r,u(s,i)}^{k}|^2 p_{r,u(s,i)}^{k} g_{u(s,i)}^{r} \tag{26b}$$

$$+ (\xi_{r,u(s,i)}^{k} - \mathfrak{v}_{r,u(s,i)}^{k} + \chi_{r,u(s,i)}) = 0. \tag{26c}$$

So, the optimal E is obtained using KKT condition as follow.

$$e_{r,u(s,i)}^{k} \times (\mathfrak{F}_{r,u(s,i)}^{k} - \mathfrak{v}_{r,u(s,i)}^{k} \tag{27}$$
$$- \mathfrak{m}_r |\mathbf{w}_{r,u(s,i)}^{k}|^2 p_{r,u(s,i)}^{k} g_{u(s,i)}^{r}) = 0.$$

where $\mathfrak{F}_{r,u(s,i)}^{k} = \bar{\mathcal{R}}_{r,u(s,i)}^{k} (\delta_s + \mathfrak{h}_{u_{(s,i)}}) + (\xi_{r,u(s,i)}^{k} + \chi_{r,u(s,i)})$. Hence, from equation (26) and (27), PRB assignment is performed as follow.

$$e_{r,u(s,i)}^{k} = \begin{cases} 1 & u(s,i) = \text{argmax} \mathfrak{Z}_{r,u(s,i)}^{k} \forall r, k \in K, s \in S, \\ 0 & \text{otherwise}, \end{cases} \tag{28}$$

where $\mathfrak{Z}_{r,u(s,i)}^{k} = (\mathfrak{F}_{r,u(s,i)}^{k} - \mathfrak{v}_{r,u(s,i)}^{k} - $

$$\mathfrak{m}_r |\mathbf{w}_{r,u(s,i)}^{k}|^2 p_{r,u(s,i)}^{k} g_{u(s,i)}^{r}). \qquad \square$$

Thus, the user in slice s that has the most considerable value of $\mathfrak{F}_{r,u(s,i)}^{k}$, 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.

Assume P and E are fixed, we want to find G. Next, we introduce a greedy algorithm that assigns an O-RU to each UE.

Greedy Algorithm Assignment (GAA): The problem can be reformulated as follow

$$\max_{G} \quad \sum_{s=1}^{S} \sum_{i=1}^{U_s} \sum_{r=1}^{R} \delta_s g_{u(s,i)}^{r} \bar{\mathcal{R}}_{u(s,k)}^{r} \tag{29a}$$

$$\text{subject to} \quad \sum_{s=1}^{S} \sum_{i=1}^{U_s} g_{u(s,i)}^{r} \psi_{r,u(s,i)} \leq \mathfrak{t}_r \quad \forall r \tag{29b}$$

$$\sum_{r} g_{u(s,i)}^{r} = 1 \quad \forall s, \forall i, \tag{29c}$$

$$g_{u(s,i)}^{r} \in \{0,1\} \quad \forall s, \forall i, \tag{29d}$$

Where $\psi_{r,u(s,i)} = \sum_{k=1}^{K_s} |\mathbf{w}_{r,u(s,i)}^{k}|^2 p_{r,u(s,i)}^{k} e_{r,u(s,i)}^{k}$ and $\mathfrak{t}_r = \zeta_r - \sigma_r$ because of the equations (20) and (8). Since we obtained (21) in (III-A), we can ignore this constraint in (29). The problem (29) is an NP-complete 0-1 multiple knapsack problem. We solve this problem using heuristic method(GAA method 1), which is a greedy algorithm [18], [38]. Firstly, we set all the variables to zero ($g_{u(s,i)}^{r} = 0, \quad \forall s, \forall i, \forall r$). Then we define the parameter $\mathfrak{B}_{u_{(s,i)}}^{rem}$. 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 ($\mathfrak{B}_{u_{(s,i)}}^{rem} = \mathcal{R}, \forall s, \forall i$). 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 ($\mathfrak{B}_{u_{(s,i)}}^{rem} = \mathfrak{B}_{u_{(s,i)}}^{rem} \setminus \{r^*\}$). Then, the O-RU with the highest achievable data rate from the new set of O-RUs $\mathfrak{B}_{u_{(s,i)}}^{rem}$ is selected. The complexity of sorting S slices based on their priority is $O(Slog(S))$. 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 $O(Slog(S)) + O(R \times \mathfrak{N})$.

