Joint Power Allocation and Network Slicing in an End-to-End ORAN System

Mojdeh Karbalaee Motalleb, Vahid Shah-Mansouri, Salar Nouri Naghadeh
School of ECE, College of Engineering, University of Tehran, Iran
Email: {mojdeh.karbalaee, vmansouri,salar.nouri@ut.ac.ir}@ut.ac.ir,

***Abstract*—Many major telecommunication companies confirmed the unification of the xRAN Group with the C-RAN Alliance to establish a more flexible and openness radio access network which is the Open-RAN (ORAN) for the fifth generation of wireless technology. To increase energy efficiency and optimize the allocation of resources, Network Slicing (NS) is considered as the best method for the fifth generation (5G) in order to virtualize the common physical network into several logical end-to-end networks. Every slice consists of a part of core network resources, network functions, and radio access network resources as a functional end-to-end network. In this paper, we elaborate joint NS in RAN and Core of ORAN system to investigate the power of each User Equipment (UE), map slices to services and also map physical Data Centers (DC) to slices to jointly maximize energy efficiency and minimize consumption power of RUs and the cost of physical resources in a downlink channel. The problem is formulated as a mixed-integer optimization problem that can be decomposed into two independent sub-problems. Heuristic algorithms are proposed to each of sub-problems in order to map slices to services, optimize power consumption and map slices to physical resources to minimize the cost of total DCs simultaneously.**

***Index Terms*—ORAN, Network Slicing, Energy Efficiency, Data Center (DC)**

I. Introduction

Open RAN (ORAN), as the integration and expansion of C-RAN and xRAN, is expected to be a key technology in 5G networks to extensively enhance the RAN performance. The core idea of C-RAN is to split the radio remote head (RRU) from the baseband unit (BBU). Several BBUs operating on a cloud server will create a BBU-Pool, providing unified baseband signal processing with powerful computing capabilities [1]–[4]. On the other hand, xRAN technology has three fundamental features. The control plane is decoupled from user plane. Besides, a modular eNB software stack is built to operate on common-off-the-shelf (COTS) hardware. Moreover, open north-bound and south-bound interfaces are introduced [5].

ORAN virtualizes the elements of the radio access networks, separate them and define appropriate open interfaces for connection of these elements. Moreover, ORAN uses machine learning techniques to develop smarter RAN layers in its architecture. In an innovative ORAN system, the programmable RAN software is decoupled from hardware [6]. Open interface is one of the most crucial properties for ORAN to enable mobile network operators (MNOs) to define their own services. The concept of software defined network (SDN), which is the separation of control plane from user plane, is deployed in an intelligent ORAN architecture. Moreover, this separation promotes RRM to use non-realtime (RT) and near-realtime (NRT) RAN Intelligent Controller (RIC). In the ORAN architecture, distributed unit (DU) is a logical node having RLC/MAC/High-PHY layer. Moreover, central unit (CU) is a logical node having RRC, SDAP and PDCP. Also, radio unit (RU) is a logical node having LOW-PHY layer and RF processing [7]. ORAN introduces interfaces such as open fronthaul interface that connects DU and RU (i.e., E2 interface), and an A1 interface between orchestration/NMS layer containing the non-real-time RIC (RIC non-RT) function and the eNB/gNB containing the near-real-time RIC (RIC near-RT) function.

To evolve servicing in 5G networks, separation of elements of software and hardware of network is employed and referred as network functions virtualization (NFV). Virtual network function (VNF) are functional blocks of the system in this case. 5G networks are expected to host several services with different requirements simultaneously. Network slicing is considered as a solution for such demand. A network slice is a logical end-to-end network which provide service with specific requirements. Multiple network slices run on the same physical network infrastructure which are managed and operate independently [8]. Slicing can be defined for the RAN, for the core network or both of these networks [9].

Based on their type of request, UEs are classified into various services, where UEs in the same service have similar service requests. In addition, each service is mapped to one or more slices based on the resources of the slices. Each VNF of the ORAN is mapped to one or some virtual machines (VMs) in the data center. For simplicity of analysis and without loss of generality, we assume a VNF is mapped to one VM and requires specific physical resources including storage, memory, and processing [2], [10], [11].

In [12], the evolution of RAN in the 5G network is described. The architecture of C-RAN and X-RAN is expressed and development of the ORAN from C-RAN and X-RAN is explained. Also requirements of the network in terms of capacity and latency is taken into account. In [13], the ORAN architecture is studied. The eight work groups is discussed including: focus areas of ORAN alliance which contains use-cases, overall architecture, open interfaces, and the virtualization and modularization of hardware and software. To the best of our knowledge, there is not work in the literature to model ORAN systems.

