Joint Power Allocation and Network Slicing In an End to End O-RAN System

1st Mojdeh Karbalaee Motalleb
Electrical and Computer Engineering
Tehran University
Tehran, Iran
mojdeh.karbalaee@ut.ac.ir

2nd Vahid Shah-Mansouri
Electrical and Computer Engineering
Tehran University
Tehran, Iran
vmansouri@ut.ac.ir

3rd Salar Nouri Naghadeh
Electrical and Computer Engineering
Tehran University
Tehran, Iran
salar.nouri@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 (O-RAN) for the fifth generation of wireless technology.

To increase enrgy 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 O-RAN 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 RRHs 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 due to the fact that sub-problems are independent. 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—O-RAN, Network Slicing, Energy Efficiency, Data Center (DC)

I. INTRODUCTION

O-RAN, which is the integration and expansion of C-RAN and xRAN, is expected to be a key technology in 5G Network, to enhance RAN performance. The idea of O-RAN comes from two opinions. Firstly, according to real-time analytic used for artificial intelligence systems, the radio access networks must be evolved to be more intelligent and flexible than before. Also O-RAN uses deep learning techniques to develop smarter layers in its architecture. Furthermore, O-RAN can virtualize elements of the network with appropriate interfaces. In an innovative O-RAN system, the programmable RAN software is decoupled from hardware, which can be run on any specific processing platforms, in order to be more flexible for mobile network operators MNOs especially for mobile virtual network operators (MVNOs) [1]. Openness is the specific feature of O-RAN. Open interface is one of the most crucial properties for O-RAN to enable MNOs and MVNOs, to define their own services. The concept of software defined network (SDN), which is the separation of control plane (CP), from user plane

(UP), is deployed in an intelligent O-RAN architecture. The seperation of CP/UP is developed through the E1 interface. Moreover, this separation, promote RRM by defining Non-RT and Near-RT) RAN Intelligent Controller (RIC) with the A1 and E2 interfaces . O-RAN introduced interfaces such as open fronthaul interface that maps DU and RRU, an 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. [2]

The core idea of C-RAN is to split the radio remote head (RRH) from baseband unit (BBU). Several BBUs operating on a cloud server will create a BBU-Pool, providing unified baseband signal processing with powerful computing capabilities [3]. To communicate between BBU-Pool and RRHs, the fronthaul fiber link interface, is assumed with limited capacity. The compression of a message passed through these links is a consequence of limited fronthaul capacity [4]–[6].

xRAN technology, released in April 2018 as the next generation of RAN, 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 [7].

To evolve servicing in 5G, separation of elements of software and hardware of network is deployed in network functions virtualization (NFV) technology. In this technology, the functionality of networks is virtualized and divided into blocks of virtual network function (VNF). The responsibility of wireless systems of the fifth generation covers wide range types of services. In order to provide the requirements of these services, network slicing (NS) is implemented to virtualize the shared physical network into several logical end-to-end networks. In the NS, a specific physical network hardware, run multiple virtual networks. Each virtual network is defined as a slice of network [8]. Three different types of NS are introduced in [9] contains Core Slicing, RAN Slicing and Core-RAN Slicing. In Core-RAN Slicing, each slice of RAN is mapped to slices of Core. Also, UEs classified into a group of services according to their requirements. In addition, each service is mapped to one or more Core-RAN slices based on the resource of slices. Using cloud-computing in
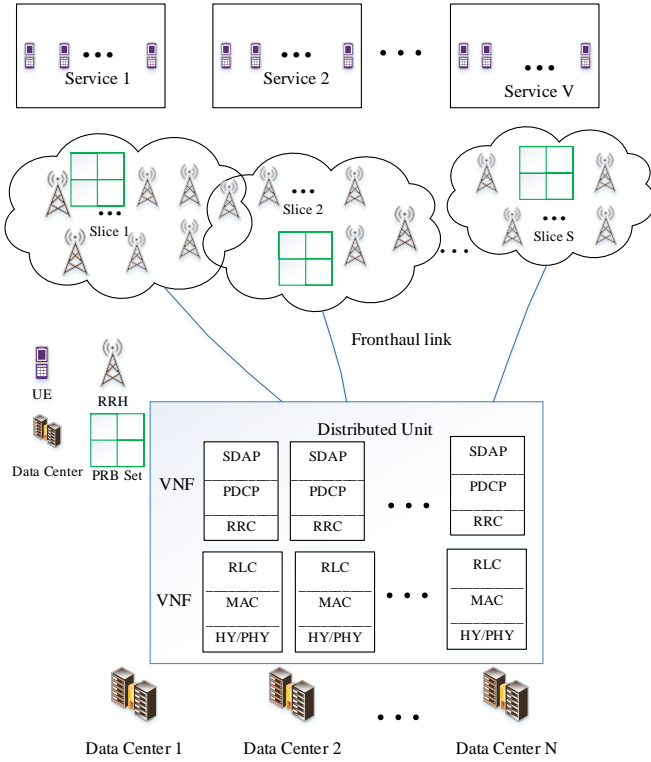
Fig. 1: Network sliced O-RAN system