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

1: Set $g_{u_{(s,i)}}^r = 0$, $\mathfrak{C}_r = \mathfrak{t}_r$, and $\mathfrak{B}_{u_{(s,i)}}^{rem} = \mathcal{R}$ $\forall s, \forall i, \forall r$.
2: Sort slices according to their priority factor (δ_s) in descending order
3: **for** $s \leftarrow 1$ to S **do**
4:     **for** $i \leftarrow 1$ to U_s **do**
5:         $RU = 0$
6:         **for** $r \leftarrow 1$ to R **do**
7:             Acquire $\mathfrak{S}_{u_{(s,i)}}^r = \bar{\mathcal{R}}_{u_{(s,i)}}^r$
8:         **end for**
9:         Obtain $r^* = \arg\max_{r \in \mathfrak{B}_{u_{(s,i)}}^{rem}} \mathfrak{S}_{u_{(s,i)}}^r$
10:         **while** $RU == 0$ **do**
11:             **if** $\mathfrak{C}_{r^*} \geq \psi_{r^*, u(s,i)}$ **then**
12:                 Set $g_{u(s,i)}^{r^*} = 1$
13:                 Set $\mathfrak{C}_{r^*} = \mathfrak{C}_{r^*} - \psi_{r^*, u(s,i)}$
14:                 Set $RU = 1$
15:             **else**
16:                 $\mathfrak{B}_{u_{(s,i)}}^{rem} = \mathfrak{B}_{u_{(s,i)}}^{rem} \setminus \{r^*\}$
17:             **end if**
18:         **end while**
19:     **end for**
20: **end for**

Algorithm 2 Iterative algorithm for the baseband resource allocation and VNF activation (IABV)

1: Set the maximum number of iterations $Iter_{max}$, convergence condition $\epsilon > 0$
2: Assign Users to O-RU randomly (Initialize G)
3: **for** $i \leftarrow 1$ to $Iter_{max}$ **do**
4:     Acquire $P^{(i)}$, $E^{(i)}$ and $M^{(i)}$ using Lagrangian function and sub-gradient method based on (III-A)
5:     Update $G^{(i)}$ based on algorithm GAAOU (1) in (III-B)
6:     **if** the algorithm converged with the tolerence of ϵ **then**
7:         Break
8:     **else**
9:         Continue the algorithm
10:     **end if**
11: **end for**

C. Iterative Proposed Algorithm

In Sections (III-A) and (III-B), the details of solving each sub-problem are depicted. Here, the iterative algorithm for the whole problem is demonstrated. Firstly, we fixed G to achieve P and E, using the Lagrangian method and the KKT conditions. Afterward, 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 $\mathfrak{N} = \sum_{s=1}^{S} \sum_{i=1}^{U_S} 1$. 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 $O(Slog(S)) + O(R \times \mathfrak{N})$. So the complexity of the main problem (13) is $O(T_{conv} \times \mathfrak{N} \times K \times (Slog(S) + R\mathfrak{N}))$.

2) Convergence Analysis: We can guarantee the convergence of the iterative algorithm if the objective function is the strictly ascending function concerning the number of iterations [39]. 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. So the the aggregate throughput is ascending function concerning the number of iterations and it will converge to the sub-optimal solution. In addition, if we assume that the interference is set to be zero $I_{r,u_{(s,i)}}^k = 0$, and we suppose that each UE has the maximum power $p_{r,u_{(s,i)}}^k = P_s^{max}$, and we consider that all PRB is assigned to all UE $e_{r,u_{(s,i)}}^k = 1 \forall s, \forall i$ 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 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.