In this paper, as depicted in Figure 1, the downlink of the ORAN system is studied. The RAN is divided into three layers including RU, DU, and CU with open interfaces. DU

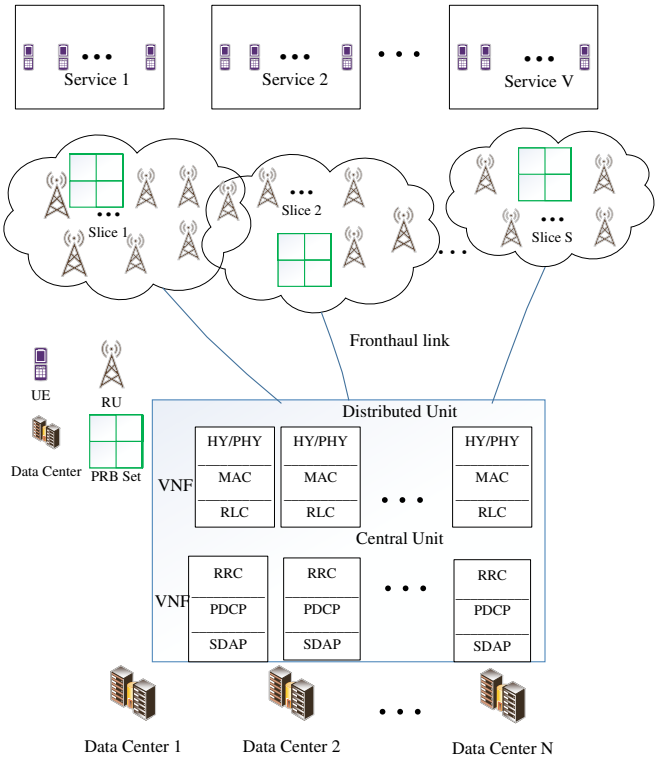

Fig. 1: Network sliced ORAN system

and CU are running on general purpose data centers. UEs are divided to different services according to their service requirements. RAN resources are decoupled to slices to provide requirements of services. Optimal power allocation and joint mapping of slices to services are applied. In addition, mapping slices to physical resources is taken to account. The contribution of the paper is expressed as follow:

- In this paper, joint network slicing and resource allocation is considered in an end to end ORAN system.
- UEs are categorized into different class of services based on their service requirements. RAN which contains RUs, PRBs, and VNFs ,in the two layer of processing, are classified into different group of slices.
- The problem is decomposed into two independent sub-problems.
- Heuristic algorithms are applied for the sub-problems to efficiently obtain the solution.

The rest of the paper is organized as follows. In Section II, system model which contains obtaining achievable rates, processing and transmission delay, and physical data center resources is expressed. Then, problem statement is explained and decomposed into two independent sub-problems. In Section III, heuristic algorithms for the sub-problems is presented. In Section IV, numerical results are provided for the problem.

II. System Model and Problem Formulation

In this section, first, we present the system model. Then, we obtain achievable 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 there are S slices Serving V services. Each Service $v \in \{1, 2, ..., V\}$ consists of U_v single-antenna user equipments (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 is mapped to a service, transmit signals cooperatively to all the UEs in a specific service [4], [15]. 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 slices.

B. The Achievable Rate

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

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

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

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

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

$$\mathbf{y}_{U_v} = \sum_{s=1}^{S} \sum_{k=1}^{K_s} \boldsymbol{H}^H_{\mathcal{R}_s,\mathcal{U}_v} \mathfrak{y}_{R_s} \zeta_{U_v,k,s} a_{v,s} + \boldsymbol{z}_{\mathcal{U}_v}, \tag{3}$$

where $\mathfrak{y}_{R_s} = \boldsymbol{W}_{\mathcal{R}_s,\mathcal{U}_v} \boldsymbol{P}^{\frac{1}{2}}_{U_v} \boldsymbol{x}_{\mathcal{U}_v} + \boldsymbol{q}_{\mathcal{R}_s}$ 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, $\boldsymbol{P}_{U_v} = \text{diag}\left(p_{u_{(v,1)}}, ..., p_{u_{(v,\mathcal{U}_v)}}\right)$.
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

$$\boldsymbol{W}_{\mathcal{R}_s,\mathcal{U}_v} = \boldsymbol{H}_{\mathcal{R}_s,\mathcal{U}_v}(\boldsymbol{H}^H_{\mathcal{R}_s,\mathcal{U}_v}\boldsymbol{H}_{\mathcal{R}_s,\mathcal{U}_v})^{-1}. \tag{4}$$

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

$$\begin{aligned} I_{u_{(v,i)}} = & \underbrace{\sum_{s=1}^{S}\sum_{n=1}^{S}\sum_{\substack{l=1 \\ l\neq i}}^{U_v}\gamma_1 p_{u_{(v,l)}} a_{v,s}\zeta_{u(v,i),n,s}\zeta_{u(v,l),n,s}}_{\text{(intra-service interference)}} \\ & + \underbrace{\sum_{\substack{y=1 \\ l\neq v}}^{V}\sum_{s=1}^{S}\sum_{n=1}^{S}\sum_{l=1}^{U_y}\gamma_2 p_{u_{(y,l)}} a_{y,s}\zeta_{u(v,i),n,s}\zeta_{u(y,l),n,s}}_{\text{(inter-service interference)}} \\ & + \underbrace{\sum_{s=1}^{S}\sum_{j=1}^{R_s}\sigma^2_{q_{r_{(s,j)}}}|\boldsymbol{h}_{r_{(s,j)},u_{(v,i)}}|^2 a_{v,s}}_{\text{(quantization noise interference)}}, \end{aligned} \tag{5}$$

