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,

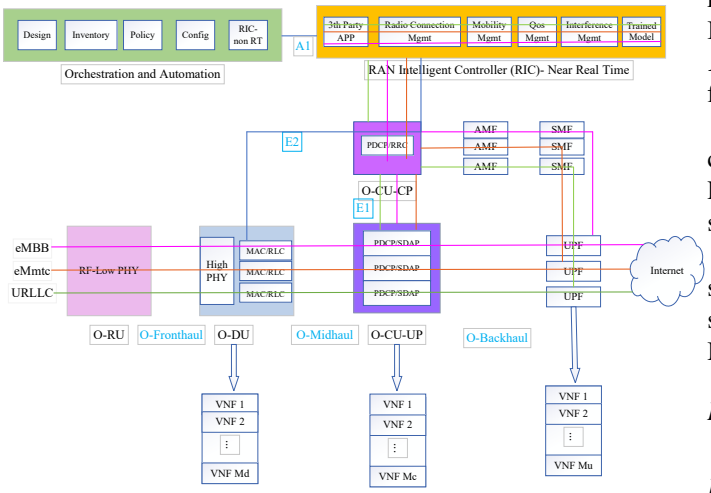

Fig. 1: Network sliced ORAN system

***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 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 preallocated slices serving these services; Each Service $s \in \{1,2,3\}$ consists of U_s request from the single-antenna UEs which require certain QoS to be able to use the requested program 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 $s_{\in}\{1,23\}$ consists of K_s preallocated virtual resource blocks that are mapped to 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 single-antenna RU that are shared between slices. All RUs $r \in \{1,2,...,R\}$ transmit signals cooperatively to all the UEs(RUs are in comp mode). Moreover, all RUs, have access to PRBs.

B. The Achievable Rate

The SNR of i^{th} UE requesting 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} 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} 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 a binary variable that illustrates whether RU r is mapped to the i^{th} UE allocate 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

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

where $\gamma_1 = e^k_{r,u(s,i)} e^k_{r',u(s,l)} a_{u(s,i)} a_{u(s,l)}$ and $\gamma_2 = e^k_{r,u(s,i)} e^k_{r',u(n,l)} a_{u(s,i)} a_{u(y,l)}$. Where $a_{u(s,i)} \in \{0,1\}$ is a binary variable to depict user admission. $e^k_{r,u(s,i)}$ 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.

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 $\mathcal{R}^e_{u(s_1,i)}$.

$$\begin{aligned} \mathcal{R}^{e,r}_{u(s_1,i)} &= \sum_{k=1}^{K_{s_1}} B \log_2(1+\rho^k_{r,u(s_1,i)}) a_{u(s_1,i)} e^k_{r,u(s_1,i)}, \\ \mathcal{R}^e_{u(s_1,i)} &= \sum_{r=1}^{R} \mathcal{R}^{e,r}_{u(s_1,i)} \end{aligned} \tag{4}$$

where B is the bandwidth of system. $\mathcal{R}^{e,r}_{u(s_1,i)}$ is the achievable rate of each RU r to UE i in slice s_1. Since the blocklength in URLLC is finite, the achievable data rate for the i^{th} UE request in the s_2^{th} application of service type 2 (URLLC) is not achieved from Shannon Capacity formula. So, for the short packet transmission the achievable data rate is approximated from follow

$$\begin{aligned} \mathcal{R}^{u,r}_{u(s_2,i)} &= \sum_{k=1}^{K_{s_2}} B(\log_2(1+\rho^k_{u(s_2,i)}) - \zeta^k_{u(s_2,i)}) \beta^k_{u(s_2,i)} \\ \mathcal{R}^e_{u(s_1,i)} &= \sum_{r=1}^{R} \mathcal{R}^{e,r}_{u(s_2,i)} \end{aligned} \tag{5}$$

Where $\beta^k_{u(s_2,i)} = a_{u(s_2,i)} e^k_{u(s_2,i)}$ and $\zeta^k_{u(s_2,i)} = log_2(e) Q^{-1}(\epsilon) \sqrt{\frac{C^k_{u(s_2,i)}}{N^k_{u(s_2,i)}}})$ Where, ϵ is the transmission probability, Q^{-1} is the inverse of Q- function (Gaussian), $C^k_{u(s_2,i)} = 1 - \frac{1}{(1+\rho^k_{u(s_2,i)})}$ depicts the channel dispersion of UE i at slice s_2, experiencing PRB k and $N^k_{u(s_2,i)}$ represents the blocklength of it. $\mathcal{R}^{e,r}_{u(s_1,i)}$ is the achievable rate of each RU r to UE i in slice s_2.

