End-to-End Resource Slicing for Coexistence of eMBB and URLLC Services in 5G-Advanced/6G Networks

Shiva Kazemi Taskou, Mehdi Rasti, *Senior Member, IEEE*, and Ekram Hossain, *Fellow, IEEE*

***Abstract*—We study the problem of end-to-end (E2E) network slicing, i.e., joint slicing of the radio access network (RAN) and core network (CN), for the coexistence of enhanced mobile broadband (eMBB) and ultra-reliable and low latency communication (URLLC) services in future generation cellular (e.g., 5G-Advanced/6G) networks. The E2E resource slicing problem is defined as a mixed-integer non-linear programming problem to minimize the E2E energy consumption and the cost of utilized resources. To overcome the difficulty of solving this problem, we decompose it into two sub-problems, namely, RAN resource allocation (RRA) and CN resource allocation (CRA) problems. In both RRA and CRA problems, the existence of binary variables makes them intractable. To tackle this difficulty, we relax the binary variables by introducing penalty functions. Then, we make the RRA and CRA problems convex by employing the majorization-minimization approximation method. Via simulation results, we compare our proposed joint RAN and CN resource allocation algorithm (JRCRA) with the disjoint solution where RAN and CN resources are allocated to users separately. The joint allocation of resources in the RAN and CN has the advantage that the E2E tolerable latency of users can be flexibly divided between RAN and CN. In contrast, if resources in RAN and CN are allocated separately, a predefined part of the E2E tolerable latency should be considered as the tolerable latency in RAN and CN. The simulation results illustrate that our proposed JRCRA algorithm obtains a 34% improvement in energy consumption and a 24% improvement in cost compared to the disjoint one. Moreover, via simulation results, we illustrate that in comparison with existing algorithms, our proposed JRCRA obtains a higher performance. Besides, simulation results confirm that JRCRA reaches a close performance to the optimal solution.**

***Index Terms*—5G-Advanced/6G, E2E network slicing, network function virtualization, service function chain, ultra-reliable and low latency communication, enhanced mobile broadband**

I. Introduction

The fifth-generation (5G) and beyond wireless networks provide three fundamental services with diverse quality of service (QoS) requirements, which are enhanced mobile broadband (eMBB) focusing on high data rate, massive machine type communication (mMTC) for connecting the huge number of connected devices, and ultra-reliable and low latency communication (URLLC) focusing on reliable and low latency communication [2]. The coexistence of different services with diverse QoS requirements, especially eMBB and URLLC, on the same frequency spectrum and infrastructure, is a major challenge [3]. To support these diverse requirements simultaneously, the wireless networks should be agile, software-based, and demand-oriented [4]. To this end, network slicing has attracted a lot of attention from both industry and academia [5]. By network slicing, the infrastructure is partitioned into several logical slices, and each of them provides a different service with diverse QoS requirements [5].

Due to the end-to-end (E2E) nature of the QoS, network slicing should be done in an E2E manner from the radio access network (RAN) to the core network (CN) [3], [6]. An E2E slice consolidates diverse resources, including radio resources in RAN, the backhaul/fronthaul links' capacity, and processing and networking resources in CN for delivering services to end-users [7]. Because of the requirement of E2E QoS, it is necessary to ensure QoS both in RAN and CN. To guarantee QoS in RAN, resources including frequency spectrum and transmit power in the RAN should be allocated to the users to send their data packets to the base station successfully. And at the CN, the necessary network resources (e.g., computing resources, link capacity) need to be allocated accordingly.

Network function virtualization (NFV) is a promising technology to realize CN slicing. In NFV, network functions are treated as virtual network functions (VNFs) performed on commodity servers provided by mobile edge or cloud computing. A sequence of VNFs connected by virtual links forms service function chains (SFCs) to provide a specific service [8]. To provide each SFC, the processing and networking resources should be allocated to them. To allocate processing resources, the CPU cycles of servers are allocated to VNFs, and to allocate networking resources, virtual links between VNFs are embedded on physical links connecting servers [8].

On the other hand, energy consumption by information and communication technology is a very pressing concern and has many environmental, economic, and performance impacts [9]. Therefore, an efficient resource allocation approach that minimizes energy consumption would be desirable [9]. Furthermore, the mobile virtual network providers (MVNOs) who lease the resources from mobile network operators (MNOs) should pay the cost of utilized resources to MNOs. Therefore, minimization of cost for the utilized-resources is an important goal in the design of resource allocation schemes.

