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 $v_i \in \{1, 2, ..., V_i\}$ consists of U_{v_i} request from the single-antenna UEs which require certain QoS to be able to use the requested program($i \in \{1, 2\}$ 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 $s_i \in \{1, 2, ..., S_i\}$, $i \in \{1, 2\}$ consists of K_{s_i}, $i \in \{1, 2\}$ preallocated virtual resource blocks that are mapped to Physical Resource Blocks (PRBs), $M_{s,1}$ VNFs for the processing of O-DU, $M_{s,2}$ VNFs for the processing of O-CU-UP and M_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 achievable data rate for the i^{th} UE request in the v_1^{th} application of service type 1 (eMBB) can be written as

$$\mathcal{R}^{e}_{u(v_1,i)} = \sum_{s1=1}^{S_1} \sum_{k=1}^{K_s} B \log_2(1 + \rho^{k,s1}_{u(v_1,i)}) a_{u(v,i)} b_{v_1,s_1} e^{s_1}_{k,u(v_1,i)}, \tag{1}$$

where B is the bandwidth of system, $a_{u(v,i)} \in \{0, 1\}$ is a binary variable to depict user admission. b_{v_1,s_1} is a binary variable that illustrates whether slice s_1 is allocated to service v_1 or not. $e^{s_1}_{k,u(v_1,i)}$ is the binary variable to show whether the UE i in service v_1 using slice s_1, is assigned to k^{th} PRB or not. and $\rho^{k,s1}_{u(v_1,i)}$ is the SNR of i^{th} UE in v_1^{th} service experienced at slice s_1 on PRB k which is obtained from

$$\rho^{k,s1}_{u(v_1,i)} = \sum_{s1=1}^{S_1} \frac{p_{u(v,i)} \sum_{s=1}^{S} |\mathbf{h}^H_{R_s,u(v,i)} \mathbf{w}_{R_s,u(v,i)}|^2 a_{v,s}}{BN_0 + I_{u(v,i)}}, \tag{2}$$

Suppose there are S slices Serving V services. Each Service $v \in \{1, 2, ..., V\}$ consists of U_v single-antenna UEs that require certain service. Each slice $s \in \{1, 2, ..., S\}$ consists of R_s RUs and K_s physical resource blocks (PRBs), one DU and one CU that contains VNFs. Slices can have shared resources. All RUs in a slice, that are mapped to a service, transmit signals cooperatively to all the UEs in a specific service [**?**], [**?**]. Each RU $r \in \{1, 2, ..., R\}$ is mapped to a DU via an optical fiber link with limited fronthaul capacity. There are two processing layers one in the DU and one in the CU of ORAN system, each represented with a VNF. The lower layer (i.e., DU) consists of high-PHY, MAC, and RLC, and the upper layer (i.e., CU) consists of RRC, PDCP and SDAP. Assume we have M_1 VNFs in the DU layer and M_2 VNFs in the CU layer for processing data. Each VNF in both layers belongs to one or more slices. So, in the s^{th} slice, there are $M_{s,1}$ VNFs in the DU layer and $M_{s,2}$ VNFs in the CU layer. The VNFs in the DU and CU layers have the computational capacity that is equal to μ_1 and μ_2, respectively. Also, RUs and PRBs can serve more than one slice.

C. The Achievable Rate

The achievable data rate for the i^{th} UE in the v^{th} service can be written as

$$\mathcal{R}_{u(v,i)} = B \log_2(1 + \rho_{u(v,i)}), \tag{3}$$

where B is the bandwidth of system and $\rho_{u(v,i)}$ is the SNR of i^{th} UE in v^{th} service which is obtained from

$$\rho_{u(v,i)} = \frac{p_{u(v,i)} \sum_{s=1}^{S} |\mathbf{h}_{R_s,u(v,i)}^H \mathbf{w}_{R_s,u(v,i)}|^2 a_{v,s}}{BN_0 + I_{u(v,i)}}, \tag{4}$$