Algorithm 3 Fast Algorithm (FA) to Check the Convergence

1: Set count = 0
2: Set $p_{r,u(s,i)}^k = 0$, $e_{r,u(s,i)}^k = 0$ and $g_{u(s,i)}^r = 0$ $\forall r, k, s, i$
3: **for** $s \leftarrow 1$ to S **do**
4:     **for** $i \leftarrow 1$ to U_s **do**
5:         count = count +1
6:         $r^* = \arg\min_r d_{r,u(s,i)}$ $\forall r$
7:         $g_{u(s,i)}^{r^*} = 1$
8:         temp =mod(count, K)
9:         **if** temp=0 **then**
10:             $e_{r^*,u(s,i)}^K = 1$
11:             Set $p_{r^*,u(s,i)}^K = \min\{P_s^{max}, P_r^{max}/\mathfrak{N}\}$
12:         **else**
13:             $e_{r^*,u(s,i)}^{temp} = 1$ $\forall r$
14:             Set $p_{r^*,u(s,i)}^{temp} = \min\{P_s^{max}, P_r^{max}/\mathfrak{N}\}$
15:         **end if**
16:     **end for**
17: **end for**

IV. NUMERICAL RESULTS AND THE FEASIBLE REGION

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
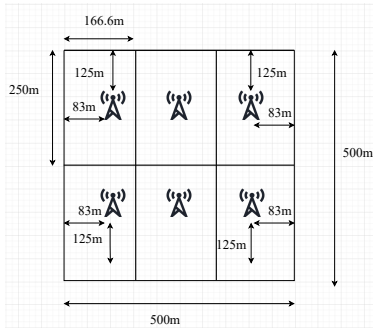
Fig. 4: O-RU placement in a cell

we have six 4-antenna O-RU (MISO) located in a place with a diameter of 500 meters as shown in figure 4. 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 $\mathbf{h}_{r,u(s,i)}^{k} = d_{r,u(s,i)}^{-\mathcal{L}} \Omega_{r,u(s,i)}^{k}$, where $d_{r,u(s,i)}^{-\mathcal{L}}$ is the distance between the O-RU r and UE i in service s and $\mathcal{L} = 3.8$ is the path-loss exponent [40]. Also, $\Omega_{r,u(s,i)}^{k}$ 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 are used in the network and the packet size for mMTC is equal to 20 bytes, and for URLLC is equal to 32 bytes. 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 $\mathfrak{q}_{r,u(s,i)}^{k} = P_s^{max}/100$. The other parameters of these simulations are depicted in Table I.

TABLE I: Simulation Parameters

Parameter	Value
Noise power	-174dBm
Bandwidth	180 KHz
Maximum transmit Power of each O-RU	40dBm
Maximum delay for eMBB	4msec
Maximum delay for URLLC	1msec
Maximum delay for mMTC	5msec
Maximum fronthaul capacity	46 bits/sec/Hz
Minimum data rate for eMBB	20 bits/sec/Hz
Minimum data rate for URLLC and mMTC	2 bits/sec/Hz
Maximum received power for mMTC	20 dBm
Maximum received power for eMBB and URLLC	33 dBm

Finding a feasible initial value is almost tricky. To overcome this challenge, we use a fast method that is discussed in IV-B. If the fast method converges and has a feasible solution, so our algorithm (IABV) can be converged too. Two different methods are used to compare with 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. So, 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 achieving power and other parameters, the achievable rate will be obtained and the optimal number of VNF is achieved from the lemma (1). The second one is similar to the fixed BBU capacity and dynamic resource allocation (FBDR) algorithm proposed in [18]. In this work, we have services with different QoS that contain UEs, which is similar to tenants with a different QoS. So, we used an algorithm similar to FBDR adapted to our conditions for comparison. We assume here that the capacity of the O-DU (which is similar to BBU in the FBDR) is as large as the capacity of our fronthaul link so that we can use this method for our system model. In this method, 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 [18] show that dynamic BBU capacity and dynamic resource allocation (DBDR) perform better than FBDR for the same priority area. We will also see that our proposed algorithm performs better than FBDR in the next subsection.