The E2E network slicing for coexisting eMBB and URLLC services, minimizing the energy consumption as well as the

Part of this paper was presented in IEEE WCNC'22 [1].

S. K. Taskou and M. Rasti are with the Department of Computer Engineering, Amirkabir University of Technology, Tehran, Iran (e-mail: shiva.kazemi.t@gmail.com). M. Rasti is also with the Center for Wireless Communications, University of Oulu, Finland (email: mehdi.rasti@oulu.fi). E. Hossain is with the Department of Electrical and Computer Engineering, University of Manitoba, Canada (email: ekram.hossain@umanitoba.ca).

cost of utilized resources are grand challenges for beyond 5G/6G networks. In this context, we study the problem of E2E network slicing to minimize the energy consumption and the cost of utilized resources for eMBB and URLLC services. The maximum E2E tolerable latency is considered to be the QoS requirement for URLLC users. And for eMBB users, two QoS metrics are defined: minimum data rate requirement in RAN and the maximum E2E tolerable latency.

A. Related Work

The existing research work on network slicing in 5G and beyond [10]– [29] can be classified into: i) work that only considered RAN slicing to guarantee QoS in RAN, ii) work that only investigated CN slicing to assure QoS in CN, and iii) work that studied E2E slicing to assure E2E QoS.

Specifically, in [10]- [15], radio resource allocation for RAN slicing was addressed. A single-cell cellular network was considered in [10]- [13]. The authors in [10] proposed a deep reinforcement learning framework to maximize the total data rate of eMBB slices such that the reliability of URLLC users is guaranteed. Also, to reduce the impact of URLLC slices on the eMBB ones, the variance of data rate for the eMBB slices is minimized. Likewise, in [11], a heuristic algorithm was proposed to allocate sub-channels to eMBB and URLLC slices in such a way that the minimum data rate is satisfied for both eMBB and URLLC users. In [12], two different resource allocation problems were defined for eMBB and URLLC slices. For the eMBB slice, the optimization problem aims at maximizing the total data rate, and the optimization problem for the URLLC users maximizes the number of admitted packets and minimizes the loss rate of eMBB users. The authors in [13] proposed an optimization-based algorithm for sub-channel allocation to eMBB users and scheduling URLLC users to maximize the total data rate of eMBB users. The RAN slicing for multi-cell cellular networks was investigated in [14]- [15]. In [14], considering two different slices, one with a data rate requirement and the other with a low-latency requirement, a deep reinforcement learning framework was proposed to maximize the total data rate. Likewise, the authors in [15] studied the coexistence of eMBB and URLLC slices in a multi-cell cellular network. A deep reinforcement learning method was proposed for sub-channel allocation to maximize the total data rate of eMBB and URLLC users.

For CN slicing, the NFV resource allocation problem was studied in [16]- [23]. In particular, in [16], the servers and links are allocated to SFCs to maximize the number of admitted SFCs and minimize utilized resources. Also, heuristic algorithms were proposed in [17]- [18] to minimize the number of utilized servers. In [17] and [18], the maximum tolerable latency in CN was defined as the sum of processing time on servers and transmission time on links. The research works [19]- [21] considered the problem of minimizing the latency in CN. The latency minimization problem in [19] was defined as a mixed-integer linear programming problem, which was solved by a heuristic algorithm. The problem of minimizing the latency and link bandwidth consumption and maximizing server load rate in [20] was defined as a multi-objective problem. The authors in [20] proposed a heuristic algorithm based on the breadth-first-search method to solve this problem. Furthermore, the authors in [21] defined an optimization problem to simultaneously minimize latency and energy consumption. Latency in [21] was defined as the summation of the processing time of VNFs and transmission time on physical links. The authors in [22] proposed an online heuristic algorithm to maximize the number of admitted SFCs satisfying the transmission latency requirement. In [23], an approximation-based algorithm and a heuristic algorithm were proposed to minimize the total energy consumption and cost of utilized resources.

