Resource Allocation in an Open RAN System using Network Slicing

1st Mojdeh Karbalaee Motalleb
Electrical and Computer Engineering
Tehran University
Tehran, Iran
mojdeh.karbalaee@ut.ac.ir

2nd Vahid Shah-Mansouri
Electrical and Computer Engineering
Tehran University
Tehran, Iran
vmansouri@ut.ac.ir

***Abstract*—Taking advantage of both virtual RAN (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 and reduces operational costs, allowing them to add new capabilities to the network more quickly. O-RAN separate 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 baseband resource allocation and virtual network function (VNF) activation in O-RAN architecture based on their service priority for different types of 5G services includes enhanced mobile broadband (eMBB), ultra-reliable low latency communications (URLLC) and massive Machine Type Communications (mMTC) services. According to the concept of network slicing, the isolation of different types of services in O-DU, O-CU, and user plane function (UPF) is performed. The limited fronthaul capacity and the restriction of end-to-end delay are considered in this problem. The optimization of baseband resources includes O-RU assignment, physical resource block (PRB), and power allocation. The main problem is the mixed-integer non-linear programming that is tremendously difficult. So we broke it down into two different steps that the iterative algorithm solves it. In the first step, we reformulated and simplified the problem to find the power allocation, PRB assignment, and the number of activated VNFs. In the second step, the O-RU association is achieved. The proposed method (IAPPVO) is confirmed by the simulation results which illustrate a higher achievable data rate than a baseline scheme that only optimizes one of the baseband resources and the FBDR algorithm described in other papers. Also, the simulation results attained less end-to-end delay for the proposed method than the FBDR and baseline scheme.**

***Index Terms*—Open Radio Access Network (O-RAN), Virtual Network Function (VNF)**

I. Introduction

One of the fifth-generation goals of the wireless system is to achieve the desired QoS (such as rate, delay, power, ...) for different types of services. Network slicing is the best solution for this aim. A network slice is an end-to-end logical network that offers services with specific needs. Multiple isolated network slices run, manage, and work independently on the same infrastructure. Several implementations of network slicing include core slicing, RAN slicing, and slicing of both sections. Different type of services includes enhanced mobile broadband (eMBB), and ultra-reliable low latency communications (URLLC), and massive Machine Type Communications (mMTC) services are introduced in 5th generation of mobile network. Each type of service requires a particular slice of network based on its QoS. eMBB requires a higher transmission data rate in comparison with 4G systems. Moreover, eMBB service satisfies the demands of high capacity and considerable coverage. URLLC requires low latency restrictions and higher reliability of the system. URLLC contains services such as autonomous vehicles, tactile internet, or remote surgery. mMTC consists of a vast number of low power internet of ihings (IoT) devices transmitting small payloads to a typical receiver. In addition, mMTC UEs include sensors for sensing, metering, and monitoring devices [1]–[6].

Recently, RAN virtualization attracts significant attention from industry and academia since it has remarkable benefits that increase flexibility and reduce operating costs such as CAPEX and OPEX and allow them to add new capabilities to the network more quickly. In addition to RAN virtualization, openness and RAN intelligence are two other fundamental points that encourage the Open Radio Access Network (O-RAN) Alliance to establish O-RAN as the next generation of RAN systems. The idea of O-RAN comes from the integration of virtual RAN (vRAN) and cloud RAN (CRAN), and it takes advantage of both. CRAN divides RAN into two parts radio remote head (RRH) and baseband unit (BBU). More than one distributed RRHs can be connected to a centralized BBU, which is named BBU-pool [7]. Unlike the previous generation of RAN that divides RAN into two parts, O-RAN separate 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 that includes higher PHY, MAC, and RLC. In addition, the O-CU depicts the logical node contains two parts, which are the O-CU user plane (O-CU-UP) and 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. One of the necessities of the new generation of wireless networks is its intelligence. Based on the requirement of an intelligent wireless network, O-RAN offers machine learning techniques. 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].

The separation of network software and hardware elements has been done and introduced as network function virtualization (NFV), and virtual network functions (VNF) are system function blocks. This technology improves the system performance in the fifth generation of telecommunications. The key idea of the implementation of NFV is to decouple software from physical hardware, dynamic scaling, and the deployment of flexible network functions. The NFV technology offers to execute VNFs on virtual machines or containers in a cloud system [15], [16]. As a result, some O-RAN components that include user plane function (UPF), O-CU, O-DU, and RAN Intelligent Controller (RIC)-near real-time, are virtualized and implemented as a VNF that can be run on virtual machines (VMs) or containers.

A. Related Works

Network slicing is increasingly receiving research attention. Many researchers studied the problem of resource allocation in network slicing for multitenant cellular networks [17]–[19]. In [18], dynamic network slicing in multitenant heterogeneous CRAN (H-CRAN) is considered. The process of allocating network resources to users is discussed. 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, in the lower level, the allocation of power and physical resource blocks (PRB) is performed. In the article [20], network slicing in the radio section is considered for fog or F-RAN structure, in which two network slices are set for hotspots and vehicle scenarios with related infrastructure. In [21], [22] the implementation of RAN level slicing is discussed in mobile network operator (MNO). Also, the problem of resource allocation is considered. Moreover, the paper faces the challenge of RAN slicing. The problem involves designing and managing multiple slices in the shared infrastructure efficiently while guaranteeing each slice's service level agreement (SLA).

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] 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 [24] the problem of power minimization for URLLC and eMBB services is presented for non-orthogonal multiple access (NOMA) and orthogonal multiple access (OMA). In [25], the authors proposed to allocate RAN resources for the network slicing system in the coexistence of eMBB and URLLC services. The system guarantees the latency, the service rate, and the maintenance of reliability.

Virtualization technique for RAN and core is one of the exciting topics. In [26], [27], the authors solve the problem of obtaining beamforming and VMs activation 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 delay are modeled based on M/M/1 queuing theory to guarantee delay for the UEs. 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.

B. Main Contribution

This study aims to optimize baseband resource allocation (power allocation, PRB allocation, and O-RUs association) and VNF activation, to develop an isolated network slicing outline for different types of services in an O-RAN platform. 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 follow:

- In this paper, a network slicing model is depicted for three different types of services introduced in 5G (URLLC, eMBB, and mMTC). 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 a 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 mean delay of the system based on the number of activated VNFs and each user's throughput. We consider the limited fronthaul capacity and accurately obtain the power and the capacity of each O-RU based on the quantization noise. We model the interference of neighboring O-RU and formulate the actual throughput for different types of services. We model the throughput of URLLC and mMTC based on their short packet transmission.
- 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.

