

# Minimizing Energy Consumption for End-to-End Slicing in 5G Wireless Networks and Beyond

Shiva Kazemi Taskou\*, Mehdi Rasti\*,†, and Pedro H. J. Nardelli†

\*Department of Computer Engineering, Amirkabir University of Technology, Tehran, Iran

†School of Energy Systems, Lappeenranta-Lahti University of Technology, Lappeenranta, Finland

Email: {shiva.kt,rasti}@aut.ac.ir, pedro.nardelli@lut.fi

**Abstract**—End-to-End (E2E) network slicing enables wireless networks to provide diverse services on a common infrastructure. Each E2E slice, including resources of radio access network (RAN) and core network, is rented to mobile virtual network operators (MVNOs) to provide a specific service to end-users. RAN slicing, which is realized through wireless network virtualization, involves sharing the frequency spectrum and base station antennas in RAN. Similarly, in core slicing, which is achieved by network function virtualization, data center resources such as commodity servers and physical links are shared between users of different MVNOs. In this paper, we study E2E slicing with the aim of minimizing the total energy consumption. The stated optimization problem is non-convex that is solved by a sub-optimal algorithm proposed here. The simulation results show that our proposed joint power control, server and link allocation (JPSLA) algorithm achieves 30% improvement compared to the disjoint scheme, where RAN and core are sliced separately.

**Index Terms**—E2E slicing, network function virtualization, wireless network virtualization, resource allocation

## I. INTRODUCTION

Network slicing (NS) allows 5G wireless networks and beyond to offer diverse services and applications on a common infrastructure [1]. Through NS, a mobile network operator (MNO) slices its resources and infrastructure and leases them to mobile virtual network operators (MVNOs) who are responsible for providing a specific service to end-users [1]. In previous generations of wireless networks, radio access network (RAN) was a bottleneck to assure the quality of service (QoS), but with the advancement of RAN technologies in 5G and beyond, QoS is affected by the allocation of resources in both RAN and core. Hence, due to the end-to-end (E2E) concept of QoS, such as E2E delay in beyond 5G and 6G, NS should be performed in an E2E manner [2]. As stated in [3], in wireless networks, a half of tolerable E2E delay is related to RAN and the other half is related to core network. An E2E slice is a logical network consisting of resources of RAN, backhaul link, and resources of core network rent to an MVNO to provide a specific service to end-users [2].

Wireless network virtualization (WNV) and network function virtualization (NFV) are key technologies to realize RAN and core slicing, respectively. Employing WNV in RAN, spectrum frequency and base station antennas owned by MNO are shared between different slices [4]. After transmitting

This work is supported by the Academy of Finland: (a) ee-IoT n.319009, (b) EnergyNet n.321265/n.328869, and (c) FIREMAN n.326270/CHISTERA-17-BDSI-003; and by JAES Foundation via STREAM project.

users' data to core networks, the required network functions, such as packet inspection, packet routing and forwarding, etc., should be performed on users' data packets. In recent years, NFV technology has been proposed for core slicing, in which network functions are decoupled from the hardware and run as virtual network functions (VNFs) commodity servers in data centers [5]. In NFV, a sequence of VNFs connected via virtual links forms a service function chain (SFC). In NFV, data center resources, including server capacity and physical link bandwidth, are allocated to SFCs [5].

## A. Related Works

Some existing works studied slicing in RAN [6]–[8] and core network [9]–[12]. For instance, in [6] a power control problem to maximize energy efficiency subject to the users' minimum data rate and delay constraints is addressed using fractional programming and Lyapunov optimization. Also, the authors in [7] and [8] studied the co-existence of enhanced mobile broadband (eMBB) and ultra-reliable low latency communication (URLLC) slices. Specifically, in [7] a deep reinforcement learning framework is proposed to maximize the total data rate of eMBB slices such that URLLC users' reliability is guaranteed. Similarly, in [8], a heuristic based sub-channel allocation algorithm is proposed to slices in such a way that the minimum data rate is satisfied for eMBB and URLLC users.

Furthermore, core slicing for assuring QoS is considered in [9]–[12]. Particularly, the authors in [9] proposed a heuristic algorithm for allocating the servers and links to SFCs to minimize the cost of servers and links. Besides, E2E delay minimization problem has been studied in [10], where waiting delay in commodity servers queue is considered as E2E delay. Moreover, in [11], a heuristic algorithm is proposed to minimize the number of utilized servers such that a maximum tolerable delay is guaranteed for SFCs. Additionally, authors in [12] proposed a heuristic algorithm for minimizing the energy consumption and cost of utilized resources in a data center.

## B. Contributions and Paper Structure

To the best of our knowledge, none of the existing work has studied E2E slicing through which RAN and core resources are jointly allocated to the end-users. Moreover, due to high energy consumption in data centers and the environmental impacts of

energy, in this paper, we study E2E slicing with the aim of minimizing total energy consumption.

The remainder of this paper is structured as follows. In Section II, the system model and notations are introduced. The optimization problem is formally stated in Section III. In Section IV, we propose an algorithm to address the stated problem. Finally, the simulation results and conclusion are presented in Section V and Section VI, respectively.

## II. SYSTEM MODEL AND NOTATIONS

Consider a multi-cell wireless network, where an MNO rents its infrastructure and spectrum to some MVNOs<sup>1</sup> providing specific service to end-users. Actually, MNO slices its resources in an E2E manner, then rents each E2E slice, including RAN and Core resources, to an MVNO. In more detail, in RAN, each user of each MVNO transmits his data to his serving base station (BS). To do so, the spectrum frequency and BSs owned by MNO are shared between users of MVNOs. The BSs transmit the received data to the core through a backhaul link. At the core, some network functions such as packet inspection, packet routing and forwarding, etc., are performed on each packet of users' data. To this end, at the core designed based on NFV, resources of data centers, including commodity servers and physical links, are shared between users of different MVNOs.