The E2E network slicing problem was studied in [24]- [33]. Specifically, in [24]– [26], the E2E slicing problem was investigated for cloud-RAN networks. In [24], network slicing was considered jointly for RAN, fronthaul link, and cloud (for embedding the baseband units' functionalities). A joint RAN bandwidth and cloud processing power allocation algorithm was proposed to minimize the E2E energy consumption subject to the E2E latency requirement for all users. Likewise, in [25], two different slices, namely, mission-critical and non-critical slices, were considered. A genetic algorithm was proposed to associate remote radio units with baseband units to minimize latency and load balancing. Also, in [26], a reinforcement learning method was proposed to minimize the operational cost of the MNO, where the agent decides about SFC embedding and configuration. Moreover, the E2E slicing, including RAN and CN, was investigated in [27]- [33]. The authors in [27] proposed a framework for jointly RAN, CN, and transport network slicing and studied the impact of this E2E slicing on RAN protocols. No mathematical framework was proposed for the E2E slicing in [27]. In [28], two QoS metrics were considered for all slices, including the data rate requirement in RAN and the E2E latency requirement. The E2E latency was obtained as the summation of RAN and CN latency; however, the CN latency was assumed to be constant. A learning approach was proposed for base station assignment to users in order to maximize the network throughput and minimize the handoff cost. In [29], two different slices were considered: a rate-constrained slice with data rate requirement and a delay-constrained slice with E2E latency requirement. In the RAN, each user is assigned to one base station, and then base stations transmit their associated users' requests to the CN. And in CN, the required VNFs are performed to provide users' SFCs. The objective of the optimization problem was to maximize the number of admitted users, and it was solved by using a deep Q-network approach.

The study in [30] focused on end-to-end slicing, which includes joint slicing of RAN and CN. Specifically, the study examined the impact of joint slicing on minimizing operational costs, where the optimization process involved the allocation of sub-channels in RAN and the selection of paths for users' packets in CN. In the study presented in [31], the authors investigated joint slicing of the RAN and CN. In this approach, base stations in the RAN are represented as virtual base stations embedding on physical base stations, and in the CN, a set of VNFs for each slice are embedded on general-purpose servers. In [32], the concept of E2E slicing is explored within the context of an information-centric network (ICN). Within

this network, the ICN cache enabler and ICN gateway are implemented as virtual functions on general-purpose servers. It is assumed that the ICN cache enabler is deployed on the mobile edge computing servers to ensure proximity to the end-user, while the ICN gateway is deployed within the cloud. The stated optimization problem in [32] involves binary variables that allocate servers for the aforementioned functions and binary variables that determine the physical links required to establish a connection between them. The objective was to minimize the utilization of processing resources in both the mobile edge computing and cloud while simultaneously maximizing service quality. The authors in [33] proposed a deep reinforcement learning (DRL) method to maximize the MNO's revenue by jointly optimizing the resources in RAN and CN. In [33], for each user, an E2E tolerable latency was considered as a QoS requirement. Also, for each slice, a minimum data rate constraint is assured such that the total achieved data rate of each slice should not be less than a minimum requirement. It is worth mentioning that a total data rate requirement was considered for each slice which may violate fairness in RAN among users of each slice.

The differences between our work and the existing works on E2E slicing are summarized in Table I.

B. Contributions

The majority of existing research on network slicing has focused only on slicing in either the RAN [10]- [15] or the CN [16]- [23]. However, given the E2E nature of 5G and beyond networks, it is a necessity to consider E2E slicing, which encompasses slicing in the RAN, backhaul or fronthaul links, and CN, in order to adequately support heterogeneous services for 5G and beyond [7]. In previous generations of wireless networks, the RAN was deemed a bottleneck in ensuring E2E QoS due to technological limitations, such as the maximum user power and frequency bandwidth. E2E QoS refers to the requirement that QoS be guaranteed across all network components, including the RAN and CN. However, in 5G, advancements in RAN technologies have significantly diminished the limitations of the RAN, thereby making both the RAN and CN influential in achieving E2E QoS. Consequently, it is necessary to study RAN and CN technologies jointly. Based on this realization, we can conclude that in order to ensure E2E QoS for users, the allocation of radio resources in the RAN and the processing resources and bandwidth of physical links in the CN should be performed in a collaborative manner. Particularly in applications with low latency requirements, as latency is a crucial aspect of E2E QoS, neither the RAN nor the CN alone can fulfill this demand. According to Ericsson [35] and Nokia [36] reports, half of the tolerable latency is related to the RAN and the other half is related to the CN. Therefore, to guarantee the E2E latency for users, it is essential to jointly slice and allocate resources in the RAN and the CN.