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 whole paper.

II. System Model and Problem Formulation

In this section, we describe the downlink system in O-RAN slicing as depicted in Figure 1. Here, 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 discuss the mean delay and the power of VNFs. Finally, the main problem is expressed.

A. System Model

Suppose we have three service types includes mMTC, eMBB, and URLLC, which support different applications.

Assume we have S_1, S_2 and S_3 different applications for the first, second and third service type, respectively ($S = S_1 + S_2 + S_3$). So, we have S preallocated slices serving these S services; 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 is served by its corresponding slice.

Each Service $s_j \in \{1, 2, ..., S_j\}$ consists of U_{s_j} request from the single-antenna UEs which require certain QoS to be able to use the requested service ($j \in \{1, 2, 3\}$ indicate 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 $s_j \in \{1, 2, ..., S_j\}$, $j \in \{1, 2\}$ consists of preallocated virtual resource blocks that are mapped to the Physical Resource Blocks (PRBs), 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.

All K PRBs can be assigned to all UE 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-RU that are shared between slices. Each O-RU $r \in \{1, 2, ..., R\}$ has J antenna for transmitting and receiving data. Also $\mathcal{R} = \{r | r \in 1, 2, ..., R\}$ depicts the set of O-RUs. Moreover, all O-RUs, have access to all PRBs.

B. The Achievable Rate

The SNR of i^{th} UE served at slice s on PRB k is obtained from

$$\rho^k_{r,u(s,i)} = \frac{|p^k_{r,u(s,i)} \mathbf{h}^{H\,k}_{r,u(s,i)} \mathbf{w}^k_{r,u(s,i)}|^2}{BN_0 + I^k_{r,u(s,i)}}, \tag{1}$$

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

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

Moreover, $g^r_{u(s,i)} \in \{0, 1\}$ is the binary variable that illustrates whether O-RU r served the i^{th} UE that is allocated to s^{th} slice or not. Also, BN_0 denotes the power of Gaussian additive noise, and $I^k_{r,u(s,i)}$ is the power of interfering signals represented as follow.

$$\begin{aligned} I^k_{r,u(s,i)} = & \underbrace{\sum_{\substack{l=1 \\ l \neq i}}^{U_s} \gamma_1 p^k_{u(s,l)} \sum_{\substack{r'=1 \\ r' \neq r}}^{R} |\mathbf{h}^{H\,k}_{r',u(s,i)} \mathbf{w}^k_{r',u(s,l)} g^{r'}_{u(s,l)}|^2}_{\text{(intra-slice interference)}} \\ & + \underbrace{\sum_{\substack{n=1 \\ n \neq s}}^{S} \sum_{l=1}^{U_s} \gamma_2 p^k_{u(n,l)} \sum_{\substack{r'=1 \\ r' \neq r}}^{R} |\mathbf{h}^{H\,k}_{r',u(s,i)} \mathbf{w}^k_{r',u(n,l)} g^{r'}_{u(n,l)}|^2}_{\text{(inter-slice interference)}} \\ & + \underbrace{\sum_{j=1}^{R} {\sigma_q}^2 |\boldsymbol{h}_{r,u(s,i)}|^2}_{\text{(Quantization Noise)}} \end{aligned} \tag{3}$$

where $\gamma_1 = e^k_{u(s,i)} e^k_{u(s,l)}$ and $\gamma_2 = e^k_{u(s,i)} e^k_{u(n,l)}$. $e^k_{u(s,i)}$ is the binary variable to show whether the k^{th} PRB is allocated to the UE i in slice s, assigned to r^{th} O-RU.

To obtain SNR as formulated in (1), let $y_{u(s,i)}$ be the received signal of UE i in s^{th} service

$$y_{u(s,i)} = \textstyle\sum_{r=1}^{R} \sum_{k=1}^{K_s} \boldsymbol{h}^{H\,k}_{r,u(s,i)} g^r_{u(s,i)} e^k_{r,u(s,i)} \mathfrak{y}^k_{r,u(s,i)} + z_{u(s,i)}, \tag{4}$$

where $\mathfrak{y}^k_{r,u(s,i)} = \boldsymbol{w}^k_{r,u(s,i)} p^{k\,\frac{1}{2}}_{r,u(s,i)} x_{u(s,i)} + \boldsymbol{q}_r$ 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. In addition, $\boldsymbol{q}_r \in \mathbb{C}^J$ indicates the quantization 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 i^{th} UE request in the s_1^{th} application of service type 1 (eMBB) can be written as $\mathcal{R}_{u(s_1,i)}$ that is formulated as below.

$$\begin{aligned} \mathcal{R}^k_{r,u(s_1,i)} &= B \log_2(1 + \rho^k_{r,u(s_1,i)}), \\ \mathcal{R}^r_{u(s_1,i)} &= \sum_{k=1}^{K} B \log_2(1 + \rho^k_{r,u(s_1,i)} e^k_{r,u(s_1,i)}), \\ \mathcal{R}_{u(s_1,i)} &= \sum_{r=1}^{R} \mathcal{R}^r_{u(s_1,i)} g^r_{u(s_1,i)} \end{aligned} \tag{5}$$

where B is the bandwidth of system. $\mathcal{R}^r_{u(s_1,i)}$ is the achievable rate of each RU r to UE i in slice s_1. Since the blocklength in URLLC and mMTC is finite, the achievable