In RAN, BSs and spectrum frequency are shared among users of different slices. We assume that a set of  $\mathcal{B} = \{1, 2, \dots, B\}$  BSs serve  $G$  slices each of them provides service to  $U_g$  end-users, where the total number of users is equal to  $U = \sum_{i \in \mathcal{G}} U_g$  (the set of users is denoted by  $\mathcal{U} = \{1, 2, \dots, U\}$ ). Besides, the whole spectrum frequency of WHz is equally divided into a set of  $\mathcal{C} = \{1, 2, \dots, C\}$  sub-channels.

If  $p_i^k$  denotes the transmit power of user  $i$  on sub-channel  $k$  and  $h_{i,m}^k$  represents the path-gain from user  $i$  to BS  $m$ , the received SINR of user  $i$  on sub-channel  $k$  is given by  $\gamma_i^k = \frac{p_i^k h_{i,b_i}^k}{\sum_{s \notin \mathcal{U}_{b_i}} p_s^k h_{s,b_i}^k + \sigma_{b_i}^2}$ , in which  $b_i$  denotes the BS associated with user  $i$ ,  $\mathcal{U}_{b_i}$  represents the users served by BS  $b_i$ , and  $\sigma_{b_i}^2$  is the noise power at BS  $b_i$ .

Let  $w^k = W/C$  be bandwidth of sub-channel  $k$ , the achieved data rate of user  $i$  based on Shannon's capacity formula is obtained by

$$R_i^k = w^k \log_2 (1 + \gamma_i^k) \quad (1)$$

After the user has successfully sent his data to his serving BS, users' data are sent to the core for performing necessary network functions. For core network, there is a data center modeled by a directed graph as  $Graph = (\mathcal{N}, \mathcal{L})$ , where  $\mathcal{N}$  and  $\mathcal{L}$  respectively denote the set of commodity servers and physical links. For practicality, we assume servers and physical links have limited capacity denoted by  $C_n^{\max}$  and  $B_l^{\max}$ , respectively.

Corresponding to the requested service, each user has an SFC represented by  $\mathcal{S}_i = \{1, 2, \dots, J_i\}$  in which  $\mathcal{S}_i[j]$  denotes  $j$ th VNF of user  $i$ 's SFC. We define binary variable  $x_{i,j}^n$  for representing the allocation of server  $n$  to  $j$ th VNF of SFC  $\mathcal{S}_i$ . Likewise, binary variable  $y_{l,j+1}^i$  denotes the allocation of physical link  $l$  to connect servers that performing  $j$ th and  $j+1$ th VNFs of SFC  $\mathcal{S}_i$ .

### A. Calculation of End-to-End Delay and Energy Consumption

In this paper, we aim at minimizing E2E energy consumption. Similar to [13], we define E2E energy consumption as product of delay and power. So, to calculate E2E energy consumption, we need to define E2E delay. E2E delay in cellular networks consists of delays in RAN, backhaul link, core, and transport networks [14]. To obtain E2E delay, we calculate one-way delay as  $T_i = T_i^{\text{RAN}} + T_i^{\text{bh}} + T_i^{\text{CN}} + T_i^{\text{TN}}$ , where  $T_i^{\text{RAN}}$ ,  $T_i^{\text{bh}}$ ,  $T_i^{\text{CN}}$ , and  $T_i^{\text{TN}}$  denote delay in RAN, backhaul link, core, and transport networks [14], respectively. It is worth noting that E2E delay is approximately obtained by  $2 \times T_i$  [14]. In what follows, we explain each component of  $T_i$  in more detail.

Delay in RAN,  $T_i^{\text{RAN}}$ , is consisting of propagation delay, transmission delay, processing delay in users' device and BS, and queuing delay [14]. In this paper, for simplicity, we consider propagation delay, processing delay in users' device and BS, and queuing delay as a constant value denoted by  $\tau_i$ . The transmission delay is obtained by  $\frac{D_i}{\sum_{k \in \mathcal{C}} R_i^k}$ , where  $D_i$  is size of each packet in bits. Hence, delay in RAN is calculated by  $T_i^{\text{RAN}} = \tau_i + \frac{D_i}{\sum_{k \in \mathcal{C}} R_i^k}$ .

We assume the backhaul link capacity share among users based on the statistical multiplexing [15]. Using statistical multiplexing, users transmit data simultaneously over a link [15]. Accordingly, the delay for transmitting a packet over backhaul link is obtained by  $T_i^{\text{bh}} = \frac{D_i}{C_{\text{bh}}^{\max}}$ , where  $C_{\text{bh}}^{\max}$  denotes the maximum capacity of backhaul link.

In core,  $T_i^{\text{CN}}$  includes two components, namely processing delay and transmitting delay over physical links. The processing delay is defined as delay to process users' data packets on selected servers and transmitting delay is delay for transmitting each packet between servers that perform necessary VNFs on the corresponding packet. Assume  $C_i$  is CPU cycles required by each bit of a packet, the processing delay to process a data packet of user  $i$  is obtained by  $\sum_{n \in \mathcal{N}} \sum_{j \in \mathcal{S}_i} x_n^{i,j} C_i D_i (1/C_n^{\max})$ . Note that  $1/C_n^{\max}$  is delay for processing one bit of each packet on server  $n$ . Similar to backhaul delay, we assume the bandwidth of each link  $l$  is shared among users according to statistical multiplexing. So, delay for transmitting a packet on physical links is calculated by  $\sum_{l \in \mathcal{L}} \sum_{j \in \mathcal{S}_i \setminus \{J_i\}} y_{l,j+1}^i D_i (1/B_l^{\max})$ , where  $1/B_l^{\max}$  is transmission delay to transmit one bit over physical link  $l$ .