Given the significance of E2E slicing, this paper investigates the effects of E2E slicing on energy consumption and the cost of utilized resources. To the best of our knowledge, this is the first work studying E2E slicing to minimize the E2E energy consumption and the cost of utilized resources considering the E2E latency for both eMBB and URLLC slices. Specifically, in this paper, we answer the question of whether joint slicing of RAN and CN can improve the network performance and by how much. In doing so, simulation results validate that E2E slicing surpasses disjoint slicing (i.e., slicing of RAN and CN independently) in terms of energy consumption and cost of utilized resources, exhibiting an improvement of 34% and 24%, respectively. In light of these findings, it is evident that E2E slicing offers superior performance in terms of energy efficiency and resource utilization cost over disjoint slicing.

More specifically, similar to [10]- [13], we consider the coexistence of eMBB and URLLC slices. Also, similar to [16]- [23], we employ NFV technology for CN slicing. We assume that in the RAN, the MNO's resources, including base stations and frequency spectrum, are shared among slices. Additionally, it is assumed that the CN resources, including servers and links, are provided by a cloud belonging to the MNO and are shared among different slices.

Complementing [10]- [23] which consider either RAN slicing or CN slicing, in this work, we study the E2E slicing with joint slicing of RAN and CN. In contrast to [27], we propose a mathematical model for E2E network slicing. In comparison with [28], we define E2E latency as the total latency in RAN, backhaul link, CN, and data transport networks. In [28], the CN latency was assumed to be constant, while in our work, the CN latency depends on the processing power and physical link allocation in CN. Although in [29] data rate requirement was considered for the rate-constrained slice and E2E latency was decomposed into disjoint RAN and CN latency for the delay-constrained slice, we define E2E latency as an E2E QoS requirement for both eMBB and URLLC slices. Compared to [33], we consider the coexistence of eMBB and URLLC slices, where an E2E latency requirement is assured for all users, and due to the high data rate requirement of eMBB users, a minimum data rate is guaranteed for each eMBB user. The major contributions of this paper are summarized as follows.

- Considering the coexistence of eMBB and URLLC services, we present a system model in which both RAN resources, including sub-channels and power levels, and CN resources, including processing and networking resources, are allocated to users to provide the E2E QoS.
- In this paper, we focus on minimizing the E2E energy consumption and the cost of utilized resources such that a minimum data rate requirement for eMBB users in RAN and the E2E latency requirement for both eMBB and URLLC users are met. To do so, we define the E2E energy consumption by summing up the energy consumption of each network component. This is calculated by multiplying the latency of each component with its corresponding power consumption. The E2E latency is the sum of latency in RAN, the backhaul link, CN, and transport networks. Moreover, the cost of utilized resources is defined as the unit price for RAN resources, including sub-channels, and CN resources, including processing and networking resources.
- The multi-objective E2E energy consumption and utilized resource cost minimization problem is formulated as

Ref	E2E slicing scenario	Objective function	Decision variables in RAN			Decision variables in CN		E2E delay		
			BS	SC	P	Processing power	Physical links	RAN	CN	
									processing	transmission
[24]	C-RAN	min. energy consumption	-	√	-	√	-	√	√	-
[25]	C-RAN	min. delay & load balancing	-	-	-	-	-	-	-	-
[26]	C-RAN	min. MNO's operational cost	-	-	-	√	√	-	-	-
[27]	RAN & CN	-	-	-	-	-	-	-	-	-
[28]	RAN & CN	max. throughput & min. handoff cost	√	-	-	-	-	√	constant	constant
[29]	RAN & CN	max. access rate	√	√		√	√	√	√	√
[30]	RAN & CN	min. operational cost	-	√	-	-	√	-	-	√
[31]	RAN & CN	min. utilized resources cost	√	-	-	√	√	-	√	√
[32]	ICN	max. service quality min. utilized resources	-	-	-	√	√	-	-	√
[33]	RAN & CN	max. MNO's utility	-	√	√	√	√	√	√	√
our work	RAN & CN	min. energy consumption min. utilized resources cost	-	√	√	√	√	√	√	√

TABLE I: Differences of our work and existing E2E slicing works

a mixed-integer non-linear programming problem. For RAN slicing, the sub-channel and power are allocated to users of different slices, and in CN, processing, and networking resources are allocated to users' data packets. To address this problem, we propose the iterative **J**oint **R**adio and **C**ore **R**esource **A**llocation (JRCRA) method.