where $p_{u(v,i)}$ represents the transmission power allocated by RUs to i^{th} UE in v^{th} service, and $\mathbf{h}_{R_s,u(v,i)} \in \mathbb{C}^{R_s}$ is the vector of channel gain of a wireless link from RUs in the s^{th} slice to the i^{th} UE in v^{th} service. In addition, $\mathbf{w}_{R_s,u(v,i)} \in \mathbb{C}^{R_s}$ depicts the transmit beamforming vector from RUs in the s^{th} slice to the i^{th} UE in v^{th} service. Moreover, BN_0 denotes the power of Gaussian additive noise, and $I_{u(v,i)}$ is the power of interfering signals. Moreover, $a_{v,s} \in \{0, 1\}$ is a binary variable that illustrates whether slice s is mapped to service v or not. If $a_{v,s} = 1$ then, v^{th} service is mapped to s^{th} slice; otherwise, it is not mapped.

To obtain SNR as formulated in (4), let $\mathbf{y}_{U_v} \in \mathbb{C}^{U_v}$ be the received signal's vector of all users in v^{th} service

$$\mathbf{y}_{U_v} = \sum_{s=1}^{S} \sum_{k=1}^{K_s} \boldsymbol{H}_{\mathcal{R}_s,\mathcal{U}_v}^H \ \mathfrak{y}_{R_s} \zeta_{U_v,k,s} a_{v,s} + \boldsymbol{z}_{\mathcal{U}_v}, \tag{5}$$

where $\mathfrak{y}_{R_s} = \boldsymbol{W}_{\mathcal{R}_s,\mathcal{U}_v} \boldsymbol{P}_{U_v}^{\frac{1}{2}} \boldsymbol{x}_{\mathcal{U}_v} + \boldsymbol{q}_{\mathcal{R}_s}$ and $\boldsymbol{x}_{\mathcal{U}_v} = [x_{u_{(v,1)}}, ..., x_{u_{(v,\mathcal{U}_v)}}]^T \in \mathbb{C}^{R_s}$ depicts the transmitted symbol vector of UEs in v^{th} set of service, $\boldsymbol{z}_{U_v}$ is the additive Gaussian noise $\boldsymbol{z}_{\boldsymbol{U_v}} \backsim \mathcal{N}(0, N_0 \boldsymbol{I}_{U_v})$ and N_0 is the noise power. In addition, $\boldsymbol{q}_{R_s} \in \mathbb{C}^{R_s}$ indicates the quantization noise, which is made from signal compression in DU. Besides, $\boldsymbol{P}_{U_v} = \text{diag}\,(p_{u_{(v,1)}}, ..., p_{u_{(v,\mathcal{U}_v)}})$.

Furthermore, $\zeta_{k,s}^{U_v} \triangleq \{\zeta_{k,s}^{u(v,1)}, \zeta_{k,s}^{u(v,2)}, ..., \zeta_{k,s}^{u(v,N_{U_v})}\}$, $\zeta_{k,s}^{u(v,i)} \in \{0, 1\}$ is a binary parameter, which demonstrates whether i^{th} UE in v^{th} service can transmit its signals through k^{th} PRB and also this PRB belongs to s^{th} slice or not. $\boldsymbol{H}_{\mathcal{R}_s,\mathcal{U}_v} = \left[\boldsymbol{h}_{\mathcal{R}_s,u_{(v,1)}}, \ldots, \boldsymbol{h}_{\mathcal{R}_s,v_{(v,\mathcal{U}_v)}}\right]^T \in \mathbb{C}^{R_s \times U_v}$ shows the channel matrix between RU set $\mathcal{R}_s$ to UE set $\mathcal{U}_v$, besides. What's more, it is assumed we have perfect channel state information (CSI).