<sup>1</sup>Hereafter, MVNO and slice are used interchangeably.

Delay on transport networks is defined as the delay for transmitting a packet from core network to the external networks such as Internet. In recent years, some architecture based on NFV and software defined networking is proposed to reduce transport delay [14]. Investigation of the external data networks is out of the scope of this paper, so we consider a constant value as transport delay.

Based on the above discussion, the one-way delay for each data packet of user  $i$  is expressed as (2).

$$\begin{aligned} T_i = \tau_i + \frac{D_i}{\sum_{k \in \mathcal{C}} R_i^k} + \frac{D_i}{C_{\text{bh}}^{\max}} + \sum_{n \in \mathcal{N}} \sum_{j \in \mathcal{S}_i} x_n^{i,j} C_i D_i (1/C_n^{\max}) \\ + \sum_{l \in \mathcal{L}} \sum_{j \in \mathcal{S}_i \setminus \{J_i\}} y_{l_j}^{i,j+1} D_i (1/B_l^{\max}) + T_i^{\text{TN}}. \end{aligned} \quad (2)$$

As aforementioned, E2E energy consumption is obtained by multiplying each component of E2E delay and its corresponding consumed power. Similar to E2E delay, E2E energy consumption consists of energy consumption in RAN, backhaul link, core, and transport networks as  $E_i = E_i^{\text{RAN}} + E_i^{\text{bh}} + E_i^{\text{CN}} + E_i^{\text{TN}}$ . Considering  $p_i^k$  as transmit power of user  $i$  for transmitting data to BS and constant energy consumption for other components denoted by  $E_i^{\text{cons}}$ , energy consumption in RAN is obtained by  $E_i^{\text{RAN}} = \sum_{k \in \mathcal{C}} p_i^k \frac{D_i}{\sum_{k \in \mathcal{C}} R_i^k} + E_i^{\text{cons}}$ .

Let  $p_{b_i}^{\text{bh}}$  denote the power consumption of BS  $b_i$  for transmitting a data packet of user  $i$  to core network through backhaul link. The energy consumption at backhaul is given by  $E_i^{\text{bh}} = p_{b_i}^{\text{bh}} \frac{D_i}{C_{\text{bh}}^{\max}}$ .

If  $\tilde{p}_n$  represents the power consumption of server  $n$  in data center to process every CPU cycle, the energy consumption in core is obtained by  $E_i^{\text{CN}} = \sum_{n \in \mathcal{N}} \sum_{j \in \mathcal{S}_i} \tilde{p}_n x_n^{i,j} C_i D_i (1/C_n^{\max})$ .

Note that similar to [12], it is assumed that there is no energy consumption on physical links in data centers. Finally, we define constant value of  $E_i^{\text{TN}}$  as energy consumption of transport networks.

Accordingly, E2E energy consumption to transmit each packet of user  $i$  is obtained by  $E_i = E_i^{\text{cons}} + \sum_{k \in \mathcal{C}} p_i^k \frac{D_i}{\sum_{k \in \mathcal{C}} R_i^k} +$

$p_{b_i}^{\text{bh}} \frac{D_i}{C_{\text{bh}}^{\max}} + \sum_{n \in \mathcal{N}} \sum_{j \in \mathcal{S}_i} \tilde{p}_n x_n^{i,j} C_i D_i (1/C_n^{\max}) + E_i^{\text{TN}}$ . It should be noted that since terms  $E_i^{\text{cons}}$ ,  $E_i^{\text{bh}}$ , and  $E_i^{\text{TN}}$  are not affected by decision variable i.e.,  $p_i^k$ ,  $x_n^{i,j}$ , and  $y_{l_j}^{i,j+1}$ , we exclude these terms from user  $i$ 's E2E energy consumption. Hence, hereafter, E2E energy consumption of user  $i$  is expressed as  $E_i = \sum_{k \in \mathcal{C}} p_i^k \frac{D_i}{\sum_{k \in \mathcal{C}} R_i^k} + \sum_{n \in \mathcal{N}} \sum_{j \in \mathcal{S}_i} \tilde{p}_n x_n^{i,j} C_i D_i (1/C_n^{\max})$ .

### III. PROBLEM STATEMENT

Given a sub-channel allocation, the problem of minimizing E2E energy consumption subject to users' E2E delay con-

