Resource Allocation in an Open RAN System using Network Slicing

Mojdeh Karbalaee Motalleb
School of ECE, College of Engineering, University of Tehran, Iran
Email: {mojdeh.karbalaee}@ut.ac.ir,

***Abstract*—Taking the advantage of both virtual RAN (v-RAN) and Cloud RAN (C-RAN), Open RAN (O-RAN) is introduced as a next generation of RAN systems which leads to increase the flexibility, Openness and reduce operator costs and allow 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 type 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 type of services in O-DU, O-CU and user plane function (UPF) is performed. Limited fronthaul capacity and the restriction of end-to-end delay is considered in the problem. 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 that is tremendously difficult. To solve the challenging problem, we broke it down into two different steps that is solved by iterative algorithm. 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, O-RU association is carried out. The proposed method is confirmed by the simulation results in a way that the simulations illustrate higher achievable data rate compared with a baseline scheme which only optimizes one of the baseband resources.**

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

I. Introduction

Nowadays, RAN virtualization attracts a lot of attention from operators since it has significant benefits which leads to increase the flexibility and reduce operator costs such as CAPEX and OPEX and also 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 Open Radio Access Network (O-RAN) Alliance to establish O-RAN as a 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 the advantage of both. CRAN, divides RAN into two parts radio remote head (RRH) and base band unit (BBU). More than one distributed RRHs can be connected to a centralized BBU which is named BBU-pool [1]. Unlike previous generation of RAN that divide 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 contains RF and lower 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 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 to each other via an open and well-defined interface F_1. Moreover, O-DU is connected to 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 new generation of wireless networks is its intelligence. Based on the requirement of smart 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 [2]–[7].

In the concept of network slicing, a slice is an end-to-end logical network which offers services with special needs. Multiple isolated network slices run and manage and work independently on the same infrastructure. There are several implementations of network slicing, including network 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 special slice of network based on its QoS [8], [9].

Some O-RAN components that includes 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.

In this paper, as depicted in Figure 1, the downlink of the ORAN system is studied.

II. System Model and Problem Formulation

In this section, first, 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 about 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.

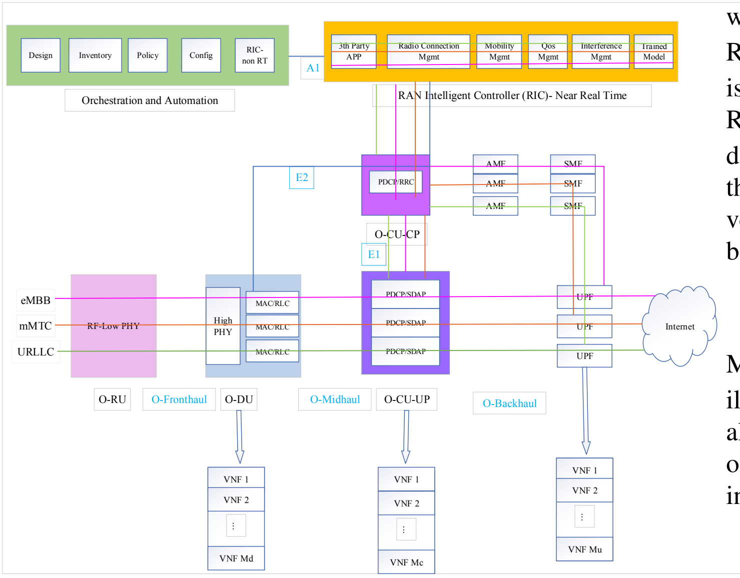

Fig. 1: Network sliced ORAN system

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 served by its corresponding slice.

Each Service $s_j \in \{1, 2, ..., S_j\}$ consists of U_s request from the single-antenna UEs which require certain QoS to be able to use the requested program($j \in \{1, 2, 3\}$ indicate service type). There are different application request which fall into one of these service categories. Each application request requires specific QoS. Based on the request for the application and QoS, 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 the all UE in each service. Also, each VNF instance is running on the virtual machine (VM) that are using resources from the data centers. Each VM, requires enough resources of CPU, memory, storage and network bandwidth.