Moreover, $\boldsymbol{W}_{\mathcal{R}_s,\mathcal{U}_v} = [\boldsymbol{w}_{\mathcal{R}_s,u(v,1)}, ..., \boldsymbol{w}_{\mathcal{R}_s,u(v,U_v)}] \in \mathbb{C}^{R_s \times U_v}$ is the zero forcing beamforming vector to minimize the interference which is indicated as below

$$\boldsymbol{W}_{\mathcal{R}_s,\mathcal{U}_v} = \boldsymbol{H}_{\mathcal{R}_s,\mathcal{U}_v} (\boldsymbol{H}_{\mathcal{R}_s,\mathcal{U}_v}^H \boldsymbol{H}_{\mathcal{R}_s,\mathcal{U}_v})^{-1}. \tag{6}$$

Hence, the interference power of i^{th} UE in v^{th} service can be represented as follow

$$\begin{aligned} I_{u_{(v,i)}} = & \underbrace{\sum_{s=1}^{S} \sum_{n=1}^{S} \sum_{\substack{l=1 \\ l \neq i}}^{U_v} \gamma_1 p_{u_{(v,l)}} a_{v,s} \zeta_{u_(v,i),n,s} \zeta_{u_(v,l),n,s}}_{\text{(intra-service interference)}} \\ & + \underbrace{\sum_{\substack{y=1 \\ l \neq v}}^{V} \sum_{s=1}^{S} \sum_{n=1}^{S} \sum_{l=1}^{U_y} \gamma_2 p_{u_{(y,l)}} a_{y,s} \zeta_{u_(v,i),n,s} \zeta_{u_(y,l),n,s}}_{\text{(inter-service interference)}} \\ & + \underbrace{\sum_{s=1}^{S} \sum_{j=1}^{R_s} \sigma_{q_{r_{(s,j)}}}^2 |\boldsymbol{h}_{r_{(s,j)},u_{(v,i)}}|^2 a_{v,s}}_{\text{(quantization noise interference)}}, \end{aligned} \tag{7}$$

where $\gamma_1 = |\boldsymbol{h}_{\mathcal{R}_s,u_{(v,i)}}^H \boldsymbol{w}_{\mathcal{R}_s,u_{(v,l)}}|^2$ and $\gamma_2 = |\boldsymbol{h}_{\mathcal{R}_s,u_{(v,i)}}^H \boldsymbol{w}_{\mathcal{R}_s,u_{(y,l)}}|^2$. Moreover, $\sigma_{q_{r_{(s,j)}}}$ is the variance of quantization noise of j^{th} RU in s^{th} slice. Interference signal for each UE is coming from UEs using the same PRB. If we replace $p_{u_{(v,l)}}$ and $p_{u_{(y,l)}}$ by P_{max}, an upper bound $\bar{I}_{u_{(v,i)}}$ is obtained for $I_{u_{(v,i)}}$. Therefore, $\bar{\mathcal{R}}_{u_{(v,i)}} \forall v, \forall i$ is derived by using $\bar{I}_{u_{(v,i)}}$ instead of $I_{u_{(v,i)}}$ in (3) and (4).

Let $\bar{p}_{r_{(s,j)}}$ denote the power of transmitted signal from the j^{th} RU in s^{th} slice. From (5), we have,

$$\bar{p}_{r_{(s,j)}} = \sum_{v=1}^{V} \boldsymbol{w}_{r_{(s,j)},\mathcal{U}_v} \boldsymbol{P}_{\mathcal{U}_v}^{\frac{1}{2}} \boldsymbol{P}_{\mathcal{U}_v}^{H\frac{1}{2}} \boldsymbol{w}_{r_{(s,j)},\mathcal{U}_v}^H a_{v,s} + \sigma_{q_{r(s,j)}}^2. \tag{8}$$

Nevertheless, the rate of users on the fronthual link between DU and the j^{th} RU in s^{th} slice is formulated as [**?**], [**?**]

$$C_{R_{(s,j)}} = \log\,(1 + \sum_{v=1}^{V} \frac{w_{r_{(s,j)},\mathcal{D}_s} \boldsymbol{P}_{\mathcal{U}_v}^{\frac{1}{2}} \boldsymbol{P}_{\mathcal{U}_v}^{H\frac{1}{2}} w_{r_{(s,j)},\mathcal{U}_v}^H a_{v,s}}{\sigma_{q_{r(s,j)}}^2}), \tag{9}$$

where, $a_{v,s}$ is a binary variable denotes whether the slice s is mapped to service v or not .

