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,

***Abstract*—**
***Index Terms*—**

I. Introduction

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 two service types includes eMBB and URLLC. Assume we have V_1 and V_2 different applications for the first and second service type, respectively ($V = V_1 + V_2$). Assume we have S preallocated slices serving V services; There are S_1 slices for the first service type (eMBB) and S_2 slices for the second service type (URLLC) ($S = S_1 + S_2$). Each Service $v_j \in \{1, 2, ..., V_j\}$ consists of U_{v_j} request from the single-antenna UEs which require certain QoS to be able to use the requested program($j \in \{1, 2\}$ 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 slice. Each slice $s_j \in \{1, 2, ..., S_j\}$, $j \in \{1, 2\}$ consists of K_{s_j}, $j \in \{1, 2\}$ 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 multi-antenna RU that are shared between slices. Each RU $r \in \{1, 2, ..., R\}$ has J antenna for transmitting and receiving data. Moreover, all RUs, have access to PRBs.

B. The Achievable Rate

The SNR of i^{th} UE in v^{th} service experienced at slice s on PRB k which is obtained from $\rho^{k,s}_{u(v,i)}$ is the SNR

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

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

$$\mathbf{w}^{k,s}_{r,u(v,i)} = \mathbf{h}^{k,s}_{r,u(v,i)} (\mathbf{h}^{H\,k,s1}_{r,u(v,i)} \mathbf{h}^{k,s}_{r,u(v,i)})^{-1} \quad (2)$$

Moreover, $g^{r}_{u(v,i)} \in \{0, 1\}$ is a binary variable that illustrates whether RU r is mapped to the i^{th} UE in v_1^{th} service or not. Also, BN_0 denotes the power of Gaussian additive noise, and $I^{k,s1}_{r,u(v,i)}$ is the power of interfering signals represented as follow

$$I^{k,s}_{r,u(v,i)} = \underbrace{\sum_{\substack{l=1 \\ l \neq i}}^{U_v} \gamma_1 p^{k,s}_{u(v,i)} \sum_{\substack{r'=1 \\ r' \neq r}}^{R} |\mathbf{h}^{H\,k,s}_{r',u(v,i)} \mathbf{w}^{k,s1}_{r',u(v,i)} g^{r'}_{u(v,i)}|^2}_{\text{(intra-service interference)}} + \underbrace{\sum_{\substack{y=1 \\ l \neq v}}^{V} \sum_{n=1}^{S} \sum_{l=1}^{U_y} \gamma_2 p^{k,n}_{u(y,l)} \sum_{\substack{r'=1 \\ r' \neq r}}^{R} |\mathbf{h}^{H\,k,s}_{r',u(v,i)} \mathbf{w}^{k,n}_{r',u(y,l)} g^{r'}_{u(y,l)}|^2}_{\text{(inter-service interference)}} \quad (3)$$

where $\gamma_1 = e^{k,s}_{r,u(v,i)} e^{k,s}_{r',u(v,l)} a_{u(v,i)} a_{u(v,l)} b_{v,s}$ and $\gamma_2 = e^{k,s}_{r,u(v,i)} e^{k,n}_{r',u(y,l)} a_{u(v,i)} a_{u(y,l)} b_{v,s} b_{y,n}$. Where $a_{u(v,i)} \in \{0, 1\}$ is a binary variable to depict user admission. $b_{v,s}$ is a binary variable that illustrates whether slice s is allocated to service v or not. $e^{k,s}_{r,u(v,i)}$ is the binary variable to show whether the k^{th} PRB is allocated to the UE i in service v, assigned to r^{th} o-RU using slice s or not.

The achievable data rate for the i^{th} UE request in the v_1^{th} application of service type 1 (eMBB) can be written as

$$\mathcal{R}^{e}_{u(v_1,i)} = \sum_{s_1=1}^{S_1} \sum_{k=1}^{K_{s_1}} B \log_2(1 + \rho^{k,s_1}_{u(v_1,i)}) a_{u(v_1,i)} b_{v_1,s_1} e^{k,s_1}_{r,u(v_1,i)}, \quad (4)$$

where B is the bandwidth of system.