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 the all PRBs.

B. The Achievable Rate

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

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

where $p_{r,u(s,i)}^k$ represents the transmission power from O-RU r to i^{th} 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 i^{th} 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 i^{th} UE in s^{th} slice that is the zero forcing beamforming vector to minimize the interference which is indicated as below

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

Moreover, $g_{u(s,i)}^r \in \{0, 1\}$ 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 $I_{r,u(s,i)}^k$ is the power of interfering signals represented as follow

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

where $\gamma_1 = e_{u(s,i)}^k e_{u(s,l)}^k$ and $\gamma_2 = e_{u(s,i)}^k e_{u(n,l)}^k$. $e_{u(s,i)}^k$ 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 user i in s^{th} service

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

where $\mathfrak{y}_{r,u(s,i)}^k = \boldsymbol{w}_{r,u(s,i)}^k p_{r,u(s,i)}^{k\,\frac{1}{2}} x_{u(s,i)} + \boldsymbol{q}_r$ 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, 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.

$$\begin{aligned}
\mathcal{R}_{r,u(s_1,i)}^k &= B \log_2(1 + \rho_{r,u(s_1,i)}^k), \\
\mathcal{R}_{u(s_1,i)}^r &= \sum_{k=1}^{K} B \log_2(1 + \rho_{r,u(s_1,i)}^k e_{r,u(s_1,i)}^k), \\
\mathcal{R}_{u(s_1,i)} &= \sum_{r=1}^{R} \mathcal{R}_{u(s_1,i)}^r g_{u(s_1,i)}^r
\end{aligned} \tag{5}$$

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 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 follow

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

Where $j \in \{1,2\}$. 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 rate of each O-RU r to UE i in slice s_j.

If we replace $p^k_{u(s,l)}$ and $p^k_{u(n,l)}$ in (3) by P_{max}, an upper bound $\bar{I}^k_{r,u(s,i)}$ is obtained for $I^k_{r,u(s,i)}$. Therefore, $\bar{\mathcal{R}}_{u_{(s,i)}} \forall s, \forall i$ 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 transmitted signal from the r^{th} O-RU to UEs served by it. From (4), we have,

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

Since we have 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

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

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 $\sigma^2_{q_r}$ is the power of quantization noise.

D. Mean Delay

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

$$\begin{aligned}
T_{tot} &= T_{process} + T_{transmission} + T_{propagation}\\
T_{process} &= T_{RU} + T_{DU} + T_{CU} + T_{UPF}\\
T_{transmission} &= T_{front} + T_{mid} + T_{back} + T_{trans2net}\\
T_{propagation} &= T_{front} + T_{mid} + T_{back} + T_{trans2net}
\end{aligned} \tag{10}$$

Total delay is sum of processing delay, transmission delay and propagation delay. The propagation delay is the time takes for a signal to reach to its destination. So it has a constant value based on the length of fiber link ($T = L/c$, where L is the length of link and c is the speed of signal). 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 the each link which is constant in this model.) Here we assume the value of propagation delay and transmission is negligible compared to the rest.

$$T_{tot} \approx T_{process} \tag{11}$$

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} slice. Therefore, the mean arrival data rate of the s^{th} slice in the UPF layer is $\alpha^1_s = \sum_{u=1}^{U_s} a_{u(s,i)}\lambda_{u(s,i)}$, where $a_{u(s,i)}$ is a binary variable which indicates whether the i^{th} UE requested s^{th} service is admitted or not.

Assume the mean arrival data rate of the UPF layer for slice s (α^U_s) is approximately equal to the mean arrival data rate of the O-CU-UP layer (α^C_s) and O-DU (α^D_s). so $\alpha_s = \alpha^U_s \approx \alpha^C_s \approx \alpha^D_s$. since, by using Burkes Theorem, the mean arrival data rate of the second and third layer 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 first and the second layer, modeled as M/M/1 queue, is formulated as follow, respectively

$$\begin{aligned}
T^s_{DU} &= \frac{1}{\mu^d_s - \alpha_s/M^d_s},\\
T^s_{CU} &= \frac{1}{\mu^c_s - \alpha_s/M^c_s}\\
T^s_{UPF} &= \frac{1}{\mu^u_s - \alpha_s/M^u_s}
\end{aligned} \tag{12}$$

Where M^d_s, M^c_s and M^u_s are the variables that depict the sum of VNFs in O-DU, O-CU-UP and UPF, respectively. Moreover, $1/\mu^d_s$, $1/\mu^c_s$ and $1/\mu^u_s$ are the mean service 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^i_s $i \in \{d,c,u\}$.