is taken to account.

The contribution of the paper is expressed as follow:

- In this paper, joint RAN slicing and resource allocation is considered in an end to end O-RAN system.
- UEs are categorized into different class of services based on their service requirements. RAN which contains RRHs, 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 resources is expressed. Also, problem statements is explained and decomposed to two different sub-problems. In Section III, heuristic algorithms for the sub-problems is presented. In Section IV, numerical results, is depicted for the sub-problems.

II. SYSTEM MODEL AND PROBLEM FORMULATION

In this section, first, we first present the system model, Then we obtain achievable rates and delays for the downlink (DL) of the O-RAN architecture. Afterward, 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 RRHs and K_s Physical Resource Blocks (PRBs). All RRHs in a slice that is mapped to a service, transmit signals cooperatively to all the UEs in specific service [6], [14]. Each RRH $r \in \{1, 2, ..., R\}$ is mapped to BBU pool via an optical fiber link with limited fronthaul capacity. Suppose we have two processing layers in the BBU-Pool of O-RAN system which contain VNFs. The lower layer consist of high-PHY and MAC, and the upper layer consist of RLC, PDCP and SDAP. Assume we have M_1 VNFs in the first layer and M_2 VNFs in the second layer for processing data . Each VNF in both layers connects to one or more slices. So in s^{th} slice, there are $M_{s,1}$ VNFs in the first layer and $M_{s,2}$ VNFs in the second layer. All VNFs in first and second layer has the computational capacity that is equal to μ_1 and μ_2, respectively. Also, RRHs and PRBs can serve more than one slices.

B. The Achievable Rate

The achievable data rate for the ith UE in the vth service can be written as

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

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

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

BBU-Pool, the performance of the system is enhanced by virtualizing resources into virtual machines (VMs). Each VM has a computing processor in order to be run by the Virtual Network Functions (VNFs) which processed arrival data. Also, VNFs are mapped to physical resources through NS techniques [4], [10], [11].

The compression of messages passed through the fronthaul link due to the limited capacity of these links is considered in [5], [12]. In [6] power allocation in C-RAN architecture, with limited fronthaul capacity in clustered network is studied. In [8], dynamic network slicing is considered in Heterogeneous CRAN (H-CRAN) to maximize the weighted sum-rate. Also, two level network slicing (upper and lower level) is done. The admission of UEs is maximized using the heuristic algorithms. In addition, the allocation of BBU capacity to UEs is considered.
In [4], [10], [11], [13] minimization of cost obtained by power consumption, also cloud processing and limited power for each RRH is considered. Furthermore, the processing delay of each VM and wireless transmission delay taken account. In [4], the processing and transmission delay is modeled as a M/M/1 queue. In [13], M/D/1 queue is modeled for transmission delay.