D. Mean 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 UEs mapped to the s^{th} slice in the CU layer is $\alpha_{s_1} = \sum_{v=1}^{V} \sum_{u=2}^{U_v} a_{v,s} \lambda_{u(v,i)}$, where $a_{v,s}$ is a binary variable which indicates whether the v^{th} service is mapped to the

s^{th} slice or not. Furthermore, the mean arrival data rate of the DU layer is approximately equal to the mean arrival data rate of the first layer $\alpha_s = \alpha_{s_1} \approx \alpha_{s_2}$ 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_s. It is assumed that there are load balancers in each layer for each slice to divide the incoming traffic to VNFs equally [**?**], [**?**], [**?**]. Suppose the baseband processing of each VNF is depicted as an M/M/1 processing queue. Each packet is processed by one of the VNFs of a slice. So, the mean delay of the s^{th} slice in the first and the second layer, modeled as M/M/1 queue, is formulated as follow, respectively

$$\begin{aligned} d_{s_1} &= \frac{1}{\mu_1 - \alpha_s/M_{s,1}}, \\ d_{s_2} &= \frac{1}{\mu_2 - \alpha_s/M_{s,2}}. \end{aligned} \tag{10}$$

where $1/\mu_1$ and $1/\mu_2$ are the mean service time of the first and the second 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 of the slice s is $\alpha_s/M_{s,i}$ $i \in \{1, 2\}$. In addition, $d_{s_{tr}}$ is the transmission delay for 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 slice [**?**]. 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

$$d_{s_{tr}} = \frac{1}{R_{tot_s} - \alpha_s}; \tag{11}$$

where, $R_{tot_s} = \sum_{v=1}^{V} \sum_{u=2}^{U_v} a_{v,s} R_{u(v,i)}$ is the total achievable rate of each slice that is mapped to specific service. Mean delay of each slice is obtained as below.

$$D_s = d_{s_1} + d_{s_2} + d_{s_{tr}} \forall s. \tag{12}$$

E. Physical Data Center Resource

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

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

where $\bar{\Omega}_s^f \in \mathbb{C}^3$ and $\Omega_{M,s}^f, \Omega_{S,s}^f, \Omega_{C,s}^f$ indicate the amount of required memory, storage, and CPU, respectively. Moreover, the total amount of required memory, storage and CPU of all VNFs of a slice is defined as

$$\bar{\Omega}_{\mathfrak{z},s}^{tot} = \sum_{f=1}^{F_s} \bar{\Omega}_{\mathfrak{z},s}^f \;\; \mathfrak{z} \in \{M, S, C\}. \tag{14}$$

Where, $F_s = M_{s_1} + M_{s_2}$ 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 and CPU is denoted by $\tau_{M_j}, \tau_{S_j}$and τ_{C_j} for the j^{th} DC, respectively

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

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 connected to the VNFs of s^{th} slice or not.

F. Problem Statement

An important criterion to measure the optimality of a system is energy efficiency represented as the sum-rate to sum-power

$$\eta(\boldsymbol{P}, \boldsymbol{A}) := \frac{\sum_{v=1}^{V} \sum_{k=1}^{U_v} \mathcal{R}_{u_{(v,k)}}}{\sum_{s=1}^{S} \sum_{i=1}^{R_s} \bar{p}_{r_{(s,i)}}} = \frac{\mathfrak{R}_{tot}(\boldsymbol{P}, \boldsymbol{A})}{P_r^{tot}(\boldsymbol{P}, \boldsymbol{A})}, \tag{15}$$

where, $P_r^{tot}(\boldsymbol{P}, \boldsymbol{A}) = \sum_{s=1}^{S} \sum_{i=1}^{R_s} \bar{p}_{r_{(s,i)}}$ is the total power consumption of all RUs in all slices. Also, $\mathfrak{R}_{tot}(\boldsymbol{P}, \boldsymbol{A}) = \sum_{v=1}^{V} \sum_{k=1}^{U_v} \mathcal{R}_{u_{(v,k)}}$ is the total rates of all UEs applied for all types of services. 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