- We compare the performance of JRCRA with the disjoint case via simulation results. The simulation results confirm that the JRCRA algorithm decreases the total energy consumption to 34% and the total cost to 24% compared to the disjoint case. These improvements are because of jointly considering the power control and sub-channel allocation in RAN and server and physical links' bandwidth allocation in CN. Furthermore, for JRCRA, the latency constraint is defined as an E2E constraint, whereas in the disjoint case, half of the E2E latency is considered for both RAN and CN. Besides, the simulation results illustrate an optimality gap of 8% to 28% when JRCRA performance is compared to the optimal solution obtained from the exhaustive search method.

C. Paper Organization

The rest of this paper is organized as follows. In Section II, we introduce the system model and notations. The E2E energy consumption model and the cost model for utilized resources are described in Section III. In Section IV, the problem of minimization of E2E energy consumption and cost of utilized resources is formally stated. The JRCRA algorithm is presented in Section V. Finally, the simulation results and conclusion are presented in Section VI and Section VII, respectively.

II. System Model, Assumptions, and Notations

We consider uplink transmissions in a cellular network in which an MNO partitions its resources to E2E slices and rents them to MVNOs to provide eMBB and URLLC services to the end-users. Each of the E2E slices consists of RAN resources, including sub-channels and power levels, and CN resources, including the processing and networking resources. In what follows, we define the notations used for RAN and CN.

A. Radio Access Network

In RAN slicing, the resources including base stations (BSs) and sub-channels (SCs) are shared among different slices. For uplink transmissions, the coverage of a specific area is provided by a set of BSs, i.e., $\mathcal{B} = \{1, 2, \cdots, B\}$. We assume that the users are already associated with the BSs based on, for example, the reference signal received power scheme. We denote the BS serving user i by b_i and the set of users served by BS $m \in \mathcal{B}$ by $\mathcal{U}^m$. The total bandwidth of W is divided into a set of SCs, $\mathcal{C} = \{1, 2, \cdots, C\}$ shared by all slices through orthogonal frequency-division multiple access (OFDMA). Accordingly, the bandwidth of each SC k is equal to $W^k = \frac{W}{C}$. There are two eMBB and URLLC slices denoted by $\mathcal{G} = \{e, u\}$, where e and u, respectively, stand for eMBB and URLLC. The BSs serve a set of eMBB users denoted by $\mathcal{U}_e = \{1, \cdots, U_e\}$ and a set of URLLC users represented by $\mathcal{U}_u = \{1 + U_e, \cdots, U_u + U_e\}$. The set of all users is denoted by $\mathcal{U} = \mathcal{U}_e \cup \mathcal{U}_u$ and the total number of users is $U = U_e + U_u$.

Let p_i^k and $h_{i,m}^k$ denote the transmit power of user i and the path-gain from user i toward BS $m \in \mathcal{B}$ at SC k, respectively. The binary variable a_i^k denotes the SC allocation to user i. If SC k is allocated to user i, $a_i^k = 1$, otherwise $a_i^k = 0$. Let us consider $\mathbf{P} = \left[p_i^k\right]^{U \times C}$ and $\mathbf{A} = \left[a_i^k\right]^{U \times C}$ as the matrices of all transmit power levels and SC allocation of users, respectively, the received SINR at BS b_i corresponding to the signal transmitted by user i on SC k would be given by $\gamma_i^k(\mathbf{A}, \mathbf{P}) = \frac{p_i^k h_{i,b_i}^k}{I_i^k + \sigma_{b_i}^2}$, where $\sigma_{b_i}^2$ is the noise power at

BS b_i and $I_i^k = \sum_{j \notin \mathcal{U}^{b_i}} a_j^k p_j^k h_{j,b_i}^k$ is the interference due to the transmissions of users in other cells.