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

$$T_{process}^{u(s,i)} = T_{RU}^{u(s,i)} + T_{DU}^{s} + T_{CU}^{s} + T_{UPF}^{s} \qquad (14)$$

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

E. VNF Power

Assume the power consumption of baseband processing at each DC d that is connected to VNFs of a slice s is depicted as ϕ_s. So the total power of the system for all active DCs that are connected to slices can be represented as

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

Where, ϕ_s is obtained from below

$$\phi_s = M_s^u \phi_s^u + M_s^c \phi_s^c + M_s^d \phi_s^d \qquad (15)$$

Moreover, ϕ_s^u, ϕ_s^c and ϕ_s^d are the static cost of energy in UPF, O-CU and O-DU, respectively.

F. Problem Statement

Suppose each slice s has priority factor δ_s where $\sum_{s=1}^{S} \delta_s = 1$. The optimization problem is formulated as follow. The aim of this paper is to maximize the sum rate of all UEs with the presence of constraints which is written as follow,

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

$$\text{subject to} \quad P_r \leq P_{max} \quad \forall r \qquad (16b)$$

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

$$\bar{\mathcal{R}}_{u_{(s_j,i)}} \geq \mathcal{R}_{min}^{s_j} \quad \forall s, j \in \{1,2,3\}, \qquad (16d)$$

$$C^r \leq C_{max}^r \quad \forall r, \qquad (16e)$$

$$T_{tot}^{u(s,i)} \leq T_{max}^{s} \quad \forall i, \forall s, \qquad (16f)$$

$$\mu_s \geq \alpha_s / M_s \quad \forall s, \qquad (16g)$$

$$\bar{\mathcal{R}}_{u_{(s,i)}} \geq \lambda_{u_{(s,i)}} \quad \forall i, \forall s, \qquad (16h)$$

$$0 \leq M_s \leq M^{max} \quad \forall s, \qquad (16i)$$

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

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

$$\sum_{s=1}^{S} \sum_{i=1}^{U_s} g_{u(s,i)}^{r} e_{r,u(s,i)}^{k} \leq 1 \quad \forall s, \forall i, \forall r \qquad (16l)$$

$$\phi_{tot} \leq \phi_{max}, \qquad (16m)$$

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

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

where $\boldsymbol{P} = [p_{r,u(s,i)}^{k}]$ $\forall s, \forall i, \forall r, \forall k$, 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 shown the number of VNFs in each layer of slice. (16b), and (16c), indicate that the power of each RU do not exceed the maximum power, and the power of each UE is a positive integer value, respectively. Also (16d) shows that the rate of each UE requesting eMBB, URLLC and mMTC is more than a threshold, respectively. (16e) and (16f) expressed the limited capacity of the fronthaul link, and the limited delay of receiving signal, respectively. (16g) and (16h) denoted the stability of the M/M/1 queue model. (16i) restricted the number of VNF in each slice due to the limited resources. (16j) and (16k) guarantee that O-RU and PRB is associated to the UE, respectively. Also, (16l) ensure that each PRB can not be assigned to more than one UE associated to the same O-RU. In addition, (16m) indicate that the static cost of energy of VNFs in each slice do not exceed from the threshold. Moreover, (16n) and (16o) 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 problem (16) is complicated due to the fact that this problem is a non-convex problem and it is a mixed integer non-linear problem (MINLP) with a binary variable and an integer variable. In the following, we apply the simplifications to reformulate MINL parts and use iterative heuristic algorithm to solve the reformulated problem. We solve this problem in two level iteratively until it converges; In the first level, parameters $(\boldsymbol{P}, \boldsymbol{E}, \boldsymbol{M})$ are obtained by relaxing and reformulating parameters and turn it to convex problem; Afterward we solve it by dual optimization problem. In the second level, finding optimal O-RU association ($\boldsymbol{G}$) is concerned with the fixed parameter of power, PRB allocation and number of 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) by P_{max}, an upper bound $\bar{I}_{r,u(s,i)}^{k}$ for $I_{r,u(s,i)}^{k}$, the lower bound $\bar{\rho}_{u(s,i)}^{k}$ for $\rho_{u(s,i)}^{k}$ and the lower bound $\bar{\mathcal{R}}_{u_{(s,i)}} \forall s, \forall i$ for $\mathcal{R}_{u_{(s,i)}}$ is obtained by replacing with $I_{r,u(s,i)}^{k}$ $\bar{I}_{r,u(s,i)}^{k}$ in (6) and (5) and make them concave.