C. 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}.

$$\begin{aligned} T_{tot} &= T_{process} + T_{transmission} + T_{propagation} \\ T_{process} &= T_{RU} + T_{DU} + T_{CU} + T_{UPF} \\ T_{transmission} &= T_{front} + T_{mid} + T_{back} + T_{trans2net} \\ T_{propagation} &= T_{front} + T_{mid} + T_{back} + T_{trans2net} \end{aligned} \tag{6}$$

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). Here we assume the value of propagation delay 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 $\alpha^1_s = \sum_{u=1}^{U_s} a_{u(s,i)} \lambda_{u(s,i)}$, 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

$$\begin{aligned} T^s_{DU} &= \frac{1}{\mu_d - \alpha_s/M^d_s}, \\ T^s_{CU} &= \frac{1}{\mu_c - \alpha_s/M^c_s} \\ T^s_{UPF} &= \frac{1}{\mu_u - \alpha_s/M^u_s} \end{aligned} \tag{7}$$

Where M^d_s, M^c_s 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^i_s $i \in \{d, c, u\}$.

In addition, T^s_{RU} 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

$$T^s_{RU} = \frac{1}{R_{tot_s} - \alpha_s}; \tag{8}$$

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

$$T^s_{process} = T^s_{RU} + T^s_{DU} + T^s_{CU} + T^s_{UPF} \tag{9}$$

2) Transmission Delay: The transmission delay is the amount of time required to push all the packets into the

fiber link. Here, we have transmission delay in fronthaul, midhaul, backhaul and the link to transmit data to internet.

$$\begin{aligned} T_{front} &= \frac{\alpha_s^f}{R_f} \\ T_{mid} &= \frac{\alpha_s^m}{R_m} \\ T_{back} &= \frac{\alpha_s^b}{R_b} \\ T_{trans2net} &= \frac{\alpha_s^t}{R_t} \end{aligned} \tag{10}$$

Where, R_f, R_m, R_b and R_t are the rate of transmission in fronthaul, midhaul, backhaul and the link to transmit data to internet, respectively. Furthermore, the mean arrival data rate of the each link (α_s^i , $i \in \{f, m, b, t\}$) is approximately equal to others ($\alpha_s \approx \alpha_s^i$, $i \in \{f, m, b, t\}$).

D. Delay for URLLC

As we know, UEs request URLLC services, require services with low latency. For the M/M/1 system, the probability of the delay for each application s in the UPF, CU, DU and RU is as follow, respectively

$$\begin{aligned} P_r\{T_{UPF}^s \geq T_{UPF}^{max}\} &= e^{-(\mu_u - \alpha_s/M_s^u)T_{UPF}^{max}} \\ P_r\{T_{CU}^s \geq T_{CU}^{max}\} &= e^{-(\mu_c - \alpha_s/M_s^c)T_{CU}^{max}} \\ P_r\{T_{DU}^s \geq T_{DU}^{max}\} &= e^{-(\mu_d - \alpha_s/M_s^d)T_{DU}^{max}} \\ P_r\{T_{RU}^s \geq T_{RU}^{max}\} &= e^{-(R_{tot_s} - \alpha_s)T_{RU}^{max}} \end{aligned} \tag{11}$$