straint is formally stated as

$$\begin{aligned} & \min_{\{p_i^k, x_n^{i,j}, y_{l_j}^{i,j+1}\}_{i \in \mathcal{U}}} \sum_{i \in \mathcal{U}} \left( \sum_{k \in \mathcal{C}} p_i^k \frac{D_i}{\sum_{k \in \mathcal{C}} R_i^k} + \sum_{n \in \mathcal{N}} \sum_{j \in \mathcal{S}_i} \tilde{p}_n x_n^{i,j} C_i D_i \left( \frac{1}{C_n^{\max}} \right) \right) \\ & \text{s.t.} \\ & \text{C1 : } \sum_{k \in \mathcal{C}} p_i^k \leq p_i^{\max}, \quad \forall i \in \mathcal{U}, \\ & \text{C2 : } \sum_{i \in \mathcal{U}} \sum_{k \in \mathcal{C}} R_i^k \leq C_{\text{bh}}^{\max}, \\ & \text{C3 : } \sum_{n \in \mathcal{N}} x_n^{i,j} = 1, \quad \forall i \in \mathcal{U}, \quad \forall j \in \mathcal{S}_i, \\ & \text{C4 : } \sum_{j \in \mathcal{S}_i} x_n^{i,j} \leq 1, \quad \forall i \in \mathcal{U}, \quad \forall n \in \mathcal{N}, \\ & \text{C5 : } \sum_{i \in \mathcal{U}} \sum_{j \in \mathcal{S}_i} x_n^{i,j} C_i \sum_{k \in \mathcal{C}} R_i^k \leq C_n^{\max}, \quad \forall n \in \mathcal{N}, \\ & \text{C6 : } \sum_{i \in \mathcal{U}} \sum_{j \in \mathcal{S}_i \setminus \{J_i\}} y_{l_j}^{i,j+1} \sum_{k \in \mathcal{C}} R_i^k \leq B_l^{\max}, \quad \forall l \in \mathcal{L}, \\ & \text{C7 : } \sum_{l \in \mathcal{L}_n^{\text{out}}} y_{l_j}^{i,j+1} - \sum_{l \in \mathcal{L}_n^{\text{in}}} y_{l_j}^{i,j+1} = x_n^{i,j} - x_n^{i,j+1}, \\ & \quad \forall i \in \mathcal{U}, \forall n \in \mathcal{N}, \forall j \in \mathcal{S}_i \setminus \{J_i\}, \\ & \text{C8 : } T_i \leq T_i^{\text{th}}, \quad \forall i \in \mathcal{U}, \\ & \text{C9 : } p_i^k \geq 0, \quad \forall i \in \mathcal{U}, \forall k \in \mathcal{C}, \\ & \text{C10 : } x_n^{i,j} \in \{0, 1\}, \quad \forall i \in \mathcal{U}, \quad \forall j \in \mathcal{S}_i \setminus \{J_i\}, \quad \forall n \in \mathcal{N}, \\ & \text{C11 : } y_{l_j}^{i,j+1} \in \{0, 1\}, \quad \forall l \in \mathcal{L}, \quad \forall i \in \mathcal{U}, \quad \forall j \in \mathcal{S}_i \setminus \{J_i\}. \end{aligned} \quad (3)$$

In problem (3), constraint C1 implies that the total transmit power of user  $i$  is limited to the maximum power budget  $p_i^{\max}$ . C2 indicates that the total rate of users does not exceed the maximum capacity of backhaul link. C3 imposes that only one server should be allocated to each VNF  $j \in \mathcal{S}_i$  for each user  $i$ . C4 represents that every VNF for each user  $i$  should be mapped to a different server. Constraints C5 and C6 demonstrate the capacity limitation of commodity servers and bandwidth limitation of physical links in data center, respectively. If  $\mathcal{L}_n^{\text{out}}$  and  $\mathcal{L}_n^{\text{in}}$  denote the outgoing links from server  $n$  and incoming links to server  $n$ , respectively, constraint C7 enforces flow conservation, i.e., the sum of all incoming and outgoing traffic in servers that do not host VNFs should be zero. C8 implies that E2E delay of user  $i$  should not be larger than maximum tolerable delay denoted by  $T_i^{\text{th}}$ . C9 represents the non-negativity of transmit power on each sub-channel. Finally, C10 and C11 indicate the binary nature of server and physical link allocation, respectively.

### IV. OUR PROPOSED JPSLA ALGORITHM

Due to (i) non-concavity of data rate function (1), (ii) existence of continuous and integer variables, and (iii) multiplying of transmit power and transmission delay (i.e.,  $\sum_{k \in \mathcal{C}} p_i^k \frac{D_i}{\sum_{k \in \mathcal{C}} R_i^k}$ ) in objective function, the optimal solution of problem (3)

cannot be obtained in polynomial time. So, in what follows, we propose a sub-optimal joint power control, server and link allocation (JPSLA) to solve problem (3).

To tackle with existence of continuous and integer variables, we decompose problem (3) into two sub-problems, namely power control problem (4) in RAN and server and link allocation problem (5) in core as

$$\begin{aligned} \min_{\{p_i^k\}} \quad & \sum_{i \in \mathcal{U}} \sum_{k \in \mathcal{C}} p_i^k \frac{D_i}{\sum_{k \in \mathcal{C}} R_i^k} \\ \text{s.t.} \quad & \text{C1, C2, C5, C6, C8, C9,} \end{aligned} \quad (4)$$

and

$$\begin{aligned} \min_{\{x_n^{i,j}, y_{l_j^{j+1}}\}} \quad & \sum_{i \in \mathcal{U}} \sum_{n \in \mathcal{N}} \sum_{j \in \mathcal{S}_i} \tilde{p}_n x_n^{i,j} C_i D_i (1/C_n^{\max}) \\ \text{s.t.} \quad & \text{C3, C4, C5, C6, C7, C8, C10, C11.} \end{aligned} \quad (5)$$

#### A. Solving Power Control Problem (4)

Power control problem (4) is a non-convex problem due to (i) non-convexity of constraints C2, C5, and C6, (ii) multiplying of transmit power and transmission delay (i.e.,  $\sum_{k \in \mathcal{C}} p_i^k \frac{D_i}{\sum_{k \in \mathcal{C}} R_i^k}$ ) in objective function, and (iii) non-concavity of data rate function (1).