Suppose $\hat{\rho}_{r,u(s,i)}^{k} = \frac{|P_{max} \mathbf{h}_{r,u(s,i)}^{H\,k} \mathbf{w}_{r,u(s,i)}^{k} g_{u(s,i)}^{r}|^2}{BN_0}$. To convexify (6) (for the short packet transmission), we replace $\rho_{r,u(s,i)}^{k}$ with $\hat{\rho}_{r,u(s,i)}^{k}$ in (7). 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 have the same processor (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 $\alpha_s = \alpha^U_s \approx \alpha^C_s \approx \alpha^D_s$. So the given assumption leads to have same energy for each layer $\phi^u_s = \phi^c_s = \phi^d_s$. As a result of these assumption, for simplicity, we can assume that $M_s = M^u_s = M^c_s = M^d_s$ Using the above assumption, we have $T^s_{DU} = T^s_{CU} = T^s_{UPF}$

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

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

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

Also from equation (16m), (16g) *and* (16i) *we have*

$$0 \leq M_s \leq \min\{M^{max}, \alpha_s/\mu_s, \phi_{max}/3\phi_s\} \tag{20}$$

We denote $\mathfrak{M}_s = \min\{M^{max}, \alpha_s/\mu_s, \phi_{max}/3\phi_s\}$*. Thus, if we restrict* (16f) *to equality we have*

$$0 \leq \frac{\alpha_s(T^s_{max}R_{u(s,i)} - T^s_{max}\lambda_{u(s,i)} - 1)}{(T^s_{max}\mu_s - 3)(R_{u(s,i)} - \lambda_{u(s,i)}) - \mu_s} \leq \mathfrak{M}_s \tag{21}$$

In (21), $0 \leq \frac{\alpha_s(T^s_{max}R_{u(s,i)} - T^s_{max}\lambda_{u(s,i)} - 1)}{(T^s_{max}\mu_s - 3)(R_{u(s,i)} - \lambda_{u(s,i)}) - \mu_s}$ *is established due to the fact that the numerator and the denominator will both have same sign. Using* (16h)*, in numerator,* $\alpha_s \geq 0$, $R_{u(s,i)} - \lambda_{u(s,i)} \geq 0$ *and to simplify the problem, assume* $(R_{u(s,i)} - \lambda_{u(s,i)})T^s_{max} \geq 1$ *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, we restrict constraint* (16h) *to* $R_{u(s,i)} \geq \lambda_{u(s,i)} + 1/T^s_{max}$*. So the numerator is positive. In denominator, it can be said approximately that* $(T^s_{max}\mu_s)(R_{u(s,i)} - \lambda_{u(s,i)}) - \mu_s \geq 0$, *since,* $(R_{u(s,i)} - \lambda_{u(s,i)}) \geq 1/T^s_{max}$ *as mentioned above. Therefore, we just need to have constraint below*

$$\frac{\alpha_s(T^s_{max}R_{u(s,i)} - T^s_{max}\lambda_{u(s,i)} - 1)}{(T^s_{max}\mu_s - 3)(R_{u(s,i)} - \lambda_{u(s,i)}) - \mu_s} \leq \mathfrak{M}_s \tag{22}$$