B. Feasible Region

Applying the correct initial point to make the system feasible is a significant step in our work. During the simulations, it can be noted that sometimes the algorithm does not converge for some of the iterations with fixed initial values. To solve this problem, we investigated the non-converging and converging simulation for models with fixed initial parameters (such as number of UEs, the threshold of power and rate,...) and random channel gains of UEs. By comparing convergent and non-convergent simulations, we experimentally found that in cases where convergence does not occur, 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. In the simulation part, if more than half of the iterations have a feasible solution for the initial condition, the simulation can be displayed as a feasible model. To remove non-converging simulations, we need to use a fast algorithm (FA) to check the convergence before the proposed algorithm (IABV). If the conditions in (21), (20), (13d) and (13c) 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 rth 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 ($\min\{P_s^{\max}, P_r^{\max}/\mathfrak{N}\}$). Moreover, the allocation of PRBs to UEs is based on dividing the number of UEs by the total number of PRBs. If the number of UEs is more than the PRBs, each UE is given exactly one PRB without interference. Otherwise, the amount of interference will not be high but more than one UE will use same PRBs.

C. Numerical Results

In Fig. 5, the aggregate throughput is demonstrated versus the different 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. Figure 6 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 figure 7, the aggregate throughput is depicted vs. the maximum power of UE for three different instances of eMBB service using proposed method (IABV), FBDR 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. In addition, We suppose that each O-RU can transmit three times of the maximum power of the UEs. 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 FBDR and the baseline scheme. Figure 8 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. In this figure, we assume that the maximum number of VNF for each slice is 50 and the maximum delay of each UE in a URLLC service is $0.5ms$. Also, the maximum number of PRB is considered to be 50. Moreover, we can see that the mean delay of a URLLC service does not reach the maximum threshold of the delay. Figure 9 is the same as figure 8 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. Figure 10, represents the aggregate throughput concerning the number of UEs in each service and the maximum power for three different mMTC service instances. mMTC service includes a large number of UEs with low data rates and low power. 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, the aggregate throughput increases. Also, by increasing the maximum power of each UE in each instance of mMTC service, the aggregate throughput rises too. Assume we have two types of eMBB service instances. In figure 11, 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. We assume that the the fronthaul link and the end-to-end delay have no restrictions. The figure 11 presented that by increasing the priority factor for one service instance, 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 figure 12, 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 figure 13, 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 other parameters are depicted in table I. 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.

In figure 14, the aggregate throughput is depicted vs. the maximum interference for two different maximum power thresholds of O-RU. We suppose that the maximum power threshold of UEs is one-third of the maximum power 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 figure 15, the aggregate throughput is shown versus the different number of UEs
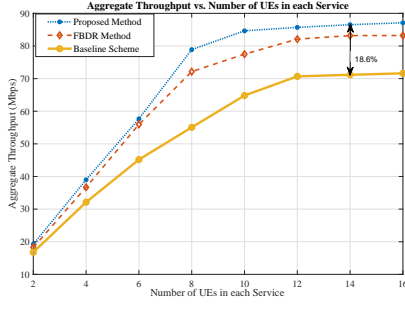
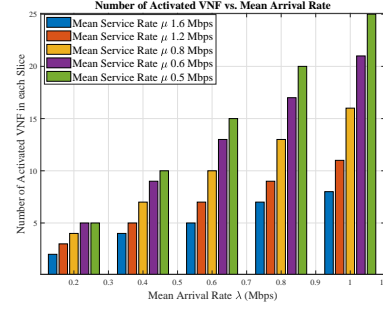Fig. 5: Aggregate throughput vs. number of UEs in each service

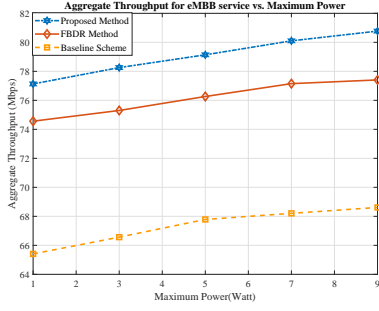Fig. 6: Number of activated VNF in each Service vs. Mean Arrival Rate(Mbps)

Fig. 7: Aggregate Throughput for eMBB vs. Maximum Transmit power for various number of UEs
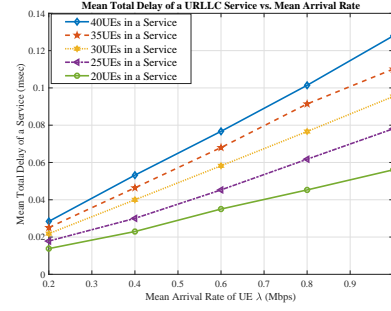