To tackle with non-convexity of constraints C2, C5, and C6, similar to [17], we define  $\nu_i^k$  as an auxiliary variable and substitute  $R_i^k$  in problem (4) by  $\nu_i^k$  and add following constraint to problem (4)

$$\text{C12 : } \nu_i^k \leq R_i^k. \quad (6)$$

While C2, C5, and C6 become convex, problem (4) is still non-convex because of non-convexity of the objective function and non-concavity of data rate in constraint C12. As aforementioned, the objective function of problem (4) is non-convex due to the existence of  $\sum_{k \in \mathcal{C}} p_i^k \frac{D_i}{\sum_{k \in \mathcal{C}} R_i^k}$ . Since the energy

consumption derivative is positive with respect to the user's transmit power, implying that energy consumption is decreased by decreasing the transmit power, the minimum amount of energy consumption can be obtained by minimizing users' total transmit power. Moreover, constraint C9 in problem (4) indicates that user  $i$ 's transmission delay to transmit a packet to BS is upper bounded by  $T_i^{\text{th}} - (\tau_i + T_i^{\text{bh}} + T_i^{\text{CN}} + T_i^{\text{TN}})$  i.e.,  $\frac{D_i}{\sum_{k \in \mathcal{C}} R_i^k} \leq T_i^{\text{th}} - (\tau_i + T_i^{\text{bh}} + T_i^{\text{CN}} + T_i^{\text{TN}})$ . Accordingly, user  $i$ 's energy consumption for sending a packet to BS is approximated as  $\sum_{k \in \mathcal{C}} p_i^k (T_i^{\text{th}} - (\tau_i + T_i^{\text{bh}} + T_i^{\text{CN}} + T_i^{\text{TN}}))$ .

Finally, to make data rate function (1) a concave one, we consider a predefined interference threshold  $I_i^{k,\text{th}}$  instead of interference terms which make data rate function (1) non-concave. To this end, we add the following constraint to problem (4)

$$\text{C13 : } \sum_{s \notin \mathcal{U}^{b_i}} p_s^k h_{s,b_i}^k \leq I_i^{k,\text{th}}, \quad \forall i \in \mathcal{U}, \quad \forall k \in \mathcal{C}. \quad (7)$$

By doing so, data rate of user  $i$  on sub-channel  $k$  is obtained by  $\tilde{R}_i^k = w^k \log_2 \left( 1 + \frac{p_i^k h_{i,b_i}^k}{I_i^{k,\text{th}} + \sigma_{b_i}^2} \right)$ . Employing the above discussed steps, problem (4) is converted to

$$\begin{aligned} \min_{\{p_i^k\}} \quad & \sum_{i \in \mathcal{U}} \sum_{k \in \mathcal{C}} p_i^k \left( T_i^{\text{th}} - (\tau_i + T_i^{\text{bh}} + T_i^{\text{CN}} + T_i^{\text{TN}}) \right) \\ \text{s.t.} \quad & \text{C1 : } \sum_{k \in \mathcal{C}} p_i^k \leq p_i^{\max}, \quad \forall i \in \mathcal{U}, \\ & \text{C2 : } \sum_{i \in \mathcal{U}} \sum_{k \in \mathcal{C}} \nu_i^k \leq C_{\text{bh}}^{\max}, \\ & \text{C5 : } \sum_{i \in \mathcal{U}} \sum_{j \in \mathcal{S}_i} x_n^{i,j} C_i \sum_{k \in \mathcal{C}} \nu_i^k \leq C_n^{\max}, \quad \forall n \in \mathcal{N}, \\ & \text{C6 : } \sum_{i \in \mathcal{U}} \sum_{j \in \mathcal{S} \setminus \{J_i\}} y_{l_j^{j+1}}^i \sum_{k \in \mathcal{C}} \nu_i^k \leq B_l^{\max}, \quad \forall l \in \mathcal{L}, \\ & \text{C8 : } \frac{D_i}{\sum_{k \in \mathcal{C}} \nu_i^k} + \tau_i + T_i^{\text{bh}} + T_i^{\text{CN}} + T_i^{\text{TN}} \leq T_i^{\text{th}}, \quad \forall i \in \mathcal{U}, \\ & \text{C9, C12, C13.} \end{aligned} \quad (8)$$

Problem (8) is a convex, so it can be optimally solved by CVX [18].

#### B. Solving Sever and Link Allocation Problem (5)

Problem (5) is an integer linear programming problem. To handle integer variables  $x_n^{i,j}$  and  $y_{l_j^{j+1}}^i$ , we replace C10 and C11 in problem (5) by following constraints

$$\begin{aligned} \text{C10.1 : } & \sum_{n \in \mathcal{N}} \sum_{i \in \mathcal{U}} \sum_{j \in \mathcal{S}_i} (x_n^{i,j} - x_n^{i,j})^2 \leq 0, \\ \text{C10.2 : } & 0 \leq x_n^{i,j} \leq 1, \quad \forall n \in \mathcal{N}, \forall i \in \mathcal{U}, \forall j \in \mathcal{S}_i. \end{aligned} \quad (9)$$

and

$$\begin{aligned} \text{C11.1 : } & \sum_{l \in \mathcal{L}} \sum_{i \in \mathcal{U}} \sum_{j \in \mathcal{S}_i} (y_{l_j^{j+1}}^i - y_{l_j^{j+1}}^i)^2 \leq 0, \\ \text{C11.2 : } & 0 \leq y_{l_j^{j+1}}^i \leq 1, \quad \forall l \in \mathcal{L}, \forall i \in \mathcal{U}, \forall j \in \mathcal{S}_i. \end{aligned} \quad (10)$$

