Network Slicing and Resource Allocation in an Open RAN System

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

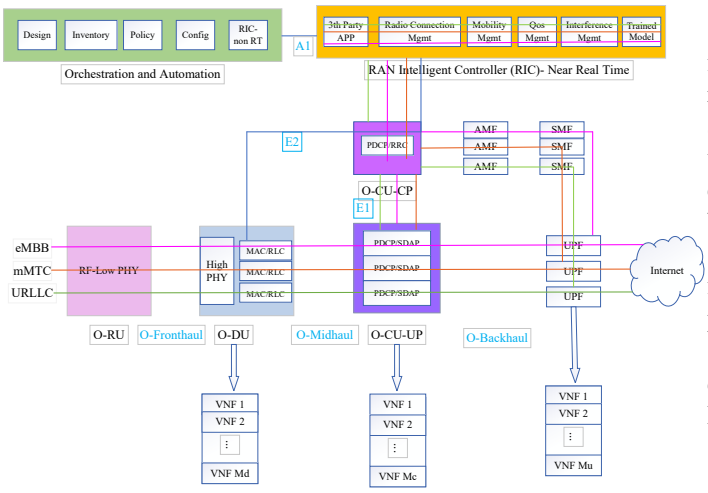

Fig. 1: Network sliced ORAN system

***Abstract*—**
***Index Terms*—**

I. Introduction

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 and delays for the downlink (DL) of the ORAN system. Afterward, we discuss about the assignment of physical data center resources. 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 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 K_{s_j}, $j \in \{1, 2, 3\}$ 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.

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

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

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

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

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

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

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 user 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, which is made from signal compression in O-DU.

The achievable data rate for the i^{th} UE request in the s^{th}_1 application of service type 1 (eMBB) can be written as $\mathcal{R}_{u(s_1,i)}$ that is formulated as below.

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

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^{th}_j ($j \in \{2,3\}$) 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

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

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,

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

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.

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 $\alpha_s^1 = \sum_{u=1}^{U_s} a_{u(s,i)}\lambda_{u(s,i)}$, 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_s^U) is approximately equal to the mean arrival data rate of the O-CU-UP layer (α_s^C) and O-DU (α_s^D). so $\alpha_s = \alpha_s^U \approx \alpha_s^C \approx \alpha_s^D$. 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

$$\begin{aligned} T_{DU}^s &= \frac{1}{\mu_d - \alpha_s/M_s^d}, \\ T_{CU}^s &= \frac{1}{\mu_c - \alpha_s/M_s^c} \\ T_{UPF}^s &= \frac{1}{\mu_u - \alpha_s/M_s^u} \end{aligned} \tag{11}$$

Where M_s^d, M_s^c and $M_s u$ are the variables that depict the sum of VNFs in O-DU, O-CU-UP and UPF, respectively. Moreover, $1/\mu_d$, $1/\mu_c$ and $1/\mu_u$ 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_{RU}^s is the mean transmission delay of s^{th} slice on the wireless link. The arrival data rate of wireless link is equal to the arrival data rate of load balancers for each service. Moreover, it is assumed that the service time of transmission queue for each slice s has an exponential distribution with mean $1/(R_{tot_s})$ and can be modeled as a M/M/1 queue. Therefore, the mean delay of the transmission layer is

$$T_{RU}^s = \frac{1}{R_{tot_s} - \alpha_s}; \tag{12}$$

where, $R_{tot_s} = \sum_{u=1}^{U_s} a_{u(s,i)} R_{u(s,i)}$ is the total achievable rate of each service. So the mean processing delay for each UE in slice s is

$$T_{process}^s = T_{RU}^s + T_{DU}^s + T_{CU}^s + T_{UPF}^s \tag{13}$$

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

$$\phi_s = M_s^u \phi_s^u + M_s^c \phi_s^c + M_s^d \phi_s^d \tag{14}$$

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

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,

$$\max_{\boldsymbol{P},\boldsymbol{E},\boldsymbol{M},\boldsymbol{G}} \quad \sum_{s=1}^{S}\sum_{i=1}^{U_s} R_{u_{(s,k)}} \tag{15a}$$

$$\text{subject to} \quad P_r \le P_{max} \quad \forall r \tag{15b}$$

$$p_{r,u(s,i)}^k \ge 0 \quad \forall i, \forall r, \forall s, \forall k, \tag{15c}$$

$$\mathcal{R}_{u_{(s_j,i)}} \ge \mathcal{R}_{min}^{s_j} \quad \forall s, j \in \{1,2,3\}, \tag{15d}$$

$$C^r \le C_{max}^r \quad \forall r, \tag{15e}$$

$$T_{tot}^s \le T_{tot}^{max,s} \quad \forall s, \tag{15f}$$

$$\sum_r g_{u(s,i)}^r \ge 1 \quad \forall s, \forall i, \tag{15g}$$

$$\sum_{k=1}^{K_s} e_{r,u(s,i)}^k \ge 1 \quad \forall s, \forall i, \tag{15h}$$

$$\phi_{tot} \le \phi_{max}, \tag{15i}$$

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. (15b), and (15c), 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 (15d) shows that the rate of each UE requesting eMBB, URLLC and mMTC is more than a threshold, respectively. (15e) and (15f) expressed the limited capacity of the fronthaul link, and the limited delay of receiving signal, respectively. (15g) and (15h) guarantee that O-RU and PRB is associated to the UE, respectively. In addition, (15i) indicate that the static cost of energy of VNFs in each slice do not exceed from the threshold.

III. Proposed Algorithm Scheme

In this section, we first apply some simplifications to the system; Solving problem (15) 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 $p^k_{u(s,l)}$ and $p^k_{u(n,l)}$ in (3) by P_{max}, an upper bound $\bar{I}^k_{r,u(s,i)}$ for $I^k_{r,u(s,i)}$ and lower bound $\bar{\mathcal{R}}_{u_{(s,i)}} \forall s, \forall i$ for $\mathcal{R}_{u_{(s,i)}}$ is obtained by replacing with $I^k_{r,u(s,i)}$ $\bar{I}^k_{r,u(s,i)}$ in (6) and (5) and make them concave.

Suppose $\hat{\rho}^k_{r,u(s,i)} = \frac{|P_{max}\mathbf{h}^{H\,k}_{r,u(s,i)}\mathbf{w}^k_{r,u(s,i)}g^r_{u(s,i)}|^2}{BN_0}$. To convexify (6), we replace $\rho^k_{r,u(s,i)}$ with $\hat{\rho}^k_{r,u(s,i)}$ in (7). So, a lower bound for (6) is given that is a concave function.

Suppose UPF, O-CU and O-DU have the same processor

Lemma 1. *In problem* (15)*,*