So the probability of coincidence of these events is as follow

$$P_r\{E_1, E_2, E_3, E_4\} = A_1 A_2 A_3 A_4, \tag{12}$$

Where $E_1 = T_{UPF}^s \geq T_{UPF}^{max}$, $E_2 = T_{CU}^s \geq T_{CU}^{max}$, $E_3 = T_{DU}^s \geq T_{DU}^{max}$ and $E_4 = T_{RU}^s \geq T_{RU}^{max}$. Also $A_1 = e^{-(\mu_u - \alpha_s/M_s^u)T_{UPF}^{max}}$, $A_2 = e^{-(\mu_c - \alpha_s/M_s^c)T_{CU}^{max}}$, $A_3 = e^{-(\mu_d - \alpha_s/M_s^d)T_{DU}^{max}}$ and $A_4 = e^{-(R_{tot_s} - \alpha_s)T_{RU}^{max}}$.

E. Physical Data Center Resource

Each VNF requires physical resources that contain memory, storage, CPU and Network Bandwidth. Let the required resources for VNF f in slice s is represented by a tuple as

$$\bar{\Omega}_s^f = \{\Omega_{M,s}^f, \Omega_{S,s}^f, \Omega_{C,s}^f, \Omega_{N,s}^f\}, \tag{13}$$

where $\bar{\Omega}_s^f \in \mathbb{C}^4$ and $\Omega_{M,s}^f, \Omega_{S,s}^f, \Omega_{C,s}^f, \Omega_{N,s}^f$ indicate the amount of required memory, storage, CPU and and Network Bandwidth, respectively. Moreover, the total amount of required memory, storage, CPU and Network Bandwidth of all VNFs of a slice in DU, CU and UPF is defined as below, respectively

$$\begin{aligned} \bar{\Omega}_{\mathfrak{z},s}^{tot,d} &= \sum_{f=1}^{M_s^d} \bar{\Omega}_{\mathfrak{z},s}^{f,d} \quad \mathfrak{z} \in \{M, S, C, N\}. \\ \bar{\Omega}_{\mathfrak{z},s}^{tot,c} &= \sum_{f=1}^{M_s^c} \bar{\Omega}_{\mathfrak{z},s}^{f,c} \quad \mathfrak{z} \in \{M, S, C, N\}. \\ \bar{\Omega}_{\mathfrak{z},s}^{tot,u} &= \sum_{f=1}^{M_s^u} \bar{\Omega}_{\mathfrak{z},s}^{f,u} \quad \mathfrak{z} \in \{M, S, C, N\}. \end{aligned} \tag{14}$$

Where, $\bar{\Omega}_{\mathfrak{z},s}^{f,d}$, $\bar{\Omega}_{\mathfrak{z},s}^{f,c}$ and $\bar{\Omega}_{\mathfrak{z},s}^{f,u}$ are the amount of resource that a VNF required in DU, CU and UPF, respectively.

$$\bar{\Omega}_{\mathfrak{z},s}^{tot} = \bar{\Omega}_{\mathfrak{z},s}^{tot,d} + \bar{\Omega}_{\mathfrak{z},s}^{tot,c} + \bar{\Omega}_{\mathfrak{z},s}^{tot,u} \tag{15}$$

Also, there are D_c data centers (DC), serving the VNFs. Each DC contains several servers that supply VNF requirements. The amount of memory, storage, CPU and and Network Bandwidth is denoted by τ_{M_j}, τ_{S_j}, τ_{C_j} and τ_{N_j} for the j^{th} DC, respectively

$$\tau_j = \{\tau_{M_j}, \tau_{S_j}, \tau_{C_j}, \tau_{N_j}\},$$

In this system model, the assignment of physical DC resources to VNFs is considered. Let $y_{s,d}$ be a binary variable indicating whether the d^{th} DC is allocated the resources to the VNFs of s^{th} slice or not.