Adding constraints C10.2 and C11.2 to problem (5) which relax binary variables  $x_n^{i,j}$  and  $y_{l_j^{j+1}}^i$  to continuous ones may make problem (5) infeasible. So, for enforcing  $x_n^{i,j}$  and  $y_{l_j^{j+1}}^i$  to binary variables, we add constraints C10.1 and C11.1 to objective function as a penalty function. By doing so, problem (5) is transformed into

$$\begin{aligned} \min_{\{x_n^{i,j}, y_{l_j^{j+1}}^i\}} \quad & \sum_{i \in \mathcal{U}} \sum_{n \in \mathcal{N}} \sum_{j \in \mathcal{S}_i} \tilde{p}_n x_n^{i,j} C_i D_i (1/C_n^{\max}) \\ & + \zeta_1 \sum_{n \in \mathcal{N}} \sum_{i \in \mathcal{U}} \sum_{j \in \mathcal{S}_i} (x_n^{i,j} - x_n^{i,j})^2 \\ & + \zeta_2 \sum_{l \in \mathcal{L}} \sum_{i \in \mathcal{U}} \sum_{j \in \mathcal{S}_i} (y_{l_j^{j+1}}^i - y_{l_j^{j+1}}^i)^2 \\ \text{s.t.} \quad & \text{C3, C4, C5, C6, C7, C8, C10.2, C11.2,} \end{aligned} \quad (11)$$

where  $\zeta_1$  and  $\zeta_2$  are penalty factors for enforcing  $x_n^{i,j}$  and  $y_{l_j^{j+1}}^i$  to 0 or 1.

The existence of  $x_n^{i,j}^2$  and  $y_{l_j^{j+1}}^i$  in the objective function make problem (11) non-convex. To deal with this difficulty, we approximate  $x_n^{i,j}^2$  and  $y_{l_j^{j+1}}^i$  by first-order Taylor approximation approach [19]. First-order Taylor approximation approximates a function  $f(x)$  by a linear function as

$$f(x) \approx f(\bar{x}) + \nabla_x f(\bar{x})(x - \bar{x}), \quad (12)$$

where,  $\bar{x}$  is a feasible initial point.

Employing (12) to approximate  $x_n^{i,j}^2$  and  $y_{l_j^{j+1}}^i$ , problem (11) is rewritten as

$$\begin{aligned} & \min_{\{x_n^{i,j}, y_{l_j^{j+1}}^i\}_{i \in \mathcal{U}}} \sum_{i \in \mathcal{U}} E_i \\ & + \zeta_1 \sum_{n \in \mathcal{N}} \sum_{i \in \mathcal{U}} \sum_{j \in \mathcal{S}_i} (x_n^{i,j} + [2x_n^{i,j}x_n^{i,j}(t-1) - (x_n^{i,j}(t-1))^2]) \\ & + \zeta_2 \sum_{l \in \mathcal{L}} \sum_{i \in \mathcal{U}} \sum_{j \in \mathcal{S}_i} (y_{l_j^{j+1}}^i + [2y_{l_j^{j+1}}^i y_{l_j^{j+1}}^i(t-1) - (y_{l_j^{j+1}}^i(t-1))^2]) \\ \text{s.t. } & C3, C4, C5, C6, C7, C8, C10.2, C11.2, \end{aligned} \quad (13)$$

Problem (13) is a linear problem that can be optimally solved by CVX [18]. Note that  $x_n^{i,j}(t-1)$  and  $y_{l_j^{j+1}}^i(t-1)$  denote the optimal solution of problem (13) at the previous iteration  $t-1$ .

#### Algorithm 1: JPSLA algorithm to solve problem (3)

- 1 Initialize the maximum number of iterations  $t^{\max}$ ,  $\zeta_1, \zeta_2 \gg 1$ , iteration index  $t = 1$ , and a feasible initial point  $x_n^{i,j}(0)$ , and  $y_{l_j^{j+1}}^i(0)$ .
- 2 **Repeat**
- 3     **Step 1: Solving power control problem (8)**  
4         Solve (8) by CVX to obtain  $p_i^k(t)$ ,  $\forall i \in \mathcal{U}$ ,  $\forall k \in \mathcal{C}$ .
- 5     **Step 2: Solving server and link allocation problem (13)**  
6         Solve (13) by CVX to obtain  $x_n^{i,j}(t)$ , and  $y_{l_j^{j+1}}^i(t)$ .
- 7     Set  $t \leftarrow t + 1$ .
- 8 **Until** convergence or  $t = t^{\max}$ .

For obtaining a local optimum for problem (3), we propose JPSLA algorithm as an iterative algorithm summarized in Algorithm 1. At each iteration of JPSLA, convex problems (8) and (13) are solved by CVX until convergence is occurred.