In this paper, as depicted in figure 1, the downlink of the O-RAN system is assumed. UEs are divided to different groups according to their service requirements. Also RAN is decoupled to slices to provide requirements of services. Optimal power allocation and joint mapping slices to services are applied. In addition, mapping slices to physical resources

where $p_{u(v,i)}$ represents the transmission power allocated by RRHs 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 RRHs in the sth slice to the ith 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 RRHs in the sth slice to the ith 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 which is given by (3)

$$\mathbf{y}_{U_v} = \sum_{s=1}^{S} \sum_{k=1}^{K_s} \mathbf{H}_{R_s,\mathcal{U}_v}^{H} (\mathbf{W}_{R_s,\mathcal{U}_v} \mathbf{P}_{U_v}^{\frac{1}{2}} \mathbf{x}_{\mathcal{U}_v} + \mathbf{q}_{R_s}) \zeta_{U_v,k,s} a_{v,s} + \mathbf{z}_{\mathcal{U}_v}; \quad (3)$$

Where $\mathbf{x}_{\mathcal{U}_v} = [x_{u(v,1)}, ..., x_{u(v,\mathcal{U}_v)}]^{T} \in \mathbb{C}^{R_s}$ depicts the transmitted symbol vector of UEs in v^{th} set of service, $\mathbf{z}_{U_v}$ is the additive Gaussian noise $\mathbf{z}_{U_v} \backsim \mathcal{N}(0, N_0 \mathbf{I}_{U_v})$ and N_0 is the noise power. In addition, $\mathbf{q}_{R_s} \in \mathbb{C}^{R_s}$ indicates the quantization noise which, is made from signal compression in BBU. Besides, $\mathbf{P}_{U_v} = \text{diag}(p_{u(v,1)}, ..., p_{u(v,\mathcal{U}_v)})$.

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. $\mathbf{H}_{R_s,\mathcal{U}_v} = [\mathbf{h}_{R_s,u(v,1)}, ..., \mathbf{h}_{R_s,v(v,\mathcal{U}_v)}]^{T} \in \mathbb{C}^{R_s \times U_v}$ shows the channel matrix between RRH 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, $\mathbf{W}_{\mathcal{R}_s,\mathcal{U}_v} = [\mathbf{w}_{\mathcal{R}_s,u(v,1)}, ..., \mathbf{w}_{\mathcal{R}_s,u(v,U_v)}] \in \mathbb{C}^{R_s \times U_v}$ is the zero forcing beamforming vector to minimize the interference which is indicated as below

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

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

$$I_{u(v,i)} = \underbrace{\sum_{s=1}^{S} \sum_{n=1}^{S} \sum_{\substack{l=1 \\ l \neq i}}^{U_v} \gamma_1 p_{u(v,l)} a_{v,s} \zeta_{u(v,i),n,s} \zeta_{u(v,l),n,s}}_{\text{(intra-service interference)}}$$

$$+ \underbrace{\sum_{\substack{y=1 \\ l \neq v}}^{V} \sum_{s=1}^{S} \sum_{n=1}^{S} \sum_{l=1}^{U_y} \gamma_2 p_{u(y,l)} a_{y,s} \zeta_{u(v,i),n,s} \zeta_{u(y,l),n,s}}_{\text{(inter-service interference)}} \quad (5)$$

$$+ \underbrace{\sum_{s=1}^{S} \sum_{j=1}^{R_s} \sigma_{q_{r(s,j)}}^{2} |\mathbf{h}_{r(s,j),u(v,i)}|^{2} a_{v,s}}_{\text{(quantization noise interference)}} .$$

where, $\gamma_1 = |\mathbf{h}_{\mathcal{R}_s,u(v,i)}^{H} \mathbf{w}_{\mathcal{R}_s,u(v,l)}|^2$ and $\gamma_2 = |\mathbf{h}_{\mathcal{R}_s,u(v,i)}^{H} \mathbf{w}_{\mathcal{R}_s,u(y,l)}|^2$. Moreover, $\sigma_{q_{r(s,j)}}$ is the variance of

quantization noise of jth RRH in s^{th} slice. As it is clear, Interference signal for each UE is comming 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 j^{th} RRH in s^{th} slice. from (3) we have,