The data rate of eMBB user $i \in \mathcal{U}_e$ on SC k is obtained by Shannon's capacity formula as

$$R_i^k = W^k \log_2(1+\gamma_i^k), \quad \forall i \in \mathcal{U}_e. \tag{1}$$

Shannon's capacity can be approached when the blocklength of channel codes goes to infinity. However, the short blocklength and reliability of URLLC services cannot be captured by Shannon's capacity. Hence, the data rate of URLLC user $i \in \mathcal{U}_u$ is obtained by finite blocklength capacity formula as [37]

$$R_i^k = W^k \left[\log_2(1+\gamma_i^k) - \sqrt{\frac{V_i^k}{L}} Q^{-1}(\epsilon) \log_2 e \right], \quad \forall i \in \mathcal{U}_u, \tag{2}$$

where $Q^{-1}(\epsilon)$ is the inverse of Gaussian Q-function, L is the blocklength in symbols, $V_i^k = 1 - \left(1+\gamma_i^k\right)^{-2}$ is the channel dispersion. Thanks to the good channel quality and high SINR in URLLC, V_i^k can be approximated as $V_i^k \approx 1$ [38]. Finally, ϵ is a predefined threshold of decoding error probability to satisfy the reliability requirement of URLLC service [38]. Specifically, there exists an inverse correlation between the value of ϵ and the level of reliability that is demanded. In order to fulfill a severe reliability requirement such as 99.9999, a low value is assigned to ϵ, i.e., $\epsilon = 10^{-5}$, thereby, the URLLC user must transmit with more power to obtain a higher data rate.

B. Core Network

For the CN, assume that there is a cloud to provide CN resources that is modeled as a directed graph $Graph = (\mathcal{V}, \mathcal{L})$, where $\mathcal{V}$ is the set of servers and $\mathcal{L}$ is the set of directed links. The server set, $\mathcal{V}$ can be further categorized into three disjoint subsets, i.e., $\mathcal{V} = \{\mathcal{AC}, \mathcal{TR}, \mathcal{N}\}$ with $\mathcal{AC}$ representing the access switches (source nodes), $\mathcal{TR}$ representing the transport switches (destination nodes)[1], and $\mathcal{N}$ as the processing servers. Each processing server $n \in \mathcal{N}$ has a maximum processing capacity, denoted by $C_n^{\max}$ defined in terms of CPU cycles per second. Also, the maximum traffic that can be carried by each bi-directional link $l \in \mathcal{L}$ is limited to $B_l^{\max}$ bit per second. The access and transport switches do not have any computation capability and they only forward traffic.

Since users of each slice receive the same service, they have the same SFC. Each SFC consists of several different VNFs in a given order interconnected by virtual links. Let $\mathcal{S}_i = \{1, 2, \cdots, J_i\}$ denote the SFC for user i, where $\mathcal{S}_i[j],\ j \in [1, 2, \cdots, J_i]$ denotes jth VNF of user i's SFC. Furthermore, each SFC $\mathcal{S}_i$ originates from a specific source node and ends at a destination node denoted by $s_i \in \mathcal{AC}$ and $d_i \in \mathcal{TR}$, respectively. Although the SFC is the same for each service and users of each slice have the same SFC, the NFV resource allocation to users' SFC will vary depending on the network conditions, users' geographical location, and users' traffic load [8]. These factors necessitate performing NFV resource allocation per user basis in the CN [8]. Hence, similar to [20] and [39], we perform NFV resource allocation per user basis in the CN.

[1]Access switches connect BS to CN and transport switches connect CN to the transport networks.

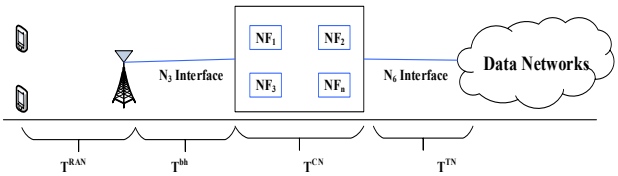

Fig. 1: Total one-way latency in the network.

To describe the embedding of the SFCs on the commodity servers and physical links, we define the binary variable $x_n^{i,j}$ to indicate the embedding of jth VNF of user i's SFC on server $n \in \mathcal{N}$. If server $n \in \mathcal{N}$ is chosen to perform jth VNF, $x_n^{i,j} = 1$, otherwise, $x_n^{i,j} = 0$. Also, $y_{l_j^{j+1}}^i$ is a binary variable that represents the allocation of physical link l to the virtual link between jth and $j+1$th VNFs of user i's SFC. If physical link l is chosen to embed the virtual link between jth and $j+1$th VNFs of user i's SFC, $y_{l_j^{j+1}}^i = 1$, otherwise, $y_{l_j^{j+1}}^i = 0$.