So by reformulating the equation (22)*, we have a new constraint* $\forall i, \forall s$ *as below,*

$$\begin{aligned}
&\mathcal{R}_{u(s,i)} \geq \frac{\mathfrak{M}_s((T^s_{max}\mu_s - 3)\lambda_{u(s,i)} + \mu_s) - \alpha_s(T^s_{max}\lambda_{u(s,i)} + 1)}{\mathfrak{M}_s(T^s_{max}\mu_s - 3) - \alpha_s T^s_{max}}, \\
&\varpi_{u(s,i)} = \frac{\mathfrak{M}_s((T^s_{max}\mu_s - 3)\lambda_{u(s,i)} + \mu_s) - \alpha_s(T^s_{max}\lambda_{u(s,i)} + 1)}{\mathfrak{M}_s(T^s_{max}\mu_s - 3) - \alpha_s T^s_{max}}, \\
&\mathcal{R}_{u(s,i)} \geq \varpi_{u(s,i)}.
\end{aligned} \tag{23}$$

In addition, we denote $\mathtt{M}_{u(s,i)} = \frac{\alpha_s(T^s_{max}R_{u(s,i)} - T^s_{max}\lambda_{u(s,i)} - 1)}{(T^s_{max}\mu_s - 3)(R_{u(s,i)} - \lambda_{u(s,i)}) - \mu_s}$*. So we have,*

$$M_s = \max\{\mathtt{M}_{u(s,i)} | i \in 1, 2, ..., U_s\} \quad \forall s. \tag{24}$$

Despite simplifying the problem (16), it is still non-convex and hard to be solved. So the simplest approach is to relax $\mathbf{E}$ into continuous value $e^k_{r,u(s,i)} \in [0,1] \, \forall s, \forall i, \forall r, \forall k$ 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 $P_r = \sigma^2_{q_r} \times 2^{C^r}$ so the constraint (16e) is changed to $P_r \leq \sigma^2_{q_r} \times 2^{C^r_{max}}$. The combination of equations (16d) and (16e) leads to the following equation

$$\begin{aligned}
&\zeta_r = min\{P_{max}, \sigma^2_{q_r} \times 2^{C^r_{max}}\}, \\
&P_r \leq \zeta_r.
\end{aligned} \tag{25}$$

Moreover, the combination of equations (16d), (16h) and (23) leads to the following equation

$$\begin{aligned}
&\eta_{u_{(s,i)}} = max\{\mathcal{R}^{max}_{u_{(s,i)}}, \lambda_{u_{(s,i)}} + 1/T^s_{max}, \varpi_{u(s,i)}\}, \\
&\bar{\mathcal{R}}_{u_{(s,i)}} \geq \eta_{u_{(s,i)}}.
\end{aligned} \tag{26}$$

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

The Lagrangian function is written as follow

$$\mathcal{L}(\boldsymbol{P}, \boldsymbol{E}; \boldsymbol{\mathfrak{v}}, \boldsymbol{\chi}, \boldsymbol{\mathfrak{h}}, \boldsymbol{\xi}, \boldsymbol{\kappa}, \boldsymbol{\mathfrak{m}}) = \sum_{s=1}^{S}\sum_{i=1}^{U_s} \delta_s \bar{\mathcal{R}}_{u_{(s,i)}} \tag{27a}$$

$$+ \sum_{s=1}^{S}\sum_{i=1}^{U_s} \mathfrak{h}_{u_{(s,i)}}(\bar{\mathcal{R}}_{u_{(s,i)}} - \eta_{u_{(s,i)}}) \tag{27b}$$

$$- \sum_{r=1}^{R} \mathfrak{m}_r(P_r - \zeta_r) \tag{27c}$$

$$+ \sum_{s=1}^{S}\sum_{i=1}^{U_s}\sum_{k=1}^{K}\sum_{r=1}^{R} \kappa^k_{r,u(s,i)} p^k_{r,u(s,i)} \tag{27d}$$

$$+ \sum_{r=1}^{R}\sum_{s=1}^{S}\sum_{i=1}^{U_s} \chi_{r,u(s,i)}(\sum_{k=1}^{K_s} e^k_{r,u(s,i)} - 1) \tag{27e}$$

$$- \sum_{s=1}^{S}\sum_{i=1}^{U_s}\sum_{k=1}^{K}\sum_{r=1}^{R} \mathfrak{v}^k_{r,u(s,i)}(e^k_{r,u(s,i)} - 1) \tag{27f}$$

$$+ \sum_{s=1}^{S}\sum_{i=1}^{U_s}\sum_{k=1}^{K}\sum_{r=1}^{R} \xi^k_{r,u(s,i)} e^k_{r,u(s,i)}. \tag{27g}$$