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

Nevertheless, the rate of users on the fronthual link between BBU-Pool and the jth RRH in s^{th} slice is formulated as [5], [12]

$$C_{R(s,j)} = \log\left(1 + \sum_{v=1}^{V} \frac{w_{r(s,j),\mathcal{D}_s} \mathbf{P}_{\mathcal{U}_v}^{\frac{1}{2}} \mathbf{P}_{\mathcal{U}_v}^{H\frac{1}{2}} w_{r(s,j),\mathcal{U}_v}^{H} a_{v,s}}{\sigma_{q_{r(s,j)}}^2}\right), \quad (7)$$

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

C. Mean Delay

Let the packet arrival of UEs have a Poisson Process with arrival rate $\lambda_{u(v,i)}$ for the ith UE of the vth service. Therefore, the mean arrival data rate of UEs map to s^{th} slice in the first layer of the O-RAN which contains SDAP, PDCP, and RLC 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 vth service is mapped to the sth slice or not. Furthermore, the mean arrival data rate of the second layer which contains MAC and HY/PHY 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 there are dispatchers in each layer for each slice to divide the incoming traffic to VNFs equally. Each dispatcher first processes the incoming traffic and divides equally to each VNF [4], [10], [11]. Suppose the baseband processing of each VNF is depicted as a M/M/1 processing queue. Each packet is processed by one of the VNFs of slices. So the mean delay of the sth slice in the first and the second layer, which is modeled as M/M/1 queue from queuing theory, is formulated as follow ,respectively

$$\begin{aligned} d_{s_1} &= \frac{1}{\mu_1 - \alpha_s/M_{s_1}} \\ d_{s_2} &= \frac{1}{\mu_2 - \alpha_s/M_{s_2}}. \end{aligned} \quad (8)$$

where, μ_1 and μ_2 is the computational capacity of the first and the second layer respectively. In addition, $d_{s_{tr}}$ is the transmission delay for s^{th} slice as a result of wireless transmission. The arrival data rate of wireless transmission is equal to the arrival data rate of dispatchers (used as a traffic divider) for each slice divide traffic data and transmit divided traffic to VNFs. 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 [4], [10], [11]. Therefore, the mean delay of the transmission layer is

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

Where, $R_{tot_s} = \sum_{v=1}^{V} \sum_{u=2}^{U_v} a_{v,s} R_{u(v,i)}$. Mean delay of each slice is

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

D. Physical Data Center Resource

Each VNF requires physical resources that contain RAM, Memory, and CPU. Let, the required resources for VNF f in slice s is represented by a three-dimensional vector as follow

$$\bar{\Omega}_{(f,s)} = \{\Omega_{R_{f,s}}, \Omega_{M_{f,s}}, \Omega_{C_{f,s}}\};$$ (11)

Where, $\bar{\Omega}_{(f,s)} \in \mathbb{C}^3$ and $\Omega_{R_{f,s}}, \Omega_{M_{f,s}}, \Omega_{C_{f,s}}$ indicate the amount of required RAM, Memory, and CPU, respectively. Moreover total amount of required RAM, Memory, and CPU of all VNFs of a slice is a three-dimensional vector which is defined as

$$\bar{\Omega}_s^{tot} = \sum_{f=1}^{M_{s_1}+M_{s_2}} \bar{\Omega}_{(f,s)};$$ (12)

Also, there are D_c data centers (DC), which served VNFs. Each DC contains several servers that supply VNF's requirements. The amount of RAM, Memory, and CPU is denoted by τ_{R_j}, τ_{M_j} and τ_{C_j} for j^{th} DC, respectively.

$$\tau_j = \{\tau_{R_j}, \tau_{M_j}, \tau_{C_j}\};$$

Also we define a weighted parameter of τ_j as follow

$$\hat{\Omega}_s^{tot} = w_R \bar{\Omega}_{R_s}^{tot} + w_M \bar{\Omega}_{M_s}^{tot} + w_C \bar{\Omega}_{C_s}^{tot}$$
$$\hat{\tau}_j = w_R \tau_{R_j} + w_M \tau_{M_j} + w_C \tau_{C_j},$$ (13)

Where, $\boldsymbol{w} = \{w_R, w_M, w_C\}$ is the weight of RAM, Memory, and CPU that is used for Algorithm 3. In this system model, the placement of physical DC resources to VNFs is considered. $y_{s,d}$ is a binary variable indicates whether d^{th} DC is connected to VNFs of sth slice or not.