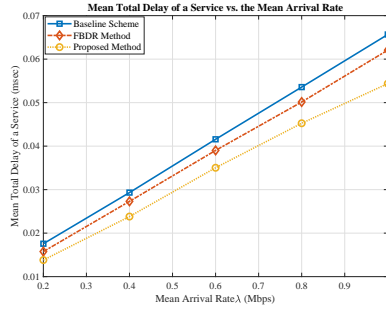Fig. 8: Mean Total Delay of a URLLC Service vs. the Mean Arrival Rate of a UE in the Service

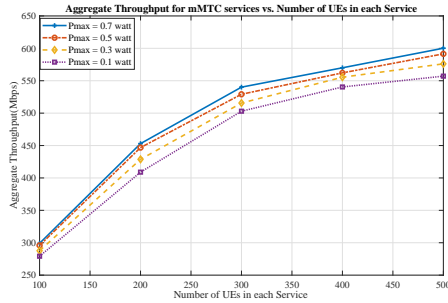Fig. 9: Mean Total Delay of a URLLC Service vs. the

Fig. 10: Aggregate Throughput vs. Number of UEs in each Service for three different mMTC services
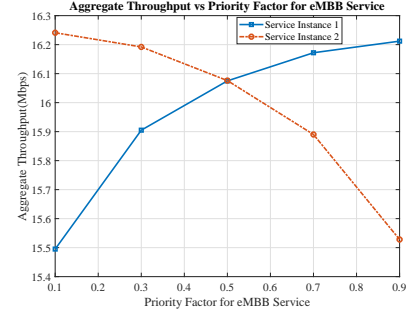

Fig. 11: Aggregate Throughput for two eMBB service instances vs. Priority of the first service instance

aggregate throughput raises. And we can see that the IABV method is better than the FA method.

V. CONCLUSION

In this paper, we modeled the downlink of the O-RAN system using network slicing for different 5G services, i.e., eMBB, mMTC, and URLLC. The isolation of various services, i.e., eMBB, mMTC, and URLLC in the O-DU, the O-CU, and the UPF, is accomplished. Also, the paper aims to obtain the number of activated VNFs in each service, RU association, power, and PRB allocation to maximize the aggregate throughput. The limited fronthaul capacity and the mean end-to-end 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

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
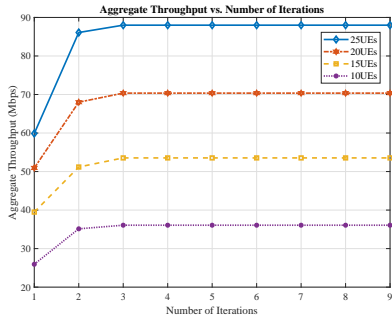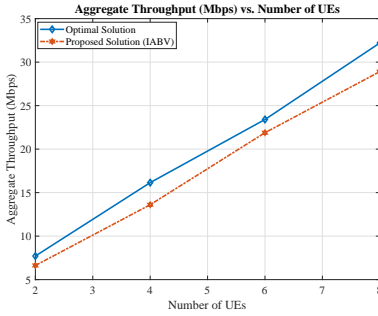
Fig. 12: Aggregate Throughput vs. Number of

Fig. 13: Aggregate Throughput vs. Number of UEs in
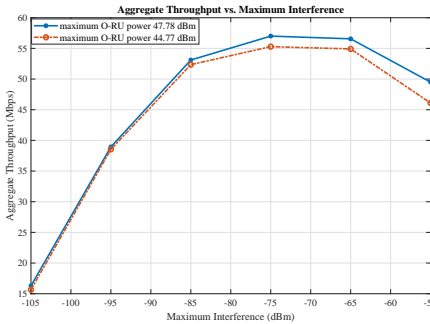one Service

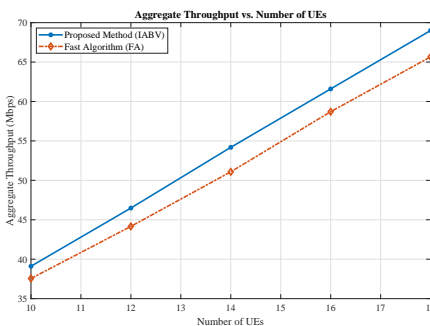

Fig. 14: Aggregate Throughput vs. Maximum

Fig. 15: Aggregate Throughput vs. the Number of UEs
in one Service

out. The performance of our proposed method (i.e., IABV) is compared with the baseline scheme and FBDR in [18]. 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 FBDR and the baseline scheme.