where $\gamma_1 = |\boldsymbol{h}^H_{\mathcal{R}_s,u_{(v,i)}}\boldsymbol{w}_{\mathcal{R}_s,u_{(v,l)}}|^2$ and $\gamma_2 = |\boldsymbol{h}^H_{\mathcal{R}_s,u_{(v,i)}}\boldsymbol{w}_{\mathcal{R}_s,u_{(y,l)}}|^2$. 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 (1) and (2).
Let $\bar{p}_{r_{(s,j)}}$ denote the power of transmitted signal from the j^{th} RU in s^{th} slice. From (3), we have,

$$\bar{p}_{r_{(s,j)}} = \sum_{v=1}^{V}\boldsymbol{w}_{r_{(s,j)},\mathcal{U}_v}\boldsymbol{P}^{\frac{1}{2}}_{\mathcal{U}_v}\boldsymbol{P}^{H\frac{1}{2}}_{\mathcal{U}_v}\boldsymbol{w}^H_{r_{(s,j)},\mathcal{U}_v}a_{v,s} + \sigma^2_{q_{r(s,j)}}. \tag{6}$$

Nevertheless, the rate of users on the fronthual link between DU and the j^{th} RU in s^{th} slice is formulated as [3], [16]

$$C_{R_{(s,j)}} = \log\left(1 + \sum_{v=1}^{V}\frac{w_{r_{(s,j)},\mathcal{D}_s}\boldsymbol{P}^{\frac{1}{2}}_{\mathcal{U}_v}\boldsymbol{P}^{H\frac{1}{2}}_{\mathcal{U}_v}w^H_{r_{(s,j)},\mathcal{U}_v}a_{v,s}}{\sigma^2_{q_{r(s,j)}}}\right), \tag{7}$$

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

C. 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 Burke's 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 [2], [10], [11]. 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

$$\begin{aligned} d_{s_1} &= \frac{1}{\mu_1 - \alpha_s/M_{s,1}}, \\ d_{s_2} &= \frac{1}{\mu_2 - \alpha_s/M_{s,2}}. \end{aligned} \tag{8}$$

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 [2]. 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 [2], [10], [11], [14]. Therefore, the mean delay of the transmission layer is

$$d_{s_{tr}} = \frac{1}{R_{tot_s} - \alpha_s}; \tag{9}$$

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

$$D_s = d_{s_1} + d_{s_2} + d_{s_{tr}}\forall s. \tag{10}$$

D. 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

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

where $\bar{\Omega}^f_s \in \mathbb{C}^3$ and $\Omega^f_{M,s}, \Omega^f_{S,s}, \Omega^f_{C,s}$ 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

$$\bar{\Omega}^{tot}_{\mathfrak{z},s} = \sum_{f=1}^{M_{s_1}+M_{s_2}}\bar{\Omega}^f_{\mathfrak{z},s} \;\; \mathfrak{z} \in \{M, S, C\}. \tag{12}$$

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 τ_{M_j}, τ_{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.

E. Problem Statement

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

$$\eta(\boldsymbol{P}, \boldsymbol{A}) := \frac{\sum_{v=1}^{V}\sum_{k=1}^{U_v} \mathcal{R}_{u_{(v,k)}}}{\sum_{s=1}^{S}\sum_{i=1}^{R_s} \bar{p}_{r_{(s,i)}}} = \frac{\mathfrak{R}_{tot}(\boldsymbol{P},\boldsymbol{A})}{P_r^{tot}(\boldsymbol{P},\boldsymbol{A})}, \tag{13}$$

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 toal 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

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

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

$$\psi_{tot} = \phi_{tot} - \nu \textstyle\sum_{d=1}^{D_c}\sum_{v=1}^{V} y_{s,d}a_{v,s} \tag{14}$$

where, ν is a design variable to value between the first term of (14) 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,

$$\max_{\boldsymbol{P},\boldsymbol{A},\boldsymbol{Y}} \quad \eta(\boldsymbol{P}, \boldsymbol{A}) + \frac{1}{\psi_{tot}(\boldsymbol{Y})} \tag{15a}$$

$$\text{subject to} \quad \bar{p}_{r_{(s,i)}} \le P_{max} \quad \forall s, \forall i, \tag{15b}$$

$$p_{u_{(v,k)}} \ge 0 \quad \forall v, \forall k, \tag{15c}$$

$$\mathcal{R}_{u_{(v,k)}} \ge \mathcal{R}^{min}_{u_{(v,k)}} \quad \forall v, \forall k, \tag{15d}$$