## V. SIMULATION RESULTS

In this section, we evaluate the performance of our proposed JPSLA algorithm. To do so, a multi-cell network consisting of  $B = 2$  BSs with a coverage area of  $500m \times 500m$  is considered. We assume in this network,  $G = 3$  slices, for instance, factory automation, health care, and intelligent transport systems are provided to end-users. Moreover, end-users of different slices are randomly distributed at each cell. Like [20], the path gain between each user and BSs is modeled by  $h_{i,m}^k = \mu^k d_{i,m}^{-\beta}$ , where  $d_{i,m}$  is the distance between user  $i$  and BS  $m$ ,  $\mu^k$  is a random value generated by the Rayleigh distribution, and  $\beta = 3$  is the path loss exponent. Additionally,

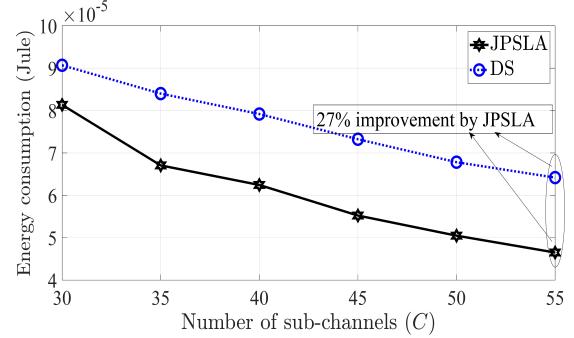


Fig. 1. Comparison of JPSLA and disjoint scheme vs. sub-channels number

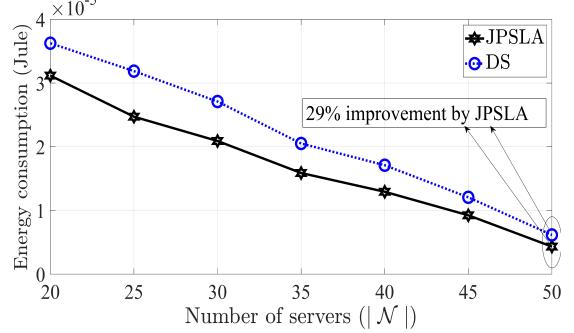


Fig. 2. Comparison of JPSLA and disjoint scheme vs. servers number

for the core network, it is assumed that commodity servers are connected to each other with randomly established physical links. Randomly setting locations and path-gain of users in each snapshot, each curve in what follows is obtained by averaging from 50 independent snapshots.

To obtain simulation figures, bandwidth of each sub-channel is set to  $w^k = 15$  KHz. Furthermore, each user  $i$  should not transmit larger than 100 mW (i.e.,  $p_i^{\max} = 100$  mW). Also, noise power at BS  $m$  is set to  $\sigma_m^2 = 10^{-14}$  W. Each user  $i$  has packets with size  $D_i = 100$  bits. Moreover, the capacity of backhaul link is limited to  $C_{bh}^{\max} = 1$  Gbps. Besides, the maximum capacity of commodity servers and physical links are randomly selected from [10, 50]MHz and [100, 500]Mbps, respectively. In addition, we set constant delay in RAN and transport network delay to  $\tau_i = 2$  ms and  $T_i^{\text{TN}} = 1$  ms, respectively. Finally, power consumption of each server  $n$  i.e.,  $\tilde{p}_n$  is randomly selected from [1, 10] W.

To show the performance of our proposed JPSLA algorithm, we compare it with the disjoint (DS) scheme, where the power control problem in RAN and server and link allocation in the core are solved disjointly. As stated in [3], half of the tolerable E2E delay for each user is related to RAN and the other half is related to the core. Firstly, in the DS scheme, the power control problem (4) is solved considering half of the maximum tolerable delay in RAN. Then, given transmit power of users, server and link allocation problem (5) is solved considering half of the maximum tolerable delay in the core.

In Fig. 1, energy consumption in JPSLA algorithm is compared with the DS scheme, where the number of sub-

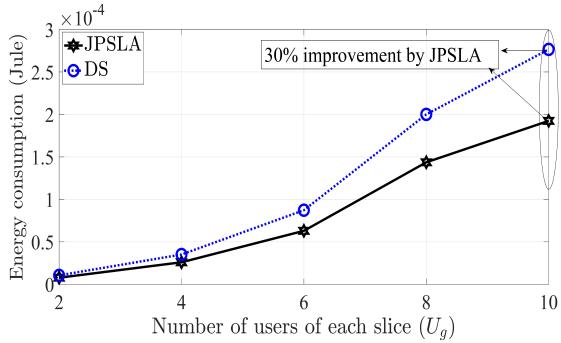


Fig. 3. Comparison of JPSLA and disjoint scheme vs. users' number

channels increases from 30 to 55. For generating this figure, we set number of servers to  $|\mathcal{N}| = 20$ . Also, a different maximum tolerable delay is considered for different slices. Specifically, the maximum tolerable delay of users in slice 1, 2, and 3 are respectively set to  $T_i^{\text{th}} = 5\text{ms}$ ,  $T_i^{\text{th}} = 10\text{ms}$ , and  $T_i^{\text{th}} = 15\text{ms}$ . Also, it is assumed that each slice provides service to  $U_g = 4$  users equally distributed in cell. It can be seen that because of joint consideration of power control, server and link allocation, as well as considering half of the tolerable delay in RAN and core in DS scheme, energy consumption in JPSLA is less than that of DS scheme. Besides, it can be observed that due to sub-channel diversity, increasing the number of sub-channel, energy consumption is decreased.

Fig. 2 illustrates the energy consumption in JPSLA algorithm in comparison with the DS scheme. To generate this figure, we set the number of sub-channels to  $C = 30$ , and the other parameters are set the same as Fig. 1. From Fig. 2, we can observe that with increasing the number of servers, the energy consumption is decreased. The reason is that commodity servers in the data centers are heterogeneous in terms of maximum capacity; also, since the maximum capacity of servers is set randomly, the probability of the existence of high-capacity servers is increased. Furthermore, the energy consumption in JPSLA is lower than that of the DS scheme because of the joint allocation of decision variables. Besides, considering half of the maximum tolerable delay in both RAN and core, the E2E delay constraint in the DS scheme is more strict than JPSLA.