E. Problem Statement

One of the most important parameters to estimate the optimality of the system is energy efficiency which is represented as a sum-rate to sum-power

$$\eta(\boldsymbol{P}, \boldsymbol{A}) := \frac{\sum_{v=1}^{V} \sum_{k=1}^{U_v} \mathcal{R}_{u(v,k)}}{\sum_{s=1}^{S} \sum_{i=1}^{R_s} \bar{p}_{r_{(s,i)}}} = \frac{\mathfrak{R}_{tot}(\boldsymbol{P}, \boldsymbol{A})}{P_{r_{tot}}(\boldsymbol{P}, \boldsymbol{A})},$$ (14)

Where, $\sum_{s=1}^{S} \sum_{i=1}^{R_s} \bar{p}_{r_{(s,i)}}$ is the total power consumption of all RRHs in all slices. Also, $\sum_{v=1}^{V} \sum_{k=1}^{U_v} \mathcal{R}_{u(v,k)}$ 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

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

In this paper, the main goal is to maximize sum-rate and minimize sum-power (the total power of all RRHs and the total power consumption of baseband processing at all DCs) simultaneously, with the presence of constraints which is written as follow,

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

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

$$p_{u(v,k)} \geq 0 \qquad \forall v, \forall k,$$ (15c)

$$\mathcal{R}_{u(v,k)} \geq \mathcal{R}_{u(v,k)}^{min} \qquad \forall v, \forall k,$$ (15d)

$$C_{r_{(s,i)}} \leq C_{r_{(s,i)}}^{max} \qquad \forall s, \forall i,$$ (15e)

$$D_s \leq D_s^{max} \qquad \forall s,$$ (15f)

$$\sum_{s=1}^{S} a_{v,s} \geq 1 \qquad \forall s,$$ (15g)

$$\sum_{d=1}^{D_c} \sum_{v=1}^{V} y_{s,d} a_{v,s} \geq 1 \times \sum_{v=1}^{V} a_{v,s} \qquad \forall s,$$ (15h)

$$\bar{\Omega}_{\mathfrak{z}(s)}^{tot} = \sum_{f=1}^{F_s} \bar{\Omega}_{\mathfrak{z}(f,s)} \leq \sum_{d=1}^{D_c} y_{s,d} \tau_{\mathfrak{z}_d} \qquad \forall \mathfrak{z}, \forall s;$$ (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 connected to a VNFs of a slice or not. (15b) and (15c) indicate that the power of each RRH 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) depicts that the capacity of the fronthaul link is limited, and the delay of receiving signal should be less than a threshold, respectively. Furthermore, (15g) ensures that each service is connected to one or more slices. Also, (15h) guarantees that each slice (VNFs in two layers of slices) has been placed to one or more physical resources of DCs. Moreover, (15i) $\mathfrak{z} \in \{M, R, C\}$, supports that we have enough physical resources for VNFs of each slice.