$$C_{r_{(s,i)}} \le C^{max}_{r_{(s,i)}} \quad \forall s, \forall i, \tag{15e}$$

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

$$\textstyle\sum_{s=1}^{S} a_{v,s} \ge 1 \quad \forall s, \tag{15g}$$

$$\textstyle\sum_{d=1}^{D_c}\sum_{v=1}^{V} y_{s,d}a_{v,s} \ge 1 \times \sum_{v=1}^{V} a_{v,s} \forall s, \tag{15h}$$

$$\bar{\Omega}^{tot}_{\mathfrak{z},s} = \textstyle\sum_{f=1}^{F_s} \bar{\Omega}^{f}_{\mathfrak{z},s} \le \sum_{d=1}^{D_c} y_{s,d}\tau_{\mathfrak{z}d} \forall s, \forall \mathfrak{z} \in \mathcal{E}; \tag{15i}$$

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. (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 is more than a threshold. (15e) and (15f) expressed the limited capacity of the fronthaul link, and the limited delay of receiving signal, respectively. Furthermore, (15g) ensures that each service is mapped to at least one slice. Also, (15h), guarantees that each slice (VNFs in two layers of slices) has been placed to one or more physical resources of DCs. Moreover, in (15i) $\mathcal{E} = \{M, S, C\}$ and the constraint supports that we have enough physical resources for VNFs of each slice.

The optimization problem in (15) can be decomposed into two independent optimization problems A and B since the variables can be obtained independently and respectively. Firstly we need to solve problem A. After obtaining $\boldsymbol{P}$ and $\boldsymbol{A}$, problem B can be solved by having the value of $\boldsymbol{A}$. The problem A is as follow

$$\max_{\boldsymbol{P},\boldsymbol{A}} \quad \eta(\boldsymbol{P}, \boldsymbol{A}) \tag{16a}$$

$$\text{subject to} \quad \bar{p}_{r_{(s,i)}} \le P_{max} \quad \forall s, \forall i, \tag{16b}$$

$$p_{u_{(v,k)}} \ge 0 \quad \forall v, \forall k, \tag{16c}$$

$$\mathcal{R}_{u_{(v,k)}} \ge \mathcal{R}^{min}_{u_{(v,k)}} \quad \forall v, \forall k, \tag{16d}$$

$$C_{r_{(s,i)}} \le C^{max}_{r_{(s,i)}} \quad \forall s, \forall i, \tag{16e}$$

$$D_s \le D_s^{max} \quad \forall s, \tag{16f}$$

$$\textstyle\sum_{s=1}^{S} a_{v,s} \ge 1 \quad \forall s. \tag{16g}$$

In problem B, $\boldsymbol{Y}$ is obtained. The problem B is

$$\min_{\boldsymbol{y}} \quad \psi_{tot}(\boldsymbol{Y}) \tag{17a}$$

$$\text{s. t.} \quad \textstyle\sum_{d=1}^{D_c}\sum_{v=1}^{V} y_{s,d}a_{v,s} \ge 1 \times \sum_{v=1}^{V} a_{v,s} \forall s, \tag{17b}$$

$$\bar{\Omega}^{tot}_{\mathfrak{z},s} = \textstyle\sum_{f=1}^{F_s} \bar{\Omega}^{f}_{\mathfrak{z},s} \le \sum_{d=1}^{D_c} y_{s,d}\tau_{\mathfrak{z}d} \forall s, \forall \mathfrak{z} \in \mathcal{E}; \tag{17c}$$

III. Heuristic Method

In this subsection, a heuristic method is proposed to solve the problem in (16) which is non-convex and computationally hard. To make it tractable, we divide problem (16) into two different part that can be solved iteratively. In the first part of sub-problem A, we obtain $\boldsymbol{A}$ by fixing $\boldsymbol{P} = P_{max}$ in (16). Also, we set $\eta = 0$. Afterward, by achieving $\boldsymbol{A}$, in the second part, we find $\boldsymbol{P}$ by using (16). This part of the problem can be approximated and converted to a convex problem, so the problem can be solved by convex methods. After solving $\boldsymbol{P}$, η is updated. Then in the next iteration, with new $\boldsymbol{P}$ and η, the two parts of the problems are solved. We repeat this procedure until the algorithm converges.

A. First Part of Sub-Problem A

Two different methods are applied to acquire $\boldsymbol{A}$. The details of the heuristic algorithm are represented in algorithm 1.

Algorithm 1 Mapping Slice to Service

Sort services according to the number of UEs in it and their requirements in the descending order.
Sort slices according to the weighted linear combination of number of PRBs, RUs and VNFs in two layers and the capacity of their resources in the descending order.
for $i \leftarrow 1$ to S **do**
  for $j \leftarrow 1$ to V **do**
    Set $a_{i,j} = 1$
    Obtain Parameters of Systems (power and rate of UEs, rate of fronhaul links, power of RUs)
    if conditions (15b), (15c), (15d) and (15e) is not applied **then**
      Set $a_{i,j} = 0$;
    else
      break from inner loop;
    end if
  end for
end for

B. Second Part of Sub-Problem A

In this part, by assuming that $\boldsymbol{A}$ is fixed, the power of UEs in each service is achieved.