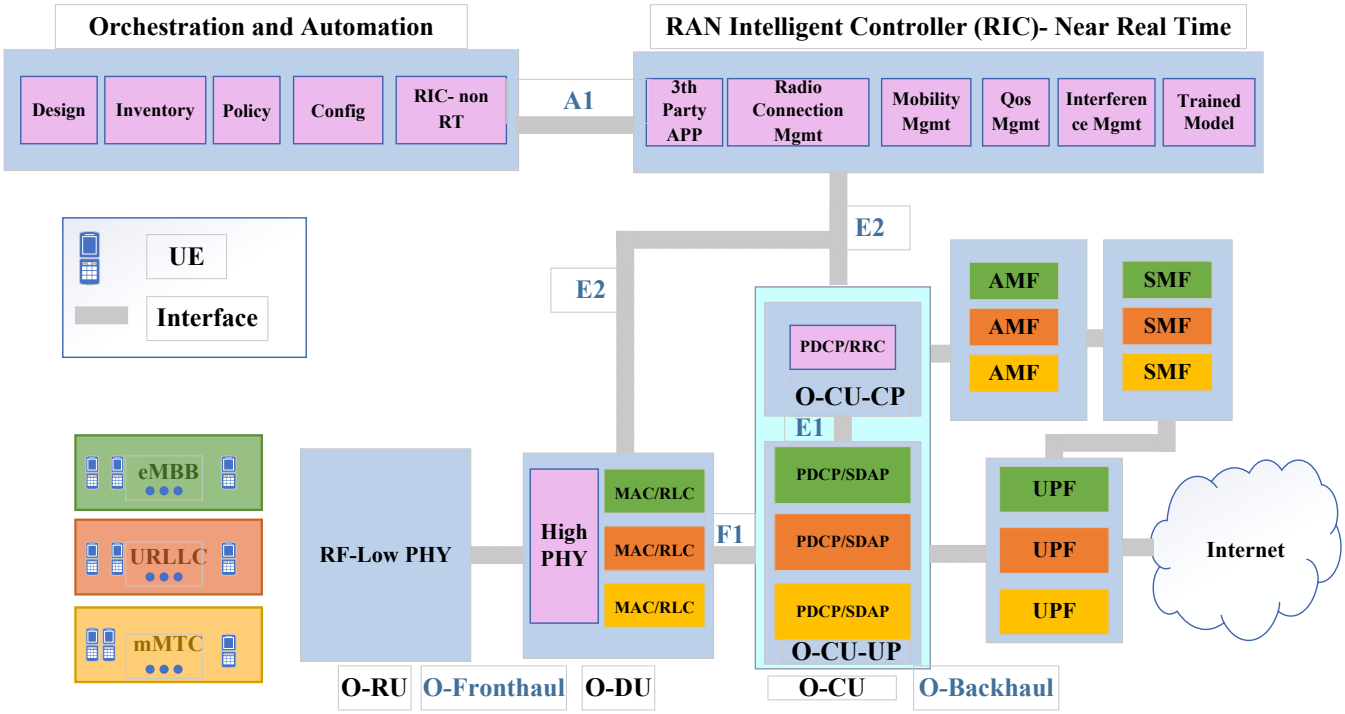

Fig. 1. Network sliced ORAN system

data rate for the i^{th} 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

$$\begin{aligned}
\mathcal{R}^k_{r,u(s_j,i)} &= B\log_2(1+\rho^k_{r,u(s_j,i)} - \zeta^k_{u(s_j,i)})e^k_{u(s_j,i)},\\
\mathcal{R}^r_{u(s_j,i)} &= \sum_{k=1}^{K} B(\log_2(1+\rho^k_{u(s_2,i)}) - \zeta^k_{u(s_j,i)})e^k_{u(s_j,i)}\\
\mathcal{R}_{u(s_j,i)} &= \sum_{r=1}^{R} \mathcal{R}^r_{u(s_j,i)} g^r_{u(s_j,i)}
\end{aligned} \tag{6}$$

Where $j \in \{2,3\}$. Also we have

$$\zeta^k_{u(s_j,i)} = log_2(e)Q^{-1}(\epsilon)\sqrt{\frac{\mathfrak{C}^k_{u(s_j,i)}}{N^k_{u(s_j,i)}}} \tag{7}$$

Where, ϵ is the transmission probability, Q^{-1} is the inverse of Q- function (Gaussian), $\mathfrak{C}^k_{u(s_j,i)} = 1 - \frac{1}{(1+\rho^k_{u(s_j,i)})^2}$ depicts the channel dispersion of UE i at slice s_j, experiencing PRB k and $N^k_{u(s_j,i)}$ represents the blocklength of it. $\mathcal{R}^{e,r}_{u(s_j,i)}$ 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^{max}_s, an upper bound $\bar{I}^k_{r,u(s,i)}$ is obtained for $I^k_{r,u(s,i)}$. Therefore, $\bar{\mathcal{R}}_{u_{(s,i)}} \forall s, \forall i$ is derived by using $\bar{I}^k_{r,u(s,i)}$ instead of $I^k_{r,u(s,i)}$ in (6) and (5).

C. Power of O-RU and Fronthaul Capacity

Let P_r denote the power of the transmitted signal from the r^{th} O-RU to UEs served by it. From (4), we have,

$$P_r = \sum_{s=1}^{S}\sum_{k=1}^{K_s}\sum_{i=1}^{U_s} |\mathbf{w}^k_{r,u(s,i)}|^2 p^k_{r,u(s,i)} g^r_{u(s,i)} e^k_{r,u(s,i)} + \sigma_q^2. \tag{8}$$

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 r^{th} O-RU is formulated as

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

Where, $\alpha^k_{r,u(s,i)} = p^k_{r,u(s,i)} g^r_{u(s,i)} e^k_{r,u(s,i)}$ 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}.

$$\begin{aligned}
T^{tot} &= T^{proc} + T^{tr} + T^{pro}\\
T^{proc} &= T^{RU} + T^{DU} + T^{CU} + T^{UPF}\\
T^{tr} &= T^{fr,t} + T^{mid,t} + T^{b,t}\\
T^{pro} &= T^{fr,p} + T^{mid,p} + T^{b,p}
\end{aligned} \tag{10}$$

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. So it has a constant value

based on the length of the fiber link ($T = L/c$, where L is the length of the link and c is the speed of signal). So the total propagation delay (T^{pro}) is the sum of the propagation delay in the fronthaul $T^{fr,p}$, the midhaul $T^{mid,p}$, and the backhaul $T^{b,p}$. Also, the transmission delay is the amount of time required to push all the packets into the fiber link. ($T = \frac{\alpha}{R}$ Where R is the rate of transmission in each link and α is the mean arrival data rate of each link which is constant in this model.) So the total transmission delay (T^{tr}) is the sum of the transmission delay in the fronthaul $T^{fr,t}$, the midhaul $T^{mid,t}$, and the backhaul $T^{b,t}$. Here we assume the propagation delay and the transmission delay are negligible compared to the processing delay.

$$T^{tot} \approx T^{proc} \tag{11}$$

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 $\alpha_s^U = \sum_{u=1}^{U_s} \lambda_{u(s,i)}$. Assume the mean arrival data rate of the UPF layer for the 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 s^{th} slice in the O-DU, the O-CU, and the UPF is modeled as M/M/1 queue, is formulated as follows, respectively.

$$\begin{aligned} T_s^{DU} &= \frac{1}{\mu_s^d - \alpha_s/M_s^d}, \\ T_s^{CU} &= \frac{1}{\mu_s^c - \alpha_s/M_s^c} \\ T_s^{UPF} &= \frac{1}{\mu_s^u - \alpha_s/M_s^u} \end{aligned} \tag{12}$$

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 $i \in \{d, c, u\}$.