The main optimizaiton problem which, is formulated as (15), can be decomposed into two independent optimization problems A and B since the variables can be obtained independently. Problem A is defined as

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

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

$$p_{u_{(v,k)}} \geq 0 \qquad \forall v, \forall k, \qquad (16c)$$

$$\mathcal{R}_{u_{(v,k)}} \geq \mathcal{R}_{u_{(v,k)}}^{min} \qquad \forall v, \forall k, \qquad (16d)$$

$$C_{r_{(s,i)}} \leq C_{r_{(s,i)}}^{max} \qquad \forall s, \forall i, \qquad (16e)$$

$$D_s \leq D_s^{max} \qquad \forall s, \qquad (16f)$$

$$\sum_{s=1}^{S} a_{v,s} \geq 1 \qquad \forall s. \qquad (16g)$$

Problem B is

$$\min_{\mathbf{y}} \quad \phi_{tot}(\mathbf{Y}) \qquad (17a)$$

$$\text{subject to} \quad \sum_{d=1}^{D_c} \sum_{v=1}^{V} y_{s,d} a_{v,s} \geq 1 \times \sum_{v=1}^{V} a_{v,s} \qquad \forall s, \qquad (17b)$$

$$\bar{\Omega}_{\mathfrak{z}(s)}^{tot} = \sum_{f=1}^{F_s} \bar{\Omega}_{\mathfrak{z}(f,s)} \leq \sum_{d=1}^{D_c} y_{s,d} \tau_{\mathfrak{z}d} \qquad \forall \mathfrak{z}, \forall s. \qquad (17c)$$

III. PROPOSED METHOD

In this subsection, the proposed method is applied to solve the optimization problem. We want to solve (16). Since the problem is non-convex and NP-Hard, iterative algorithm is applied. To solve the problem and obtain $\mathbf{A}$ and $\mathbf{P}$ we divide problem (16) into two different part that can be solved iteratively.

A. First Part of Sub-Problem A

Firstly, we need to obtain $\mathbf{A}$ by fixing $\mathbf{P}$ in the problem (16) and updating this parameter at the end of each iteration. Two different methods are applied to acquire $\mathbf{A}$. The details of the heuristic algorithm are represented in **Algorithm** (1).

Algorithm 1 Mapping Slice to Service

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

B. Second Part of Sub-Problem A

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

Theorem 1. η^* *which is the optimum energy efficiency can be achieved if*

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

Proof. See [15], Appendix A   □

The second subproblem can be solved using the Lagrangian function and iterative algorithm. Since, Interference is a function of the power of UEs, for simplicity, we assume an upper bound $\bar{I}_{u_{(v,i)}}$ for interference (the worst-case). 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). The Lagrangian function is written as follow

$$\mathcal{L}(\mathbf{P}; \boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\xi}, \boldsymbol{\kappa}) = \sum_{v=1}^{V} \sum_{k=1}^{U_v} \bar{\mathcal{R}}_{u_{(v,k)}} - \eta \sum_{v=1}^{V} \sum_{i=1}^{R_s} \bar{p}_{r_{(s,i)}}$$
$$+ \sum_{s=1}^{S} \sum_{k=1}^{U_v} \lambda_{u_{(v,k)}} (\bar{\mathcal{R}}_{d_{(s,k)}} - \mathcal{R}_{u_{(v,k)}}^{max})$$
$$- \sum_{s=1}^{S} \sum_{i=1}^{R_s} \mu_{r_{(s,i)}} (\bar{p}_{r_{(s,i)}} - P_{max})$$
$$- \sum_{s=1}^{S} \sum_{i=1}^{R_s} \xi_{r_{(s,i)}} (\bar{p}_{r_{(s,i)}} - \sigma_{q_{r(s,j)}}^2 2^{C_{r(s,i)}^{max}}).$$
$$+ \sum_{v=1}^{V} \sum_{k=1}^{U_v} \kappa_{u_{(v,k)}} \sum_{s=1}^{S} (R_{u_{(v,k)}} - \mathfrak{D}_\mathfrak{s}) a_{v,s}. \qquad (19)$$

Where, $\mathfrak{D}_\mathfrak{s} = \frac{1}{D_s^{max} - d_{s_1} - d_{s_2}} + \alpha_s$. Also, $\boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\xi},$ and $\boldsymbol{\kappa}$ are the matrix of Lagrangian multipliers that have non-zero positive elements. Optimal power is obtained from equation (19) as follow

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

where, $\mathfrak{y}_{u(v,i)} = (\lambda_{u(v,i)} + \kappa_{u(v,k)} + 1) \frac{B}{Ln_2}$ and $\mathfrak{w}_{u(v,i)} = \sum_{s=1}^{S} |\mathbf{h}_{R_s,u(v,i)}^H \mathbf{w}_{R_s,u(v,i)}|^2 a_{v,s}$. 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 $\mathbf{P}$ is obtained.