Theorem 1. *The optimum energy efficiency is achieved if*

$$\max_{\boldsymbol{P}}(\mathfrak{R}_{tot}(\boldsymbol{P})-\eta^* P_{r_{tot}}(\boldsymbol{P})) = \mathfrak{R}_{tot}(\boldsymbol{P}^*)-\eta^* P_{r_{tot}}(\boldsymbol{P}^*) = 0. \tag{18}$$

Proof. See [17, Appendix A] □

The second sub-problem can be solved using the Lagrangian function and iterative algorithm. Since, Interference is a function of the power of UEs, to make it tractable, we assume an upper bound $\bar{I}_{u_{(v,i)}}$ for interference. In order to make (16) as a standard form of a convex optimization problem, it is required to change the variable of equations (16e) and (16f) ($p_{r_{(s,i)}} = \sigma^2_{q_{r(s,j)}} \times 2^{C_{r_{(s,i)}}}$ and $1/(D_s - d_{s_1} + d_{s_2}) + \alpha_s$ respectively). Assume $\boldsymbol{\lambda}$, $\boldsymbol{\mu}$, $\boldsymbol{\xi}$, and $\boldsymbol{\kappa}$ are the matrix of Lagrangian multipliers that have non-zero positive elements. The Lagrangian function is written as follow

$$\mathcal{L}(\boldsymbol{P};\boldsymbol{\lambda},\boldsymbol{\chi},\boldsymbol{\mu},\boldsymbol{\xi},\boldsymbol{\kappa}) = \sum_{v=1}^{V}\sum_{k=1}^{U_v}\bar{\mathcal{R}}_{u_{(v,k)}} - \eta\sum_{v=1}^{V}\sum_{i=1}^{\mathcal{R}_s}\bar{p}_{r_{(s,i)}} \tag{19a}$$

$$+\sum_{v=1}^{V}\sum_{k=1}^{U_v}\lambda_{u_{(v,k)}}(\bar{\mathcal{R}}_{u_{(v,k)}} - \mathcal{R}^{max}_{u_{(v,k)}}) \tag{19b}$$

$$-\sum_{s=1}^{S}\sum_{i=1}^{R_s}\mu_{r_{(s,i)}}(\bar{p}_{r_{(s,i)}} - P_{max}) \tag{19c}$$

$$-\sum_{s=1}^{S}\sum_{i=1}^{R_s}\xi_{r_{(s,i)}}(\bar{p}_{r_{(s,i)}} - \sigma^2_{q_{r(s,i)}}2^{C^{max}_{r_{(s,i)}}}). \tag{19d}$$

$$+\sum_{v=1}^{V}\sum_{k=1}^{U_v}\kappa_{u_{(v,k)}}\sum_{s=1}^{S}(R_{u_{(v,k)}} - \mathfrak{D}_{\mathfrak{s}})a_{v,s}. \tag{19e}$$

where, $\mathfrak{D}_{\mathfrak{s}} = \frac{1}{D^{max}_s - d_{s_1} - d_{s_2}} + \alpha_s$. Optimal power is obtained from (19)

$$p^*_{u(v,i)} = [\frac{\mathfrak{y}_{u(v,i)}\mathfrak{w}_{u(v,i)} - \mathfrak{x}_{u(v,i)}\mathfrak{z}_{u(v,i)}}{\mathfrak{x}_{u(v,i)}\mathfrak{w}_{u(v,i)}}]^+ \tag{20}$$

where, $\mathfrak{y}_{u(v,i)} = (\lambda_{u(v,i)} + \kappa_{u_{(v,k)}} + 1)\frac{B}{Ln2}$ and $\mathfrak{w}_{u(v,i)} = \sum_{s=1}^{S}|\mathbf{h}^H_{R_s,u(v,i)}\mathbf{w}_{R_s,u(v,i)}|^2 a_{v,s}$. Also $\mathfrak{z}_{u(v,i)} = BN_0 + \bar{I}_{u(v,i)}$ and $\mathfrak{x}_{u(v,i)} = \sum_{s=1}^{S}\sum_{i=1}^{R_s}(\mu_{r_{u(s,i)}} + \xi_{r_{(s,i)}} + \eta)||w_{r_{(s,j)},u_{(v,i)}}||^2$. By using sub-gradient method, the optimal power $\boldsymbol{P}$ is obtained [15].