In addition, $T_{u(s,i)}^{RU}$ is the mean transmission delay of i^{th} UE in s^{th} service on the wireless link. The arrival data rate of wireless link for each UE i in 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.

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

$$T_{u(s,i)}^{RU} = \frac{1}{R_{u(s,i)} - \lambda_{u(s,i)}}; \tag{13}$$

So the mean processing delay for each UE i in slice s is

$$T_{u(s,i)}^{proc} = T_{u(s,i)}^{RU} + T_s^{DU} + T_s^{CU} + T_s^{UPF} \tag{14}$$

Hence, $T_{u(s,i)}^{tot} \approx T_{u(s,i)}^{proc}$

E. VNF Power

Assume the power consumption of the baseband processing, the VNFs cost of a slice s is depicted as ϕ_s. So the system's total cost of energy of all slices can be represented as follows.

$$\phi_{tot} = \textstyle\sum_{s=1}^{S} \phi_s.$$

Where, ϕ_s is obtained from below

$$\phi_s = M_s^u \phi_s^u + M_s^c \phi_s^c + M_s^d \phi_s^d \tag{15}$$

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.

F. Problem Statement

Suppose each slice s has priority factor δ_s where $\sum_{s=1}^{S} \delta_s = 1$. This paper aims to maximize the sum rate of all UEs with the presence of constraints as follows.

$$\max_{\boldsymbol{P},\boldsymbol{E},\boldsymbol{M},\boldsymbol{G}} \quad \sum_{s=1}^{S} \sum_{i=1}^{U_s} \delta_s \bar{\mathcal{R}}_{u_{(s,i)}} \tag{16a}$$

$$\text{subject to} \quad P_r \le P_r^{max} \quad \forall r \tag{16b}$$

$$p_{r,u(s,i)}^k \ge 0 \quad \forall i, \forall r, \forall s, \forall k, \tag{16c}$$

$$p_{r,u(s,i)}^k \le P_s^{max} \quad \forall i, \forall r, \forall s, \forall k, \tag{16d}$$

$$\bar{\mathcal{R}}_{u_{(s,i)}} \ge \mathcal{R}_s^{min} \quad \forall s, \tag{16e}$$

$$C_r \le C_r^{max} \quad \forall r, \tag{16f}$$

$$T_{u(s,i)}^{tot} \le T_s^{max} \quad \forall i, \forall s, \tag{16g}$$

$$\mu_s \ge \alpha_s / M_s \quad \forall s, \tag{16h}$$

$$\bar{\mathcal{R}}_{u_{(s,i)}} \ge \lambda_{u_{(s,i)}} \quad \forall i, \forall s, \tag{16i}$$

$$0 \le M_s \le M_s^{max} \quad \forall s, \tag{16j}$$

$$\sum_{r} g_{u(s,i)}^r = 1 \quad \forall s, \forall i, \tag{16k}$$

$$\sum_{k=1}^{K_s} g_{u(s,i)}^r e_{r,u(s,i)}^k \ge 1 \quad \forall s, \forall i, \forall r \tag{16l}$$

$$\sum_{s=1}^{S} \sum_{i=1}^{U_s} g_{u(s,i)}^r e_{r,u(s,i)}^k \le 1 \quad \forall s, \forall i, \forall r \tag{16m}$$

$$\phi^{tot} \le \phi^{max}, \tag{16n}$$

$$g_{u(s,i)}^r \in \{0, 1\} \quad \forall s, \forall i, \tag{16o}$$

$$e_{r,u(s,i)}^k \in \{0, 1\} \quad \forall s, \forall i, \tag{16p}$$

where $\boldsymbol{P} = [p_{r,u(s,i)}^k], \quad \forall s, \forall i, \forall r, \forall k$, is the matrix of power for UEs, $\boldsymbol{E} = [e_{r,u(s,i)}^k], \quad \forall s, \forall i, \forall r, \forall k$ indicate the binary variable for PRB association. Moreover,

$\boldsymbol{G} = [g^r_{u(s,i)}],\ \forall s, \forall i, \forall r$ is a binary variable for O-RU association. Furthermore, $M = [M^d_s, M^c_s, M^u_s],\ \forall s$ is the matrix that shown the number of VNFs in each layer of slice. (16b), (16c) and (16d), 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, (16e) shows that the rate of each UE requesting each type of service (eMBB, URLLC, and mMTC) is more than a threshold, respectively. (16f) and (16g) expressed the limited fronthaul capacity and the limited end-to-end delay of the received signal, respectively. (16h) and (16i) denoted the stability of the M/M/1 queue model. (16j) restricted the number of VNF in each slice due to the limited resources. (16k) and (16l) guarantee that O-RU and PRB are associated with the UE, respectively. Also, (16m) ensures that each PRB can not be assigned to more than one UE associated with the same O-RU. In addition, (16n) indicates that the fixed cost of energy of VNFs in each slice does not exceed the threshold. Moreover, (16o) and (16p) depict that $\boldsymbol{E}$ and $\boldsymbol{G}$ are matrix of binary variables.

III. Proposed Algorithm Scheme

In this section, we first apply some simplifications to the system; Solving the problem (16) is complicated since this problem is non-convex and it is a mixed-integer non-linear problem (MINLP) with a binary variable and an integer variable. We applied some simplifications and used an iterative heuristic algorithm to solve the problem. We solve this problem in two-level iteratively until it converges [32].

In the first level, the principal 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 want to obtain the variables $\boldsymbol{P}, \boldsymbol{E}$, and $\boldsymbol{M}$. Despite the simplification of the problem (16), it is still NP-hard and challenging to solve. Therefore, we relax the variable $\boldsymbol{E}$ [18], [32] and reformulating the constraint (16g), 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 the (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 the (6) and (5) and make these equations become concave functions.