C. Solving two part of Sub-problem A iteratively

In (III-A) and (III-B), the details of solving each part of the subproblem are depicted. Firstly, by the fixing power of each UE, $\mathbf{A}$ is obtained using Algorithm (1). Then by using the sub-gradient method, (III-B) is solved and iteratively solves these problems until variables converge. Here, the algorithm of solving sub-problem A is shown in **Algorithm** (2)

Algorithm 2 Joint Network Slicing and Power Allocation

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

TABLE I: Simulation Parameter

Parameter	Value
Noise power	-174dBm
Bandwidth	120 KHZ
Maxmimun transmit Power of each RRH	40dBm
Maxmimun transmit Power of each UE	30dBm
Minimum delay	300usec
Maximum fronthaul capacity	20bits/sec/Hz
Minimum data rate	120 kbits/sec

D. Sub-Problem B

In this subsection, we want to solve (17), which is the placement of virtual resources to physical resources in order to minimize the number of using DCs. 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. Secondly, we start mapping from the most needed slices to the DC with the most physical resources. After mapping DCs to slices, if some slices are not admitted, we start admitting remaining slices to more than one DCs. In the end, if DC with the lowest physical resources is free and can be served instead of DC with the highest physical resource, the slices remapped to new DC with the lowest physical resource since it has the lowest power consumption. If a slice does not admit to a specific DC, it remains for the next placement. In next placement the remaining slices, map to more than one DC according to their requirements.

IV. SIMULATION

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 0.016 bits/J/Hz better than the proposed method for $V = 6$ and $E[U_v] = 10$, and also, 0.011 bits/J/Hz better than the proposed method for $V = 3$ and, $E[U_v] = 10$. As it is shown, the Energy Efficiency is

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 $\bar{\Omega}_{\hat{\jmath}(s)}^{tot} \leq \tau_{\hat{\jmath}} \forall \hat{\jmath}, \forall s$ **then**
7:             Set $y_{s,d} = 1$;
8:             $\tau_j \leftarrow \tau_j - \bar{\Omega}_s^{tot}$
9:         **end if**
10:     **end for**
11: **end for**
12: $\{ind_{remaining} = s | (\sum_{d=1}^{D_c} y_{s,d} == 0)\}$
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 $S_{remaining}$ **do**
16:     **for** $n \leftarrow 1$ to D_c **do**
17:         Set $y_{s,d} = 1$;
18:         $\bar{\Omega}_s^{tot} \leftarrow \bar{\Omega}_s^{tot} - \tau_j$
19:         **if** $\bar{\Omega}_s^{tot} == 0$ **then**
20:             Set $y_{s,d} = 1$;
21:             $\tau_j \leftarrow \tau_j - \bar{\Omega}_s^{tot}$
22:             break inner loop
23:         **end if**
24:     **end for**
25: **end for**
26: Remapping DCs must be done to prevent wasting Energy

TABLE II: Simulation Parameter

Parameter	Value
Mean of CPU for DCs	25.6GHz
Mean of RAM for DCs	128G
Mean of Memory for DCs	10T
Mean of CPU for Slices	3.2GHz
Mean of RAM for Slices	16G
Mean of Memory for Slices	1T