C. Solving two part of Sub-problem A iteratively

In (III-A) and (III-B), the details of solving each part of the sub-problem are depicted. Firstly, we obtain $\boldsymbol{A}$ by fixing $\boldsymbol{P} = P_{max}$ in the problem (16) and using algorithm (1). Also we fixed $\eta = 0$. Afterward, $\boldsymbol{A}$ is achieved from algorithm 1. Then, in the second part, we acquire $\boldsymbol{P}$ using the sub-gradient method. After solving $\boldsymbol{P}$, η is updated. Then in the next iteration, with new $\boldsymbol{P}$ and η the two parts of the problems is solved until the algorithm converges. Here, the algorithm of solving sub-problem A is shown in algorithm (2)

Algorithm 2 Joint Network Slicing and Power Allocation

Set the maximum number of iterations I_{max}, convergence condition ϵ_η and the initial value $\eta^{(1)} = 0$
Set $\boldsymbol{P} = \boldsymbol{P}_{max}$
for $counter \leftarrow 1$ to I_{max} **do**
  Achieve $\boldsymbol{A}$ by applying Algorithm (1)
  Obtain $\boldsymbol{P}$ by using sub-gradient method which is mentioned in (III-B).
  if $\mathfrak{R}_{tot}(\boldsymbol{P}^{(i)}, \boldsymbol{A}^{(i)}) - \eta^{(i)} P_{r_{tot}}(\boldsymbol{P}^{(i)}, \boldsymbol{A}^{(i)}) < \epsilon_\eta$ **then**
    Set $\boldsymbol{P}^* = \boldsymbol{P}^{(i)}$, $\boldsymbol{A}^* = \boldsymbol{A}^{(i)}$ and $\eta^* = \eta^{(i)}$;
    break;
  else
    $i = i + 1$, Setting $\boldsymbol{P} = \boldsymbol{P}^{(i)}$;
  end if
end for

D. Sub-Problem B

In this subsection, we would like to solve (17), which is the placement of virtual resources to physical resources in order to minimize the cost function ψ_{tot}. To achieve optimum $\boldsymbol{Y}$ heuristic algorithm is applied. The details of the heuristic algorithm are written in algorithm (3). In this algorithm, firstly, we sort slices and DCs according to their sum-weighted of their requirements (line 1 and line 2 of algorithm 3). We define a weighted parameter for $\Omega^{tot}_{\mathfrak{z},s}$ and $\tau^{\mathfrak{z}}_j$ as follow

$$\begin{aligned}\hat{\Omega}^{tot}_s &= w_M\bar{\Omega}^{tot}_{M,s} + w_S\bar{\Omega}^{tot}_{S,s} + w_C\bar{\Omega}^{tot}_{C,s}\\ \hat{\tau}_j &= w_M\tau^M_j + w_S\tau^S_j + w_C\tau^C_j,\end{aligned} \tag{21}$$

where, $\boldsymbol{w} = \{w_M, w_S, w_C\}$ is the weight of memory, storage and CPU. Secondly, we start mapping from the most needed slices to the DC with the most physical resources (from line 4 to line 11 of algorithm 3). After mapping DCs to slices, If a slice does not admit to a specific DC, it remains for the next placement. In the next placement the remaining slices, map to more than one DC according to their requirements. (from line 12 to line 25 of algorithm 3). At the end, if DC with the lowest physical resources is free and can be served the slices of a DC with the highest physical resource, the slices remapped to new DC with the lowest physical resource since it has the lowest power consumption (line 26 of algorithm 3).

Algorithm 3 Plecement of Physical resources into Virtual resources

1: Sort Slices according to $\hat{\Omega}_s^{tot}, \forall s$ in descending order.
2: Sort DCs according to $\hat{\tau}_j, \forall j$ in descending order.
3: $\boldsymbol{Y} = \boldsymbol{0}$
4: **for** $d \leftarrow 1$ to D_c **do**
5: **for** $s \leftarrow 1$ to S **do**
6: **if** $\sum_{d=1}^{D_c} y_{s,d} == 0$ and $\bar{\Omega}_{\mathfrak{z}(s)}^{tot} \leq \tau_{\mathfrak{z}_j} \forall \mathfrak{z}, \forall s$ **then**
7: Set $y_{s,d} = 1$;
8: $\tau_j^{\mathfrak{z}} \leftarrow \tau_j^{\mathfrak{z}} - \bar{\Omega}_{\mathfrak{z},s}^{tot}$ $\mathfrak{z} \in \{M, S, C\}$
9: **end if**
10: **end for**
11: **end for**
12: $ind_{rem} = \{s | (\sum_{d=1}^{D_c} y_{s,d} == 0)\}$
13: Sort remaining amount of DCs same as before in descending order.
14: Sort remaining slices same as before in descending order.
15: **for** $r \leftarrow 1$ to S_{rem} **do**
16: **for** $n \leftarrow 1$ to D_c **do**
17: Set $y_{s,d} = 1$;
18: $\bar{\Omega}_{\mathfrak{z},s}^{tot} \leftarrow \bar{\Omega}_{\mathfrak{z},s}^{tot} - \tau_j^{\mathfrak{z}}$
19: **if** $\bar{\Omega}_s^{tot} == 0$ **then**
20: Set $y_{s,d} = 1$;
21: $\tau_j^{\mathfrak{z}} \leftarrow \tau_j^{\mathfrak{z}} - \bar{\Omega}_{\mathfrak{z},s}^{tot}$ $\mathfrak{z} \in \{M, S, C\}$
22: break inner loop
23: **end if**
24: **end for**
25: **end for**
26: Remapping DCs must be done to prevent wasting Energy