III. Modeling of E2E Energy Consumption and Cost of Utilized Resources

A. E2E Energy Consumption

We define the E2E energy consumption as the product of power consumption and latency [40]. In cellular networks, the users' E2E latency is the latency for receiving a packet by the destination [41] which involves a latency in all network components including RAN, backhaul, CN, and transport network (TN) [41]. To the best of our knowledge, none of the existing research works has investigated E2E latency constraint, including RAN, backhaul, CN, and TN in resource allocation problems for eMBB and URLLC services.

As shown in Fig. 1, the one-way latency of user i is a summation of latency in RAN for transmitting a packet from user i to BS (T_i^{RAN}), backhaul latency for transmitting a packet from BS to CN (T_i^{bh}), the latency in CN for processing and transmitting a packet (T_i^{CN}), and latency of the data packet transportation to the external data networks through TNs (T_i^{TN}) [41]. Therefore, user i's one-way latency is given by $T_i = T_i^{\mathrm{RAN}} + T_i^{\mathrm{bh}} + T_i^{\mathrm{CN}} + T_i^{\mathrm{TN}}$ [41]. In what follows, we describe the calculation of each component of one-way latency in more detail.

1) T_i^{RAN} is the sum of propagation latency, processing time at user i and BS b_i, queuing latency, and transmission time. The transmission time is the time for transmission of each packet between user i and BS b_i. Assuming D_i as the packet size in bits, the transmission time for each data packet is given by $\frac{D_i}{\sum_{k \in \mathcal{C}} a_i^k R_i^k}$. Therefore, considering a constant value τ_i for other components of RAN latency, we have $T_i^{\mathrm{RAN}} = \frac{D_i}{\sum_{k \in \mathcal{C}} a_i^k R_i^k} + \tau_i$.

2) The backhaul latency is equal to the time for transmitting a data packet from BS to CN and given by $T_i^{\mathrm{bh}} = D_i/Cbh^{\max}$ in which $Cbh^{\max}$ is the maximum capacity of the backhaul link.
3) To obtain T_i^{CN} as the time to perform required network functions (per packet) in CN, we need to calculate the latency to process each data packet, $T_i^{\mathrm{CN,pc}}$ and latency of transmitting packets from the access switch to the transport switch, $T_i^{\mathrm{CN,link}}$, that is $T_i^{\mathrm{CN}} = T_i^{\mathrm{CN,pc}} + T_i^{\mathrm{CN,link}}$. To calculate $T_i^{\mathrm{CN,pc}}$, similar to [42], for simplicity but without loss of generality, we assume that the processing latency of server n for a CPU cycle is given by $T_{n,\mathrm{bit}}^{\mathrm{CN,pc}} = 1/C_n^{\max}$, where $C_n^{\max}$ is the maximum capacity of server n in terms of CPU cycles per second. Assuming each bit of a packet requires C_i CPU cycles, the processing latency to process a data packet of user i in CN is obtained by $T_i^{\mathrm{CN,pc}} = \sum_{n\in\mathcal{N}} \sum_{j\in\mathcal{S}_i} x_n^{i,j} C_i D_i T_{n,\mathrm{bit}}^{\mathrm{CN,pc}}$. Furthermore, $T_i^{\mathrm{CN,link}}$ is the latency of physical links for transmission of a packet between access and transport switches. We assume that the links' bandwidth are allocated to users according to statistical multiplexing[2] [43]. Therefore, similar to [42], the transmission latency for transmitting one bit of each packet is calculated as $T_{l,\mathrm{bit}}^{\mathrm{CN,link}} = 1/B_l^{\max}$, where, $B_l^{\max}$ is the maximum bandwidth of physical link $l \in \mathcal{L}$. Accordingly, the latency on links between access and transport switches is given by $T_i^{\mathrm{CN,link}} = \sum_{l\in\mathcal{L}} \sum_{j\in\mathcal{S}_i\setminus\{J_i\}} y_{l_j^{j+1}}^i D_i T_{l,\mathrm{bit}}^{\mathrm{CN,link}}$.
4) T_i^{TN} is the latency of data transmission between CN and the external data networks[3]. For simplicity, we assume T_i^{TN} as a constant value.