Since the blocklength in URLLC is finite, the achievable data rate for the i^{th} UE request in the v_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

$$\mathcal{R}^{u}_{u(v_2,i)} = \sum_{s_2=1}^{S_2}\sum_{k=1}^{K_{s_2}} B(\log_2(1+\rho^{k,s_2}_{u(v_2,i)}) - \zeta^{k,s_2}_{u(v_2,i)})\beta^{k,s_2}_{u(v_2,i)} \tag{5}$$

Where $\beta^{k,s_2}_{u(v_2,i)} = a_{u(v_2,i)}b_{v_2,s_1}e^{k,s_2}_{u(v_2,i)}$ and $\zeta^{k,s_2}_{u(v_2,i)} = log_2(e)Q^{-1}(\epsilon)\sqrt{\frac{C^{k,s_2}_{u(v_2,i)}}{N^{k,s_2}_{u(v_2,i)}}})$ Where, ϵ is the transmission probability, Q^{-1} is the inverse of Q- function (Gaussian), $C^{k,s_2}_{u(v_2,i)} = 1 - \frac{1}{(1+\rho^{k,s_2}_{u(v_2,i)})}$ depicts the channel dispersion of UE i requesting service v_2, experiencing PRB k at slice s_2 and $N^{k,s_2}_{u(v_2,i)}$ represents the blocklength of it.

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

Assume the mean arrival data rate of the UPF layer (α^U_v) is approximately equal to the mean arrival data rate of the O-CU-UP layer (α^C_v) and O-DU (α^D_v). so $\alpha_v = \alpha^U_v \approx \alpha^C_v \approx \alpha^D_v$. since, by using Burkes Theorem, the mean arrival data rate of the second layer which is processed in the first layer is still Poisson with rate α_v. 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 v^{th} service in the first and the second layer, modeled as M/M/1 queue, is formulated as follow, respectively

$$\begin{aligned} T^v_{DU} &= \frac{1}{\mu_d - \alpha_v/M_{v,d}}, \\ T^v_{CU} &= \frac{1}{\mu_c - \alpha_v/M_{v,c}} \\ T^v_{UPF} &= \frac{1}{\mu_u - \alpha_v/M_{v,u}} \end{aligned} \tag{7}$$

Where $M_{v,d} = \sum_{s=1}^{S} M^d_s b_{v,s}$, $M_{v,c} = \sum_{s=1}^{S} M^c_s b_{v,s}$ and $M_{v,u} = \sum_{s=1}^{S} M^u_s b_{v,s}$ are the sum of VNFs in O-DU, O-CU-UP and UPF, respectively. Since, each service may use more than one slice, the data packets of each service is divided between VNFs of slices allocated to service. 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_v 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 service v is $\alpha_v/M_{v,i}$ $i \in \{d, c, u\}$.

In addition, T^v_{RU} is the mean transmission delay of v^{th} service 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 $1/(R_{tot_v})$ and can be modeled as a M/M/1 queue. Therefore, the mean delay of the transmission layer is

$$T^v_{RU} = \frac{1}{R_{tot_v} - \alpha_v}; \tag{8}$$

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

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

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.

$$\begin{aligned} T_{front} &= \frac{\alpha^f_v}{R_f} \\ T_{mid} &= \frac{\alpha^m_v}{R_m} \\ T_{back} &= \frac{\alpha^b_v}{R_b} \\ T_{trans2net} &= \frac{\alpha^t_v}{R_t} \end{aligned} \tag{10}$$

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^i_v , $i \in \{f, m, b, t\}$) is approximately equal to others ($\alpha_v \approx \alpha^i_v$, $i \in \{f, m, b, t\}$).