REFERENCES

[1] X. Shen, J. Gao, W. Wu, K. Lyu, M. Li, W. Zhuang, X. Li, and J. Rao, "Ai-assisted network-slicing based next-generation wireless networks," *IEEE Open Journal of Vehicular Technology*, vol. 1, pp. 45–66, Jan. 2020.

[2] M. Setayesh, S. Bahrami, and V. W. Wong, "Joint prb and power allocation for slicing embb and urllc services in 5g c-ran," in *IEEE Global Communications Conference (GLOBECOM 2020)*, 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. 55 765–55 779, Sep 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. 67 512–67 547, Oct. 2020.

[5] R. Kassab, O. Simeone, and P. Popovski, "Coexistence of urllc and embb services in the c-ran uplink: An information-theoretic study," in *2018 IEEE Global Communications Conference (GLOBECOM)*, Abu Dhabi, United Arab Emirates, Dec. 2018, pp. 1–6.

[6] 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 Transactions on Wireless Communications*, vol. 20, no. 7, pp. 4585 – 4600, Feb. 2021.

[7] B. Han, L. Liu, J. Zhang, C. Tao, C. Qiu, T. Zhou, Z. Li, and Z. Piao, "Research on resource migration based on novel rrh-bbu mapping in cloud radio access network for hsr scenarios," *IEEE Access*, vol. 7, pp. 108 542–108 550, Aug. 2019.

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

[9] S. Niknam, A. Roy, H. S. Dhillon, S. Singh, R. Banerji, J. H. Reed, N. Saxena, and S. Yoon, "Intelligent o-ran for beyond 5g and 6g wireless networks," *arXiv preprint arXiv:2005.08374*, May 2020.

[10] N. Kazemifard and V. Shah-Mansouri, "Minimum delay function placement and resource allocation for open ran (o-ran) 5g networks," *Computer Networks*, vol. 188, p. 107809, Apr. 2021.

[11] C. B. Both, J. Borges, L. Gonçalves, C. Nahum, C. Macedo, A. Klautau, and K. Cardoso, "System intelligence for uav-based mission critical with challenging 5g/b5g connectivity," *arXiv preprint arXiv:2102.02318*, Feb. 2021.

[12] "O-ran architecture description," O-RAN Alliance, Tech. Rep., 2020.

[13] O.-R. W. G. 2, "Ai/ml workflow description and requirements," O-RAN Alliance, Tech. Rep., Mar. 2020.

[14] B.-S. Lin, "Toward an ai-enabled o-ran-based and sdn/nfv-driven 5g& iot network era," *Network and Communication Technologies*, vol. 6, no. 1, pp. 6–15, Jun. 2021.

[15] 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 Communications surveys & tutorials*, vol. 18, no. 1, pp. 236–262, Sep. 2015.

[16] Z. Luo and C. Wu, "An online algorithm for vnf service chain scaling in datacenters," *IEEE/ACM Transactions on Networking*, vol. 28, no. 3, pp. 1061–1073, Mar. 2020.

[17] 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. 34 538–34 551, Feb. 2020.

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

[19] Y. L. Lee, J. Loo, and T. C. Chuah, "A new network slicing framework for multi-tenant heterogeneous cloud radio access networks," in *2016 International Conference on Advances in Electrical, Electronic and Systems Engineering (ICAEES)*. Putrajaya, Malaysia: IEEE, Nov. 2016, pp. 414–420.

[20] H. Xiang, S. Yan, and M. Peng, "A realization of fog-ran slicing via deep reinforcement learning," *IEEE Transactions on Wireless Communications*, vol. 19, no. 4, pp. 2515–2527, Jan. 2020.