In Fig. 3, we increase the number of users served by each slice from  $U_g = 2$  to  $U_g = 10$ . To generate this figure, we assume that  $C = 50$  sub-channels are equally allocated to users at each cell. Fig. 3 demonstrates that the energy consumption in both JPSLA and DS scheme is increased when the number of users is increased. Obviously, with increasing the number of users, since more users cause interference on each other, users should transmit with higher transmit power resulting in more energy consumption.

## VI. CONCLUSION

In this paper, we studied E2E slicing so that RAN and core resources are jointly allocated to end-users of MVNOs. Due to the importance of energy consumption in wireless networks and its environmental impacts, we aimed at minimizing E2E

energy consumption. To address the stated problem, which was non-convex, we proposed JPSLA algorithm. Via simulation results, we illustrated that JPSLA improves energy consumption by 30% in comparison with the disjoint scheme, where RAN and core are sliced separately.

## REFERENCES

- [1] V. N. Ha and L. B. Le, "End-to-End Network Slicing in Virtualized OFDMA-Based Cloud Radio Access Networks," *IEEE Access*, vol. 5, pp. 18675-18691, 2017.
- [2] I. Afolabi, T. Taleb, K. Samdanis, A. Ksentini, and H. Flinck, "Network Slicing and Softwarization: A Survey on Principles, Enabling Technologies, and Solutions," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 3, pp. 2429-2453, 2018.
- [3] White Paper, "Cellular IoT in the 5G era," Ericsson, 2020.
- [4] K. Zhu and E. Hossain, "Virtualization of 5G Cellular Networks as a Hierarchical Combinatorial Auction," *IEEE Transactions on Mobile Computing*, vol. 15, no. 10, pp. 2640-2654, 2016.
- [5] J. Gil Herrera and J. F. Botero, "Resource Allocation in NFV: A Comprehensive Survey," *IEEE Transactions on Network and Service Management*, vol. 13, no. 3, pp. 518-532, 2016.
- [6] Q. Shi, L. Zhao, Y. Zhang, G. Zheng, F. R. Yu, and H. Chen, "Energy-Efficiency Versus Delay Tradeoff in Wireless Networks Virtualization," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 1, pp. 837-841, 2018.
- [7] M. Alsenwi, N. H. Tran, M. Bennis, S. R. Pandey, A. K. Bairagi, and C. S. Hong, "Intelligent Resource Slicing for eMBB and URLLC Coexistence in 5G and Beyond: A Deep Reinforcement Learning Based Approach," Available online at: <https://arxiv.org/pdf/2003.07651.pdf>, 2020.
- [8] P. Korrai, E. Lagunas, S. K. Sharma, S. Chatzinotas, A. Bandi, and B. Ottersten, "A RAN Resource Slicing Mechanism for Multiplexing of eMBB and URLLC Services in OFDMA Based 5G Wireless Networks," *IEEE Access*, vol. 8, pp. 45674-45688, 2020.
- [9] V. Eramo and F. G. Lavacca, "Computing and Bandwidth Resource Allocation in Multi-Provider NFV Environment," *IEEE Communications Letters*, vol. 22, no. 10, pp. 2060-2063, 2018.
- [10] G. Sun, Z. Chen, H. Yu, X. Du, and M. Guizani, "Online Parallelized Service Function Chain Orchestration in Data Center Networks," *IEEE Access*, vol. 7, pp. 100147-100161, 2019.
- [11] M. Savi, M. Tornatore, and G. Verticale, "Impact of Processing-Resource Sharing on the Placement of Chained Virtual Network Functions," *IEEE Transactions on Cloud Computing*, 2019, early access.
- [12] S. Kazemi Taskou, M. Rasti, and P. H. J. Nardelli, "Energy and Cost Efficient Resource Allocation for Blockchain-Enabled NFV," *IEEE Transactions on Services Computing*, early access, 2021.
- [13] M. Chen and Y. Hao, "Task Offloading for Mobile Edge Computing in Software Defined Ultra-Dense Network," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 3, pp. 587-597, 2018.
- [14] I. Parvez, A. Rahmati, I. Guvenc, A. I. Sarwat, and H. Dai, "A Survey on Low Latency Towards 5G: RAN, Core Network and Caching Solutions," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 4, pp. 3098-3130, 2018.
- [15] D. P. Bertsekas, R. G. Gallager, and P. Humblet, "Data networks," *Prentice-Hall International New Jersey*, 1992, vol. 2.
- [16] H. Alameddine, M. H. K. Tushar, and C. Assi, "Scheduling of Low Latency Services in Softwarized Networks," *IEEE Transactions on Cloud Computing*, no. pp, vol. , 2019.
- [17] P. Luong, F. Gagnon, C. Despins, and L. 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, April 2018.
- [18] M. Grant and S. Boyd, "CVX: Matlab Software for Disciplined Convex Programming, version 2.1," [Online] <http://cvxr.com/cvx>, Mar. 2014.
- [19] Y. Sun, P. Babu and D. P. Palomar, "Majorization-Minimization Algorithms in Signal Processing, Communications, and Machine Learning," *IEEE Transactions on Signal Processing*, vol. 65, no. 3, pp. 794-816, 1 Feb. 2017.
- [20] S. Kazemi and M. Rasti, "Joint power control and sub-channel allocation for co-channel OFDMA femtocells," in *Proceeding of 2016 IEEE Symposium on Computers and Communication (ISCC)*, Messina, 2016, pp. 1171-1176.