D. 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

$$\bar{\Omega}^f_s = \{\Omega^f_{M,s}, \Omega^f_{S,s}, \Omega^f_{C,s}, \Omega^f_{N,s}\}, \tag{11}$$

where $\bar{\Omega}^f_s \in \mathbb{C}^4$ and $\Omega^f_{M,s}, \Omega^f_{S,s}, \Omega^f_{C,s}, \Omega^f_{N,s}$ 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 is defined as

$$\bar{\Omega}^{tot}_{\mathfrak{z},s} = \sum_{f=1}^{F_s} \bar{\Omega}^{f}_{\mathfrak{z},s} \quad \mathfrak{z} \in \{M, S, C, N\}. \tag{12}$$

Where, $F_s = M^d_s + M^c_s + M^u_s$ 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 $\tau_{M_j}, \tau_{S_j}, \tau_{C_j}$ 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.

E. Problem Statement

In this system, the goal is to minimize the cost of the system. Power of each O-RU is obtained as below

$$P_r = \sum_{v=1}^{V} \sum_{i=1}^{U_v} \sum_{s=1}^{S} \sum_{k=1}^{K_s} p^{k,s}_{r,u(v,i)} a_{u(v,i)} b_{v,s} e^{k,s}_{r,u(v,i)} g^{r}_{u(v,i)}. \tag{13}$$

The total power cost of O-RUs for transmitting data to UE is depicted as follow

$$P_{tot} = \sum_{r=1}^{R} P_r \tag{14}$$

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

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

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.

$$z_d = \begin{cases} 1 & \sum_{s=1}^{S} y_{s,d} \geq 1 \\ 0 & \text{otherwise} \end{cases} \tag{15}$$

In addition, $\phi_{s,d}$ is obtained from below

$$\phi_s = M^u_s \phi^u_s + M^c_s \phi^c_s + M^d_s \phi^d_s \tag{16}$$

So the optimization problem is formulated as follow The aim of this paper is minimize the total power of all RUs and the total power consumption of baseband processing at all DCs simultaneously, with the presence of constraints which is written as follow,

$$\min_{\boldsymbol{P},\boldsymbol{A},\boldsymbol{E},\boldsymbol{M},\boldsymbol{G},\boldsymbol{Y}} \quad P_{tot} + \varphi \phi_{tot} \tag{17a}$$

$$\text{subject to} \quad P_r \leq P_{max} \quad \forall r \tag{17b}$$

$$p^{k,s}_{r,u(v,i)} \geq 0 \quad \forall v, \forall i, \forall r, \forall s, \forall k, \tag{17c}$$

$$\mathcal{R}^{e}_{u_{(v_1,k)}} \geq \mathcal{R}^{v_1,e}_{min} \quad \forall v_1, \tag{17d}$$

$$\mathcal{R}^{u}_{u_{(v_2,k)}} \geq \mathcal{R}^{v_2,u}_{min} \quad \forall v_2, \tag{17e}$$

$$T^{v}_{tot} \leq T^{max,v}_{tot} \quad \forall v, \tag{17f}$$

$$a_{u(v,i)} \leq a_{u(v,i)} \sum_{r} g^{r}_{u(v,i)} \leq 1 \quad \forall v, \forall i \tag{17g}$$

$$\sum_{s=1}^{S} \sum_{v=1}^{V} \sum_{i=1}^{U_v} a_{u(v,i)} g^{r}_{u(v,i)} e^{k,s}_{r,u(v,i)} \leq 1 \quad \forall r, \tag{17h}$$

$$\sum_{s=1}^{S} b_{v,s} \geq 1 \quad \forall v, \tag{17i}$$

$$\sum_{d=1}^{D_c} \sum_{v=1}^{V} y_{s,d} a_{v,s} \geq 1 \times \sum_{v=1}^{V} a_{v,s} \forall s, \tag{17j}$$

$$\sum_{s=1}^{S} y_{s,d} \bar{\Omega}^{tot}_{\mathfrak{z},s} \leq \tau_{\mathfrak{z}d} \forall d, \forall \mathfrak{z} \in \mathcal{E}; \tag{17k}$$