Lemma 2. *By taking derivatives of* (27) *(the lagrangian function), with respect to the* $\boldsymbol{P}$ *and the* $\boldsymbol{E}$*, these two variables are obtained. Assume,* $e^k_{r,u(s,i)} = 1$

$$\frac{\partial \mathcal{L}}{\partial p^k_{r,u(s,i)}} = (\delta_s + \mathfrak{h}_{u_{(s,i)}})\mathfrak{B}^k_{r,u(s,i)} + (\kappa^k_{r,u(s,i)} - \mathfrak{m}_r \mathfrak{D}^k_{r,u(s,i)}) = 0 \tag{28}$$

Where

$$\mathfrak{D}^k_{r,u(s,i)} = |\mathbf{w}^k_{r,u(s,i)}|^2 g^r_{u(s,i)} e^k_{r,u(s,i)},$$
$$\mathfrak{B}^k_{r,u(s,i)} = \frac{B|\mathbf{h}^{H\,k}_{r,u(s,i)} \mathbf{w}^k_{r,u(s,i)}|^2 g^r_{u(s,i)} e^k_{r,u(s,i)}}{\ln(2)} \mathfrak{S}^k_{r,u(s,i)},$$
$$\mathfrak{S}^k_{r,u(s,i)} = \frac{1}{|\mathbf{h}^{H\,k}_{r,u(s,i)} \mathbf{w}^k_{r,u(s,i)}|^2 \mathfrak{k}^k_{r,u(s,i)} + BN_0 + I^k_{r,u(s,i)}}. \tag{29}$$

where $\mathfrak{k}^k_{r,u(s,i)} = g^r_{u(s,i)} e^k_{r,u(s,i)} p^k_{r,u(s,i)}$*. Thus, from equation* (28)*, optimal power is obtained and power is allocated. We denote* $\mathfrak{j}^k_{r,u(s,i)} = g^r_{u(s,i)} e^k_{r,u(s,i)}$.

$$p^k_{r,u(s,i)} = \Big[\frac{(\delta_s + \mathfrak{h}_{u_{(s,i)}})B\mathfrak{j}^k_{r,u(s,i)}}{\kappa^k_{r,u(s,i)} - \mathfrak{m}_r \mathfrak{D}^k_{r,u(s,i)}} - \frac{BN_0 + I^k_{r,u(s,i)}}{|\mathbf{h}^{H\,k}_{r,u(s,i)} \mathbf{w}^k_{r,u(s,i)}|^2 \mathfrak{j}^k_{r,u(s,i)}}\Big]^+. \tag{30}$$

Also $[a]^+ = \max(0, a)$*. In addition, PRB assignment is obtained as follow*

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial e^k_{r,u(s,i)}} &= \bar{\mathcal{R}}^k_{r,u(s,i)}(\delta_s + \mathfrak{h}_{u_{(s,i)}}) \\ &- \mathfrak{m}_r |\mathbf{w}^k_{r,u(s,i)}|^2 p^k_{r,u(s,i)} g^r_{u(s,i)} \\ &+ (\xi^k_{r,u(s,i)} - \mathfrak{v}^k_{r,u(s,i)} + \chi_{r,u(s,i)}) = 0. \end{aligned} \tag{31}$$

Using KKT conditions, we have

$$e^k_{r,u(s,i)} \times (\mathfrak{F}^k_{r,u(s,i)} - \mathfrak{v}^k_{r,u(s,i)} - \mathfrak{m}_r |\mathbf{w}^k_{r,u(s,i)}|^2 p^k_{r,u(s,i)} g^r_{u(s,i)}) = 0. \tag{32}$$