IV. Numerical Results

In this section, simulation and numerical results for the main problem are depicted. In Fig. 2, energy efficiency is depicted for two different numbers of services with the different mean number of UEs in each service (the parameters for simulation listed in table I). The optimal method is obtained by using MOSEK toolbox to obtain $\boldsymbol{A}$ and CVX toolbox to obtain $\boldsymbol{P}$ and iteratively updates $\boldsymbol{A}$, η and $\boldsymbol{P}$. The optimal method is 0.1 bits/J/Hz better than the proposed method for $V = 6$ and $E[U_v] = 10$,

TABLE I: Simulation Parameter

Parameter	Value
Noise power	-174dBm
Bandwidth	120 KHZ
Maxmimun transmit Power of each RU	40dBm
Minimum delay	300usec
Maximum fronthaul capacity	200 bits/sec/Hz
Minimum data rate	10 bits/sec/Hz

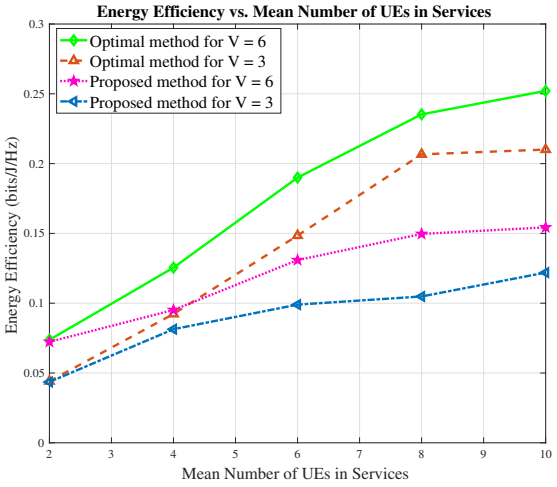

Fig. 2: Energy Efficiency vs. Mean Number of UEs in each Service

and also, 0.09 bits/J/Hz better than the proposed method for $V = 3$ and, $E[U_v] = 10$. As it is shown, the Energy Efficiency is increased as a mean number of UEs rises. In Fig. 3, the ratio of admitted slices is demonstrated for two different numbers of DCs with the different number of slices (the parameters for simulation listed in table II and also we set $w_C = 320$, $w_S = 100$, $w_M = 1$). In this simulation, the second term of (14), is important and the designed parameter ν is supposed to be high. Also, it is assumed that just one DC can serve each slice, and each slice is not admitted by more than one DC. The proposed method is based on Algorithm 3, and the optimal method is done by the MOSEK toolbox. When we have two DCs, the proposed method and optimal method have approximately the same ratio of admitted slices. But by increasing the number of DCs to five, the performance of the proposed method reduced. Using five DCs, the difference between the

TABLE II: Simulation Parameter

Parameter	Value
Mean of CPU for DCs	320GHz
Mean of Memory for DCs	1T
Mean of Storage for DCs	100T
Mean of CPU for Slices	32GHz
Mean of Memory for Slices	100G
Mean of Storage for Slices	10T

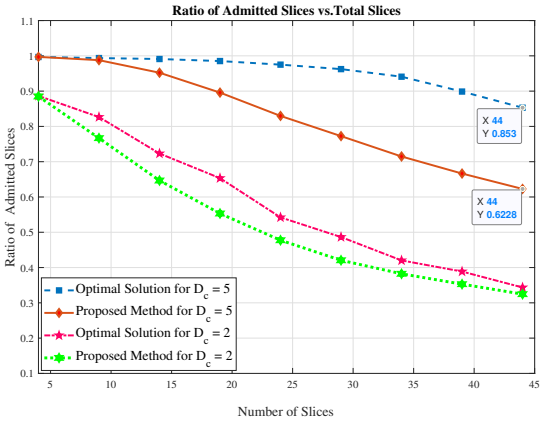

Fig. 3: Ratio of Admitted Slices connected to just one DC vs. Total slices

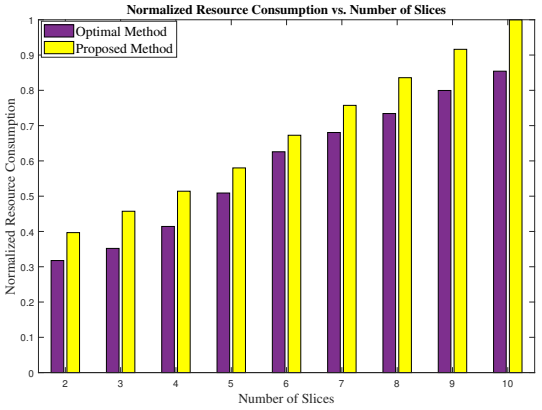

Fig. 4: Normalized Resource Consumption vs. Number of Slices

proposed method and the optimal method in the worst case (44 slices) is about 23 percentage. In Fig. 4, the normalized resource consumption is depicted due to the number of slices (the parameters for simulation listed in table II and also we set $w_C = 320$, $w_S = 100$, $w_M = 1$). In this simulation, suppose the number of DCs is entirely enough to cover all slices and also, the designed parameter ν is assumed to be low, so we focused on the first term of (14). However, if any slice remains, it can be served by more than one DCs. The optimality of the placement of DCs to slices is measured based on the power consumption of DCs. It is shown that how much resources of active DCs are not used. For ten slices, the difference between the optimal solution and the proposed solution is about 15 percent.