Suppose $\hat{\rho}^k_{r,u(s,i)} = \frac{|P^{max}_s \mathbf{h}^{H\,k}_{r,u(s,i)} \mathbf{w}^k_{r,u(s,i)} g^r_{u(s,i)}|^2}{BN_0}$. we replace $\rho^k_{r,u(s,i)}$ with $\hat{\rho}^k_{r,u(s,i)}$ in the (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.

$$
\begin{aligned}
\bar{\mathcal{R}}^r_{u(s_j,i)} &= \sum_{k=1}^{K_{s_j}} B(\log_2(1+\bar{\rho}^k_{u(s_2,i)}) - \hat{\zeta}^k_{u(s_j,i)}) e^k_{u(s_j,i)} \\
\bar{\mathcal{R}}_{u(s_j,i)} &= \sum_{r=1}^{R} \bar{\mathcal{R}}^r_{u(s_j,i)} \\
\hat{\zeta}^k_{u(s_j,i)} &= log_2(e) Q^{-1}(\epsilon) \sqrt{\frac{\hat{\mathfrak{C}}^k_{u(s_j,i)}}{N^k_{u(s_j,i)}}}) \\
\hat{\mathfrak{C}}^k_{u(s_j,i)} &= 1 - \frac{1}{(1+\hat{\rho}^k_{u(s_j,i)})^2}
\end{aligned}
\tag{17}
$$

Consider UPF, O-CU and O-DU use the processors with the same processing capability. (for simplification), so 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 approximately equal to the mean arrival data rate of the O-CU-UP layer (α^C_s) and O-DU (α^D_s). so $\alpha_s = \alpha^U_s \approx \alpha^C_s \approx \alpha^D_s$. Moreover, the given assumption leads to have same processing power for each layer $\phi^u_s = \phi^c_s = \phi^d_s$. As a result of these assumption, we can assume that $M_s = M^u_s = M^c_s = M^d_s$ Using the above assumption, we have $T^{DU}_s = T^{CU}_s = T^{UPF}_s$

$$
\begin{aligned}
T^{proc}_s &= T^{RU}_s + T^{DU}_s + T^{CU}_s + T^{UPF}_s \\
T^{proc}_s &= T^{RU}_s + 3 \times T^{DU}_s.
\end{aligned}
\tag{18}
$$

Lemma 1. *In the problem* (16)*, the constraint* (16g) *can be reformulated as below* $\forall i, \forall s$

$$
\begin{aligned}
T^{max}_s &\geq \frac{1}{R_{u(s,i)} - \lambda_{u(s,i)}} + \frac{3}{\mu_s - \alpha_s/M_s} \\
M_s &\geq \frac{\alpha_s (T^{max}_s R_{u(s,i)} - T^{max}_s \lambda_{u(s,i)} - 1)}{(T^{max}_s \mu_s - 3)(R_{u(s,i)} - \lambda_{u(s,i)}) - \mu_s}
\end{aligned}
\tag{19}
$$

Also from equation (16n)*,* (16h) *and* (16j) *we have*

$$
\alpha_s/\mu_s \leq M_s \leq \min\{M^{max}, \phi_{max}/3\phi_s\} \tag{20}
$$

We denote $\mathfrak{M}_s = \min\{M^{max}, \phi_{max}/3\phi_s\}$*. Thus, if we restrict constraint* (16g) *to equality, constraint* (16g) *is still valid. Also, we have the following inequality.*

$$
\alpha_s/\mu_s \leq \frac{\alpha_s (T^{max}_s R_{u(s,i)} - T^{max}_s \lambda_{u(s,i)} - 1)}{(T^{max}_s \mu_s - 3)(R_{u(s,i)} - \lambda_{u(s,i)}) - \mu_s} \leq \mathfrak{M}_s \tag{21}
$$

In equation (21)*,* $0 \leq \frac{\alpha_s (T^{max}_s R_{u(s,i)} - T^{max}_s \lambda_{u(s,i)} - 1)}{(T^{max}_s \mu_s - 3)(R_{u(s,i)} - \lambda_{u(s,i)}) - \mu_s}$ *is established due to the fact that the numerator and the denominator will both have the same sign. In the numerator, according to the* (16i)*,* $R_{u(s,i)} - \lambda_{u(s,i)} \geq 0$*, 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* $(R_{u(s,i)} - \lambda_{u(s,i)}) T^{max}_s \geq 1$*, the numerator will be positive.* $(R_{u(s,i)} - \lambda_{u(s,i)}) T^{max}_s \geq 1$ *since the order of* T^{max}_s *is about milli second and the difference between achievable rate and packet rate can be more than* $1/T^{max}_s$*.*

Therefore, to ensure that this constraint will be valid, we restrict constraint (16i) *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 (21), *leads to* $R_{u(s,i)} \geq \lambda_{u(s,i)}$ *that is the constraint* (16i). *For the right side, by reformulating the equation* (21), *we have a new constraint* $\forall i, \forall s$ *as below,*

$$\begin{aligned}
\mathcal{R}_{u(s,i)} &\geq \frac{\mathfrak{M}_s \wr_{u(s,i)} - \alpha_s(T_s^{max}\lambda_{u(s,i)} + 1)}{\mathfrak{M}_s(T_s^{max}\mu_s - 3) - \alpha_s T_s^{max}}, \\
\wr_{u(s,i)} &= (T_s^{max}\mu_s - 3)\lambda_{u(s,i)} + \mu_s, \\
\varpi_{u(s,i)} &= \frac{\mathfrak{M}_s \wr_{u(s,i)} - \alpha_s(T_s^{max}\lambda_{u(s,i)} + 1)}{\mathfrak{M}_s(T_s^{max}\mu_s - 3) - \alpha_s T_s^{max}}, \\
\mathcal{R}_{u(s,i)} &\geq \varpi_{u(s,i)}.
\end{aligned} \tag{22}$$

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 follow.*

$$M_s = \max\{\mathtt{M}_{u(s,i)} | i \in 1, 2, ..., U_s\} \quad \forall s. \tag{23}$$

Despite simplifying the problem (16), it is still non-convex and hard to be solved. So 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 (16) 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 (16f) is changed to $P_r \leq \sigma_{q_r}^2 \times 2^{C_{max}^r}$. The combination of equations (16b) and (16f) leads to the following equation

$$\begin{aligned}
\zeta_r &= min\{P_{max}, \sigma_{q_r}^2 \times 2^{C_{max}^r}\}, \\
P_r &\leq \zeta_r.
\end{aligned} \tag{24}$$