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* (31) *and* (32)*, PRB assignment is performed as follow.*

$$e^k_{r,u(s,i)} = \begin{cases} 1 & u(s,i) = \mathit{argmax} \mathfrak{F}^k_{r,u(s,i)} \forall s, \forall r, \forall k \\ 0 & \mathit{otherwise} \end{cases} \tag{33}$$

Thus the user in each slice s that have the largest value of $\mathfrak{F}^k_{r,u(s,i)}$, should be allocated to the PRB k; Due to the fact that just one PRB can be allocated to a UE between those UEs (regardless to the services) that are associated to the same O-RU.

B. Sub-Problem 2

After power allocation and PRB assignment, the remaining problem is to assign O-RU to the 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 assign one O-RU to each UE.

Greedy Algorithm for Non-Comp O-RU Assignment: The problem can be reformulated as follow

$$\max_{\boldsymbol{G}} \quad \sum_{s=1}^{S} \sum_{i=1}^{U_s} \sum_{r=1}^{R} \delta_s g^r_{u(s,i)} \bar{\mathcal{R}}^r_{u_{(s,k)}} \tag{34a}$$

$$\text{subject to} \quad \sum_{s=1}^{S} \sum_{i=1}^{U_s} g^r_{u(s,i)} \psi_{r,u(s,i)} \leq \mathfrak{t}_r \quad \forall r \tag{34b}$$

$$\sum_{r} g^r_{u(s,i)} = 1 \quad \forall s, \forall i, \tag{34c}$$

$$g^r_{u(s,i)} \in \{0,1\} \quad \forall s, \forall i, \tag{34d}$$

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 $\mathfrak{t}_r = \zeta_r - \sigma_r$ Since we obtained (26) in (III-A), we can ignore this constraint in (34). The problem (34) is an NP-complete 0-1 multiple knapsack problem. We solve this problem using GAAOU which is a greedy algorithm (1) as follow [10], [11]. Firstly, we set all variables $g^r_{u(s,i)} = 0, \quad \forall s, \forall i, \forall r$. Then we define $\mathfrak{B}^{rem}_{u_{(s,i)}} = \mathcal{R}\ \forall s, \forall i$ and $\mathfrak{C}_r = \mathfrak{t}_r, \forall r$ as a set of all O-RUs and value of each O-RU, respectively. Next, we sort all slices based on their priority. Afterward, we assign the O-RU that provides the highest achievable rate for each UE (we start from the UEs on the slices with highest priority) on the condition that it does not exceed the value of each O-RU (that is a function of maximum power and capacity of O-RU). If it exceeds the value of O-RU, then O-RU with the next highest achievable rate is selected. The complexity of sorting S slices based on their priority is $O(Slog(S))$. Depict $\mathfrak{N} = \sum_{s=1}^{S} \sum_{i=1}^{U_s} 1$. The complexity of this algorithm is about $O(Slog(S)) + O(R \times \mathfrak{N})$.

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

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
                $\mathfrak{B}^{rem}_{u_{(s,i)}} = \mathcal{R} \setminus \{r^*\}$
            end if
        end while
    end for
end for

C. Iterative Proposed Algorithm

In (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 $\boldsymbol{G}$, then $\boldsymbol{P}$ and $\boldsymbol{E}$ is achieved using Lagrangian method. Afterward, $\boldsymbol{G}$ is updated using GAAOU algorithm. This process is repeated until it converges. The whole algorithm is depicted as follow (Algorithm (2)).

Algorithm 2 Iterative Algorithm for Power Allocation, PRB, VNF and O-RU Association (IAPPVO)

```
Set the maximum number of iterations Iter_max, convergence condition ε > 0
Assign Users to O-RU randomly (Initialize G)
for i ← 1 to Iter_max do
    Acquire P^(i), E^(i) and M^(i) using Lagrangian function and sub-gradient method based on (III-A)
    Update 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
```

References

[1] 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.

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

[3] 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.

[4] 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.

[5] 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.

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

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

[8] 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.

[9] 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.

[10] 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.

[11] 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.