$$\phi_{tot} = \sum_{s=1}^{S} \sum_{d=1}^{D_c} y_{s,d} \phi_{s,d}.$$

Also, a cost function for the placement of VNFs into DCs is defined as

$$\psi_{tot} = \phi_{tot} - \nu \sum_{d=1}^{D_c} \sum_{v=1}^{V} y_{s,d} a_{v,s} \tag{16}$$

where, ν is a design variable to value between the first term of (16) which is the total power consumption of physical resources and the second term that is shown the amount of admitted slices to have physical resources. Our goal is to maximize sum-rate and minimize sum-power (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,

$$\max_{\boldsymbol{P}, \boldsymbol{A}, \boldsymbol{Y}} \quad \eta(\boldsymbol{P}, \boldsymbol{A}) + \varphi \frac{1}{\psi_{tot}(\boldsymbol{Y})} \tag{17a}$$

$$\text{subject to} \quad \bar{p}_{r_{(s,i)}} \leq P_{max} \quad \forall s, \forall i, \tag{17b}$$

$$p_{u_{(v,k)}} \geq 0 \quad \forall v, \forall k, \tag{17c}$$

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

$$C_{r_{(s,i)}} \leq C_{r_{(s,i)}}^{max} \quad \forall s, \forall i, \tag{17e}$$

$$D_s \leq D_s^{max} \quad \forall s, \tag{17f}$$

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

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

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

where $\boldsymbol{P} = [p_{u(v,k)}] \;\; \forall v, \forall k$, is the matrix of power for UEs, $\boldsymbol{A} = [a_{v,s}] \;\; \forall v, \forall s$ denotes the binary variable for connecting slices to services 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, φ is weighted variable to value between first and second term of objective function.

(17b), and (17c), 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 (17d) shows that the rate of each UE is more than a threshold. (17e) and (17f) expressed the limited capacity of the fronthaul link, and the limited delay of receiving signal, respectively. Furthermore, (17g) ensures that each service is mapped to at least one slice. Also, (17h), guarantees that each slice (VNFs in two layers of slices) has been placed to one or more physical resources of DCs. Moreover, in (17i) $\mathcal{E} = \{M, S, C\}$ and the constraint supports that we have enough physical resources for VNFs of each slice.

The optimization problem in (17) can be decomposed into two independent optimization sub-problems 1 and 2 since the variables can be obtained independently and respectively. Firstly we need to solve sub-problem 1. After obtaining $\boldsymbol{P}$ and $\boldsymbol{A}$, sub-problem 2 can be solved by having the value of $\boldsymbol{A}$. The sub-problem 1 is as follow

$$\max_{\boldsymbol{P},\boldsymbol{A}} \quad \eta(\boldsymbol{P}, \boldsymbol{A}) \tag{18a}$$

$$\text{subject to} \quad \bar{p}_{r_{(s,i)}} \leq P_{max} \qquad \forall s, \forall i, \tag{18b}$$

$$p_{u_{(v,k)}} \geq 0 \qquad \forall v, \forall k, \tag{18c}$$

$$\mathcal{R}_{u_{(v,k)}} \geq \mathcal{R}^{min}_{u_{(v,k)}} \qquad \forall v, \forall k, \tag{18d}$$

$$C_{r_{(s,i)}} \leq C^{max}_{r_{(s,i)}} \qquad \forall s, \forall i, \tag{18e}$$

$$D_s \leq D^{max}_s \qquad \forall s, \tag{18f}$$

$$\textstyle\sum_{s=1}^{S} a_{v,s} \geq 1 \qquad \forall s. \tag{18g}$$

In sub-problem 2, $\boldsymbol{Y}$ is obtained. The sub-problem 2 is

$$\min_{\boldsymbol{y}} \quad \psi_{tot}(\boldsymbol{Y}) \tag{19a}$$

$$\text{s. t.} \quad \textstyle\sum_{d=1}^{D_c} \sum_{v=1}^{V} y_{s,d} a_{v,s} \geq 1 \times \sum_{v=1}^{V} a_{v,s} \forall s, \tag{19b}$$

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

III. Numerical Results

IV. Conclusion