V. Conclusion

In this paper, joint network slicing and power allocation were considered in an ORAN system. It is assumed that UEs are classified based on their requirements into services. Also, there is a number of slices serving the services. Each slice consists of PRBs, RUs, and VNFs that support CU and DU. The limited fronthaul capacity is considered for the fiber links between RUs and DU. The target was to maximize the sum-rate and minimize the power consumption and energy cost of data centers simultaneously. The problem is decomposed into two sub-problems. Each sub-problem is solved separately by a heuristic algorithm. Using numerical results, we validate the heuristic method and study the performance of the algorithms.

References

[1] J. Wu, Z. Zhang, Y. Hong, and Y. Wen, "Cloud radio access network (c-ran): a primer," *IEEE Network*, vol. 29, no. 1, pp. 35–41, 2015.

[2] J. Tang, W. P. Tay, T. Q. Quek, and B. Liang, "System cost minimization in cloud ran with limited fronthaul capacity," *IEEE Transactions on Wireless Communications*, vol. 16, no. 5, pp. 3371–3384, 2017.

[3] O. Simeone, J. Kang, J. Kang, and S. Shamai, "Cloud radio access networks: Uplink channel estimation and downlink precoding," *arXiv preprint arXiv:1608.07358*, 2016.

[4] M. K. Motalleb, A. Kabiri, and M. J. Emadi, "Optimal power allocation for distributed mimo c-ran system with limited fronthaul capacity," in *2017 Iranian Conference on Electrical Engineering (ICEE)*. IEEE, 2017, pp. 1978–1982.

[5] (2018) xran forum merges with c-ran alliance to form oran alliance. [Online]. Available: https://www.businesswire.com/news/home/20180227005673/en/xRAN-Forum-Merges-C-RAN-Alliance-Form-ORAN

[6] B. Fletcher. (2019) Vodafone initiates first open ran trials in the u.k., challenging traditional vendors. [Online]. Available: https://www.fiercewireless.com/tech/vodafone-initiates-first-open-ran-trials-uk-challenging-traditional-vendors

[7] I. Scales. (2018) The open ran (oran) alliance formed to lever open 5g for 'other' technologies? and much more... [Online]. Available: https://www.o-ran.org/resources

[8] 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.

[9] X. Zhou, R. Li, T. Chen, and H. Zhang, "Network slicing as a service: enabling enterprises' own software-defined cellular networks," *IEEE Communications Magazine*, vol. 54, no. 7, pp. 146–153, 2016.

[10] P. Luong, C. Despins, F. Gagnon, and L.-N. Tran, "A novel energy-efficient resource allocation approach in limited fronthaul virtualized c-rans," in *2018 IEEE 87th Vehicular Technology Conference (VTC Spring)*. IEEE, 2018, pp. 1–6.

[11] P. Luong, F. Gagnon, C. Despins, and L.-N. Tran, "Joint virtual computing and radio resource allocation in limited fronthaul green c-rans," *IEEE Transactions on Wireless Communications*, vol. 17, no. 4, pp. 2602–2617, 2018.

[12] P. Sehier, P. Chanclou, N. Benzaoui, D. Chen, K. Kettunen, M. Lemke, Y. Pointurier, and P. Dom, "Transport evolution for the ran of the future," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 11, no. 4, pp. B97–B108, 2019.

[13] S. A. T. Kawahara and A. U. R. Matsukawa, "O-ran alliance standardization trends," 2019.

[14] K. Guo, M. Sheng, J. Tang, T. Q. Quek, and Z. Qiu, "Exploiting hybrid clustering and computation provisioning for green c-ran," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 12, pp. 4063–4076, 2016.

[15] P.-R. Li, T.-S. Chang, and K.-T. Feng, "Energy-efficient power allocation for distributed large-scale mimo cloud radio access networks," in *2014 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2014, pp. 1856–1861.

[16] S.-H. Park, O. Simeone, O. Sahin, and S. S. Shitz, "Fronthaul compression for cloud radio access networks: Signal processing advances inspired by network information theory," *IEEE Signal Processing Magazine*, vol. 31, no. 6, pp. 69–79, 2014.

[17] D. W. K. Ng, E. S. Lo, and R. Schober, "Energy-efficient resource allocation for secure ofdma systems," *IEEE Transactions on Vehicular Technology*, vol. 61, no. 6, pp. 2572–2585, 2012.