Moreover, the combination of equations (16e), (16i) and (22) leads to the following equation

$$\begin{aligned}
\eta_{u_{(s,i)}} &= max\{\mathcal{R}_{u_{(s,i)}}^{min}, \lambda_{u_{(s,i)}} + 1/T_{max}^s, \varpi_{u(s,i)}\}, \\
\bar{\mathcal{R}}_{u_{(s,i)}} &\geq \eta_{u_{(s,i)}}.
\end{aligned} \tag{25}$$

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

The Lagrangian function is written as follow

$$\mathcal{L}(\boldsymbol{P}, \boldsymbol{E}; \mathfrak{v}, \boldsymbol{\chi}, \mathfrak{h}, \boldsymbol{\xi}, \boldsymbol{\kappa}, \mathfrak{m}) = \sum_{s=1}^{S}\sum_{i=1}^{U_s} \delta_s \bar{\mathcal{R}}_{u_{(s,i)}} \tag{26a}$$

$$+ \sum_{s=1}^{S}\sum_{i=1}^{U_s} \mathfrak{h}_{u_{(s,i)}}(\bar{\mathcal{R}}_{u_{(s,i)}} - \eta_{u_{(s,i)}}) \tag{26b}$$

$$- \sum_{r=1}^{R} \mathfrak{m}_r(P_r - \zeta_r) \tag{26c}$$

$$+ \sum_{s=1}^{S}\sum_{i=1}^{U_s}\sum_{k=1}^{K}\sum_{r=1}^{R} \kappa_{r,u(s,i)}^k p_{r,u(s,i)}^k \tag{26d}$$

$$+ \sum_{s=1}^{S}\sum_{i=1}^{U_s}\sum_{k=1}^{K}\sum_{r=1}^{R} \mathfrak{q}_{r,u(s,i)}^k (P_s^{max} - p_{r,u(s,i)}^k) \tag{26e}$$

$$+ \sum_{r=1}^{R}\sum_{s=1}^{S}\sum_{i=1}^{U_s} \chi_{r,u(s,i)} (\sum_{k=1}^{K_s} e_{r,u(s,i)}^k - 1) \tag{26f}$$

$$- \sum_{s=1}^{S}\sum_{i=1}^{U_s}\sum_{k=1}^{K}\sum_{r=1}^{R} \mathfrak{v}_{r,u(s,i)}^k (e_{r,u(s,i)}^k - 1) \tag{26g}$$

$$+ \sum_{s=1}^{S}\sum_{i=1}^{U_s}\sum_{k=1}^{K}\sum_{r=1}^{R} \xi_{r,u(s,i)}^k e_{r,u(s,i)}^k. \tag{26h}$$

Lemma 2. *The derivatives of the Lagrangian function* (26) *with respect to the* $\boldsymbol{P}$ *and* $\boldsymbol{E}$ *give the Karush-Kuhn-Tucker (KKT) conditions to obtain the optimal value of these two variables [18], [32].*

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

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial p_{r,u(s,i)}^k} &= (\delta_s + \mathfrak{h}_{u_{(s,i)}})\mathfrak{B}_{r,u(s,i)}^k \\
&\quad + (\mathfrak{s}_{r,u(s,i)}^k - \mathfrak{D}_{r,u(s,i)}^k) = 0
\end{aligned} \tag{27}$$

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 follow.*

$$\begin{aligned}
\mathfrak{D}_{r,u(s,i)}^k &= \mathfrak{m}_r |\mathbf{w}_{r,u(s,i)}^k|^2 g_{u(s,i)}^r e_{r,u(s,i)}^k, \\
\mathfrak{B}_{r,u(s,i)}^k &= \frac{B|\mathbf{h}_{r,u(s,i)}^{H\,k} \mathbf{w}_{r,u(s,i)}^k|^2 g_{u(s,i)}^r e_{r,u(s,i)}^k}{\ln(2)} \mathfrak{S}_{r,u(s,i)}^k, \\
\mathfrak{S}_{r,u(s,i)}^k &= \frac{1}{|\mathbf{h}_{r,u(s,i)}^{H\,k} \mathbf{w}_{r,u(s,i)}^k|^2 \mathfrak{k}_{r,u(s,i)}^k + BN_0 + I_{r,u(s,i)}^k}.
\end{aligned} \tag{28}$$

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* (27), *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.*

$$\begin{aligned}
p_{r,u(s,i)}^k = [&\frac{(\delta_s + \mathfrak{h}_{u_{(s,i)}})B\mathfrak{j}_{r,u(s,i)}^k}{Ln2 \times (\kappa_{r,u(s,i)}^k - \mathfrak{D}_{r,u(s,i)}^k)} \\
&- \frac{BN_0 + I_{r,u(s,i)}^k}{|\mathbf{h}_{r,u(s,i)}^{H\,k} \mathbf{w}_{r,u(s,i)}^k|^2 \mathfrak{j}_{r,u(s,i)}^k}]^+.
\end{aligned} \tag{29}$$

Also $[a]^+ = \max(0, a)$. *In addition, PRB assignment can be achieved from the derivatives of the Lagrangian function* (26) *with respect to the* $\boldsymbol{E}$ *as follow.*

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial e_{r,u(s,i)}^k} &= \bar{\mathcal{R}}_{r,u(s,i)}^k (\delta_s + \mathfrak{h}_{u_{(s,i)}}) \\
&\quad - \mathfrak{m}_r |\mathbf{w}_{r,u(s,i)}^k|^2 p_{r,u(s,i)}^k g_{u(s,i)}^r \\
&\quad + (\xi_{r,u(s,i)}^k - \mathfrak{v}_{r,u(s,i)}^k + \chi_{r,u(s,i)}) = 0.
\end{aligned} \tag{30}$$

So, the optimal $\boldsymbol{E}$ *is obtained using KKT condition as follow.*

$$\begin{aligned}
e_{r,u(s,i)}^k \times (&\mathfrak{F}_{r,u(s,i)}^k - \mathfrak{v}_{r,u(s,i)}^k \\
&- \mathfrak{m}_r |\mathbf{w}_{r,u(s,i)}^k|^2 p_{r,u(s,i)}^k g_{u(s,i)}^r) = 0.
\end{aligned} \tag{31}$$