F. Problem Statement

Power of each O-RU is obtained as below

$$P_r = \sum_{s=1}^{S} \sum_{i=1}^{U_v} \sum_{k=1}^{K_s} p_{r,u(s,i)}^k a_{u(s,i)} e_{r,u(s,i)}^k g_{u(s,i)}^r. \tag{16}$$

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

$$\phi_{tot} = \textstyle\sum_{s=1}^{S} \phi_s + \sum_{d=1}^{D_c} z_d \psi_d.$$

Where,z_d is shown that whether the d^{th} DC is turned on or not and ψ_d is a static cost when a DC is active.

$$z_d = \begin{cases} 1 & \sum_{s=1}^{S} y_{s,d} \geq 1 \\ 0 & \text{otherwise} \end{cases} \tag{17}$$

In addition, ϕ_s is obtained from below

$$\phi_s = M_s^u \phi_s^u + M_s^c \phi_s^c + M_s^d \phi_s^d \tag{18}$$

Where, ϕ_s^u, ϕ_s^c and ϕ_s^d are the static cost of energy in UPF, CU and DU, respectively. So 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,

$$\max_{\boldsymbol{P},\boldsymbol{A},\boldsymbol{E},\boldsymbol{M},\boldsymbol{G},\boldsymbol{Y}} \quad \sum_{s=1}^{S} \sum_{i=1}^{U_s} R_{u_{(s,k)}} \tag{19a}$$

$$\text{subject to} \quad P_r \leq P_{max} \quad \forall r \tag{19b}$$

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

$$\mathcal{R}_{u_{(s_1,i)}}^e \geq a_{u(s_1,i)} \mathcal{R}_{min}^{s_1,e} \quad \forall s_1, \tag{19d}$$

$$\mathcal{R}_{u_{(s_2,i)}}^u \geq a_{u(s_2,i)} \mathcal{R}_{min}^{s_2,u} \quad \forall s_2, \tag{19e}$$

$$\sum_{s=1}^{S} \sum_{i=1}^{U_s} R_{u_{(s,k)}}^r \leq C_{max}^r \quad \forall r, \tag{19f}$$

$$T_{tot}^s \leq T_{tot}^{max,s} \quad \forall s, \tag{19g}$$

$$P_r\{E_1, E_2, E_3, E_4\} \leq \epsilon_s \quad \forall s_2, \tag{19h}$$

$$a_{u(s,i)} \leq a_{u(s,i)} \sum_r g_{u(s,i)}^r \quad \forall s, \forall i, \tag{19i}$$

$$a_{u(s,i)} g^r_{u(s,i)} \leq a_{u(s,i)} g^r_{u(s,i)} \sum_{k=1}^{K_s} e^k_{r,u(s,i)} \quad \forall s, i, \tag{19j}$$

$$\phi_{tot} \leq \phi_{max}, \tag{19k}$$

$$\textstyle\sum_{s=1}^{S} y_{s,d} \bar{\Omega}^{tot}_{\mathfrak{z},s} \leq \tau_{\mathfrak{z}_d} \quad \forall d, \forall \mathfrak{z} \in \mathcal{E}; \tag{19l}$$

where $\boldsymbol{P} = [p_{u(s,i)}]\ \forall s, \forall i$, is the matrix of power for UEs, $\boldsymbol{A} = [a_{u(s,i)}]\ \forall s, \forall i$ denotes the binary variable for UE admission, $\boldsymbol{E} = [e^k_{r,u(s,i)}]\ \forall s, \forall i \forall r, \forall k$ 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 and $\boldsymbol{Y} = [y_{s,d}]\ \forall s, \forall d$ is a binary variable shown whether the physical DC is mapped to a VNFs of a slice or not. Also, η is weighted variable to value between the benefit and the cost term of objective function. (19b), and (19c), 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 (19d) and (19e) shows that the rate of each UE requesting eMBB and URLLC is more than a threshold, respectively. (19f) and (19g) expressed the limited capacity of the fronthaul link, and the limited delay of receiving signal, respectively. (19g) is a reliability condition that the delay in each layer should be less than threshold. (19i) and (19j) guarantee that if a UE in admitted by the system, O-RU and PRB is associated to it, respectively. In addition, (19k) indicate that the static cost of energy of VNFs in each slice do not exceed from the threshold. Moreover, in (19l) $\mathcal{E} = \{M, S, C, N\}$ and the constraint supports that we have enough physical resources for VNFs of each slice.