[21] S. E. Elayoubi, S. B. Jemaa, Z. Altman, and A. Galindo-Serrano, "5g ran slicing for verticals: Enablers and challenges," *IEEE Communications Magazine*, vol. 57, no. 1, pp. 28–34, Jan. 2019.

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

[23] P. Yang, X. Xi, T. Q. Quek, J. Chen, X. Cao, and D. Wu, "How should i orchestrate resources of my slices for bursty urllc service provision?" *IEEE Transactions on Communications*, vol. 69, no. 2, pp. 1134–1146, Nov. 2020.

[24] F. Saggese, M. Moretti, and P. Popovski, "Power minimization of downlink spectrum slicing for embb and urllc users," *arXiv preprint arXiv:2106.08847*, Jun. 2021.

[25] P. Korrai, E. Lagunas, S. K. Sharma, S. Chatzinotas, A. Bandi, and B. Ottersten, "A ran resource slicing mechanism for multiplexing of embb and urllc services in ofdma based 5g wireless networks," *IEEE Access*, vol. 8, pp. 45 674–45 688, Mar. 2020.

[26] J. Tang, W. P. Tay, T. Q. Quek, and B. Liang, "System cost minimization in cloud ran with limited fronthaul capacity," *IEEE Transactions on Wireless Communications*, vol. 16, no. 5, pp. 3371–3384, May 2017.

[27] K. Guo, M. Sheng, J. Tang, T. Q. Quek, and Z. Qiu, "Exploiting hybrid clustering and computation provisioning for green c-ran," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 12, pp. 4063–4076, Nov. 2016.

[28] P. Luong, F. Gagnon, C. Despins, and L.-N. Tran, "Joint virtual computing and radio resource allocation in limited fronthaul green c-rans," *IEEE Transactions on Wireless Communications*, vol. 17, no. 4, pp. 2602–2617, Feb. 2018.

[29] P. Luong, C. Despins, F. Gagnon, and L.-N. Tran, "A novel energy-efficient resource allocation approach in limited fronthaul virtualized c-rans," in *2018 IEEE 87th Vehicular Technology Conference (VTC Spring)*, Jun. 2018, pp. 1–6.

[30] S. Ali, A. Ahmad, and A. Khan, "Energy-efficient resource allocation and rrh association in multitier 5g h-crans," *Transactions on Emerging Telecommunications Technologies*, vol. 30, no. 1, p. e3521, Jan. 2019.

[31] S. Ali, A. Ahmad, Y. Faheem, M. Altaf, and H. Ullah, "Energy-efficient rrh-association and resource allocation in d2d enabled multi-tier 5g c-ran," *Telecommunication Systems*, vol. 74, no. 2, pp. 1–15, Jun. 2019.

[32] S. Ali, A. Ahmad, R. Iqbal, S. Saleem, and T. Umer, "Joint rrh-association, sub-channel assignment and power allocation in multi-tier 5g c-rans," *IEEE Access*, vol. 6, pp. 34 393–34 402, Jun. 2018.

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

[34] 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 Transactions on Communications*, vol. 69, no. 9, pp. 6063–6078, 2021.

[35] J. Cavazos. (2020) 5g testing: What is o-ran? part 2. [Online]. Available: https://blogs.keysight.com/blogs/inds.entry.html

[36] *Queueing Systems. Volume 1: Theory 1st Edition.* A Wiley-Interscience Publication, 1975.

[37] D. Marabissi and R. Fantacci, "Highly flexible ran slicing approach to manage isolation, priority, efficiency," *IEEE Access*, vol. 7, pp. 97 130–97 142, 2019.

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

[39] N. Gholipoor, S. Parsaeefard, M. R. Javan, N. Mokari, H. Saeedi, and H. Pishro-Nik, "Resource management and admission control for tactile internet in next generation of radio access network," *IEEE Access*, vol. 8, pp. 136 261–136 277, Jul. 2020.

[40] N. Gholipoor, S. Parsaeefard, M. R. Javan, and N. Mokari, "Cloud-based queuing model for tactile internet in next generation of ran," in *2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring)*, Antwerp, Belgium, Jun. 2020, pp. 1–6.