Where $\mathfrak{F}^k_{r,u(s,i)} = \bar{\mathcal{R}}^k_{r,u(s,i)}(\delta_s + \mathfrak{h}_{u_{(s,i)}}) + (\xi^k_{r,u(s,i)} + \chi_{r,u(s,i)})$. *Hence, from equation* (30) *and* (31)*, PRB assignment is performed as follow.*

$$e^k_{r,u(s,i)} = \begin{cases} 1 & u(s,i) = argmax \mathfrak{F}^k_{r,u(s,i)} \forall s, \forall r, \forall k \\ 0 & otherwise \end{cases} \quad (32)$$

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

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 $\boldsymbol{P}$ and $\boldsymbol{E}$ are fixed, 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 problem can be reformulated as follow

$$\max_{\boldsymbol{G}} \quad \sum_{s=1}^{S} \sum_{i=1}^{U_s} \sum_{r=1}^{R} \delta_s g^r_{u(s,i)} \bar{\mathcal{R}}^r_{u_{(s,k)}} \quad (33a)$$

$$\text{subject to} \quad \sum_{s=1}^{S} \sum_{i=1}^{U_s} g^r_{u(s,i)} \psi_{r,u(s,i)} \leq \mathfrak{t}_r \quad \forall r \quad (33b)$$

$$\sum_{r} g^r_{u(s,i)} = 1 \quad \forall s, \forall i, \quad (33c)$$

$$g^r_{u(s,i)} \in \{0,1\} \quad \forall s, \forall i, \quad (33d)$$

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 (24) and (8). Since we obtained (25) in (III-A), we can ignore this constraint in (33). The problem (33) 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], [33]. Firstly, we set all the variables to zero ($g^r_{u(s,i)} = 0, \quad \forall s, \forall i, \forall r$). 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 ($\mathfrak{B}^{rem}_{u_{(s,i)}} = \mathcal{R}, \forall s, \forall i$). 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 ($\mathfrak{B}^{rem}_{u_{(s,i)}} = \mathfrak{B}^{rem}_{u_{(s,i)}} \setminus \{r^*\}$). 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 $O(Slog(S))$. 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)

Set $g^r_{u(s,i)} = 0, \quad \forall s, \forall i, \forall r.$
Set $\mathfrak{C}_r = \mathfrak{t}_r, \forall r$
Set $\mathfrak{B}^{rem}_{u_{(s,i)}} = \mathcal{R}\ \forall s, \forall i$
Sort slices according to their priority factor (δ_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

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 $\boldsymbol{G}$ 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). The number of UEs are $\mathfrak{N} = \sum_{s=1}^{S} \sum_{i=1}^{U_S} 1$.

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

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

As we mentioned before, the complexity order of the first

TABLE I
SIMULATION PARAMETER

Parameter	Value
Noise power	-174dBm
Bandwidth	180 KHz
Maximum transmit Power of each O-RU	38dBm
Maximum delay for eMBB	4msec
Maximum delay for URLLC	1msec
Maximum delay for mMTC	5msec
Maximum fronthaul capacity	200 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

sub-problem is about $O(N_u \times K)$ 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 (16) is $O(\mathfrak{N} \times K \times (Slog(S) + R\mathfrak{N}))$.

IV. NUMERICAL RESULTS

In this section, numerical results for the main problem are depicted to evaluate the performance of the algorithms. We consider three network slices, for eMBB, URLLC and mMtc services. Assume we have six 4-antenna O-RU (MISO) located in a cell with a diameter of 500 meters. We consider 25 PRB in the network. 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 $\lambda = 0.2Mbps$. The other parameters of these simulations are depicted in Table I. Two different methods are used to compare with the performance of the proposed method (IABV). The first one is a baseline scheme, which used random PRB allocation. The association of O-RU is based on the distance, channel quality, and the fronthaul capacity. Also, the number of activated VNF for each slice is fixed. The second one is similar to the FBDR algorithm proposed in [18]. In this method, PRB and power are dynamically allocated, and the number of VNFs is obtained from the simulation, and the UEs are associated with the O-RU based on the quality of their channels, the fronthaul capacity, and channel distance.

In Fig. 2 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. Here, we did not consider the priority. The figure presented that the proposed method (IABV) is 18.6% higher 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 3, depicts the number of activated VNF for the five different mean service times of one URLLC service vs. the mean arrival time for 12 UEs. This figure presented that as the mean arrival rate increases, the number of activated

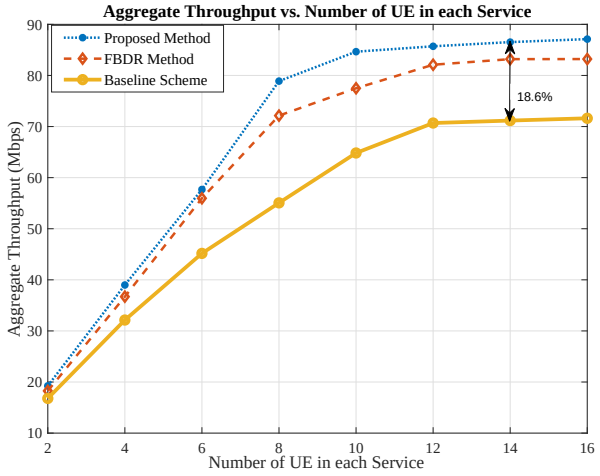

Fig. 2. Aggregate Throughput vs. Number of UE in each Service

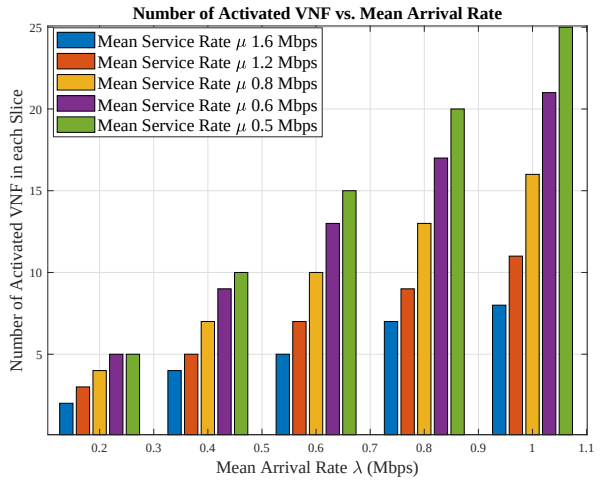

Fig. 3. Number of activated VNF in each Service vs. Mean Arrival Rate(Mbps)

VNF increases. Moreover, the number of activated VNFs decreased when the mean service rate increases.