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 III). In this simulation, it is assumed that just one DC can serve each slice, and it 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 proposed method and the optimal method in the worst case (44 slices) is about 23 percentage. In Fig. 4, the normalized resource consumption
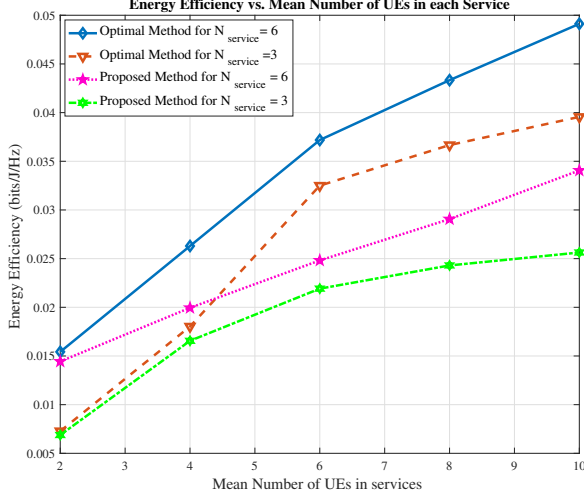
Fig. 2: Energy Efficiency vs. Mean Number of UEs in each Service

TABLE III: Simulation Parameter

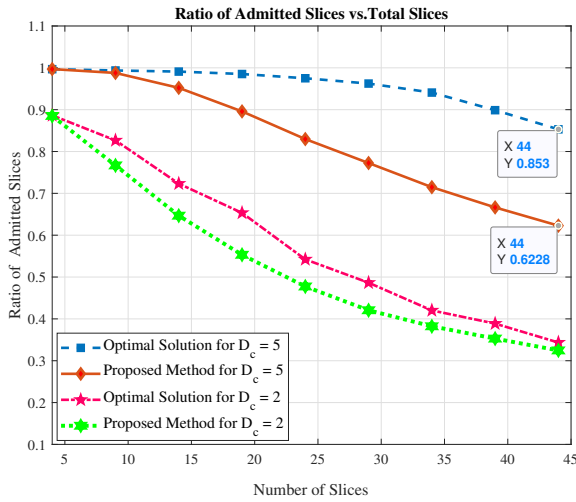| Parameter	Value
w_C	320
w_R	64
w_M	1

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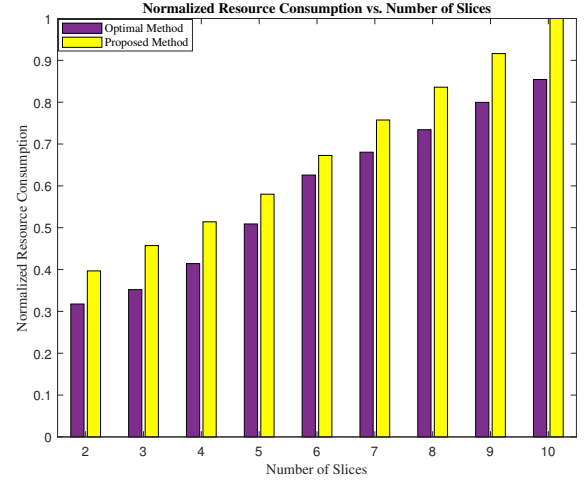

Fig. 4: Normalized Resource Consumption vs. Number of Slices

is depicted due to the number of slices (the parameters for simulation listed in table II and III). In this simulation, it is assumed that the number of DCs is entirely enough to cover all slices. The optimality of the placement of DCs to slices is measured. 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 is considered in an end to end O-RAN system. It is assumed that UEs are classified based on their service requirements. Also, there is a number of slices served the services. Each slice, is consist of PRBs, RRHs, and VNFs that run on VMs. The limited fronthaul capacity is considered for the fiber links between RRHs and BBU-pool. The target is 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-problems are solved separately by a heuristic algorithm. For the sub-problem A, energy efficiency vs. mean number of UEs in each service is depicted. The energy efficiency exceeded by increasing the mean number of UEs in each service. For the sub-problem B, two figures are shown. In the first one, the ratio of admitted slices that connect to just one DC, for a different number of slices is denoted. In the second figure, the normalized resource consumption of DCs is depicted. In each figure, the heuristic algorithm is compared with the optimal method and the difference between them is discussed.

REFERENCES

[1] 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

[2] 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

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] (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

[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] 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.

[13] 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.

[14] 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.

[15] 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.