In figure 4, 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 5, 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

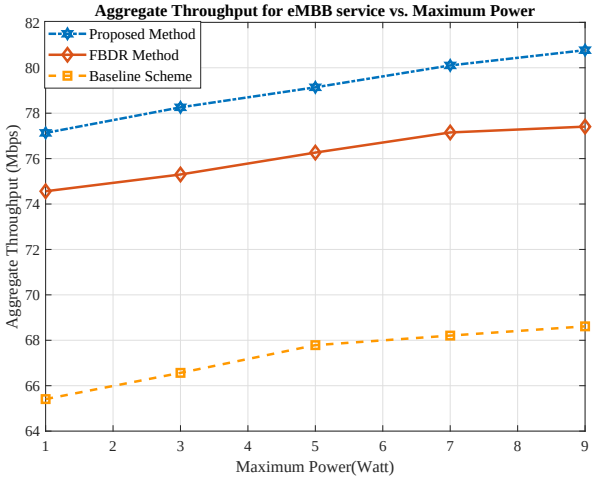

Fig. 4. Aggregate Throughput for eMBB vs. Maximum Transmit power for various number of UEs

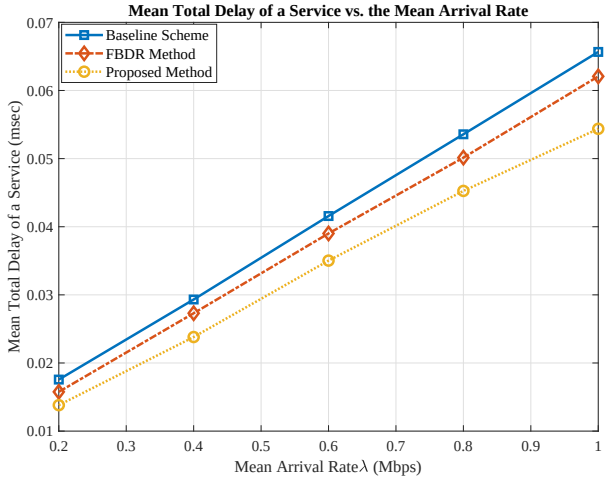

Fig. 6. Mean Total Delay of a URLLC Service vs. the Mean Arrival Rate of a UE in the Service for different methods

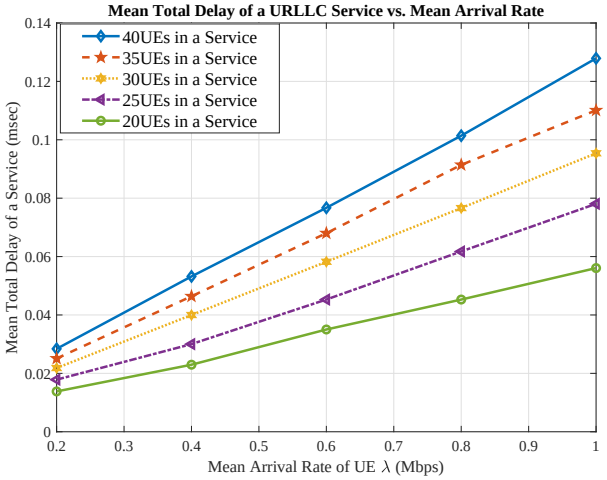

Fig. 5. Mean Total Delay of a URLLC Service vs. the Mean Arrival Rate of a UE in the Service for various number of UEs

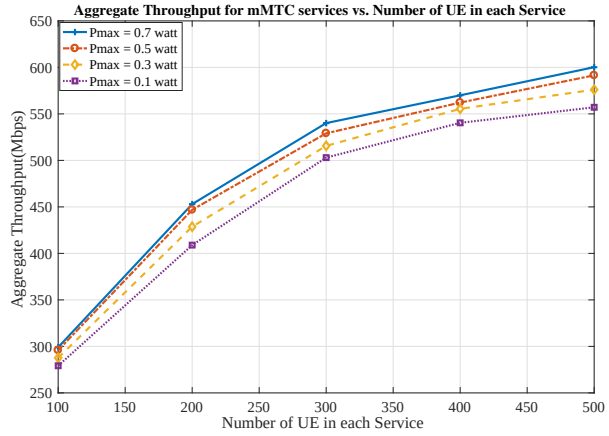

Fig. 7. Aggregate Throughput for mMTC services vs. Number of UE in each Service for three different mMTC service instances

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 6 is the same as figure 5 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 7, 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 $1bits/sec/Hz$ data rate and is not sensitive to the end-to-end delay. 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 8, 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 figure 8 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.

V. Conclusion

This paper proposed the downlink of the O-RAN system using network slicing for different types of 5G services (eMBB, mMTC, and URLC). The isolation of various

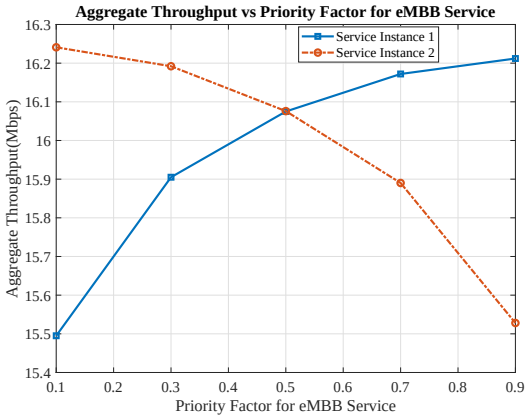

Fig. 8. Aggregate Throughput for two eMBB service instances vs. Priority of the first service instance

types of services (eMBB, mMTC, and URLC) in the O-DU, the O-CU, and the user plane function (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 the 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 obtain PRB association and power allocation using the Lagrangian method. Then in the second step, the O-RU association is acquired. The performance of our proposed method (IABV) is compared with the baseline scheme and FBDR in [18]. Also, we assume distinct scenarios for each service (eMBB, URLLC, and mMTC) based on their requirement QoS. Simulation results depict that the proposed method (IABV) achieved 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, 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 *GLOBECOM 2020-2020 IEEE Global Communications Conference*. IEEE, 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, 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, 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)*. IEEE, 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*, 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, 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, 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*, 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, 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*, 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., 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, 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, 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, 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, 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, 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)*. IEEE, 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, 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, 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, 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, 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*, 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, 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, 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, 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, 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)*. IEEE, 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, 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*, pp. 1–15, 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, 2018.

[33] 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, 2007.