Accordingly, the one-way latency for each data packet of user i can be expressed as

$$T_i = \underbrace{\tau_i + \frac{D_i}{\sum_{k\in\mathcal{C}} a_i^k R_i^k}}_{T_i^{\mathrm{RAN}}} + \underbrace{\frac{D_i}{Cbh^{\max}}}_{T_i^{\mathrm{bh}}} + \underbrace{T_i^{\mathrm{CN,pc}} + T_i^{\mathrm{CN,link}}}_{T_i^{\mathrm{CN}}} + T_i^{\mathrm{TN}}. \tag{3}$$

To calculate energy consumption, we multiply the latency of each component in (3) and its corresponding consumption power. The energy consumption to transmit a data packet of user i, E_i is obtained by $E_i = E_i^{\mathrm{RAN}} + E_i^{\mathrm{bh}} + E_i^{\mathrm{CN}} + E_i^{\mathrm{TN}}$. The energy consumption for user i in RAN, E_i^{RAN} consists of energy consumption to transmit a data packet from user i to BS b_i and a constant energy consumption denoted by E_i^{cons} for the other components in RAN. Therefore, the corresponding energy consumption for RAN is calculated by $E_i^{\mathrm{RAN}} = \frac{D_i}{\sum_{k\in\mathcal{C}} a_i^k R_i^k} \sum_{k\in\mathcal{C}} a_i^k p_i^k + E_i^{\mathrm{cons}}$, where $\sum_{k\in\mathcal{C}} a_i^k p_i^k$ is the total transmit power of user i for transmitting a data packet to BS b_i.

[2]In a statistically multiplexed link, users' traffic can be transmitted simultaneously [43].

[3]Note that, if transmitting and receiving users are served by the same BS, T_i^{TN} is set to zero [41].

The energy consumption of user i in the backhaul is expressed as $E_i^{\mathrm{bh}} = \frac{D_i}{Cbh^{\max}} p_{b_i}^{\mathrm{bh}}$, where $p_{b_i}^{\mathrm{bh}}$ is the transmit power of BS b_i to transmit a data packet to CN via backhaul link. The energy consumption for network functions in CN is given by $E_i^{\mathrm{CN}} = T_i^{\mathrm{CN,pc}} \widetilde{p}_n$, where $\widetilde{p}_n$ represents the power consumption of server $n \in \mathcal{N}$[4]. Additionally, E_i^{TN} is assumed to be the energy consumption of TNs for transmitting data packets from CN to external data networks. Note that the constant energy consumption components, including E_i^{cons}, E_i^{bh}, and E_i^{TN} do not depend on the decision variables, so they do not have any impact on the objective function of the optimization problem, therefore, for simplicity, we drop these constant components in E_i hereafter. Accordingly, the energy consumption can be expressed as

$$E_i = \frac{D_i}{\sum_{k\in\mathcal{C}} a_i^k R_i^k} \sum_{k\in\mathcal{C}} a_i^k p_i^k + T_i^{\mathrm{CN,pc}} \widetilde{p}_n. \tag{4}$$

B. Cost of Utilized Resources

To provide end-users with the required service, a MVNO should lease RAN and CN resources from the MNO. To calculate the cost of utilizing RAN resources including SCs, we introduce a unit price component for allocation of each SC k to users of slice g as $\mathrm{cost}_{g,k}^{\mathrm{sc}}$. Hence, slice g should pay $\sum_{k\in\mathcal{C}} \mathrm{cost}_{g,k}^{\mathrm{sc}} a_i^k$ for each end-user i. Therefore, $\mathrm{cost}_g^{\mathrm{RAN}} = \sum_{k\in\mathcal{C}} \mathrm{cost}_{g,k}^{\mathrm{sc}} \sum_{i\in\mathcal{U}_g} a_i^k$.

Moreover, to calculate the cost of utilized CN resources, the cost of CN resources is expressed as the unit price of each CPU cycle of servers' processing capacity (denoted by $\mathrm{cost}_{g,n}^{\mathrm{cpu}}$), and the unit price for transmitting each bit per second over physical links is denoted by $\mathrm{cost}_{g,l}^{\mathrm{link}}$. Hence, the cost of CN resources used by each slice g's users is calculated as $\mathrm{cost}_g^{\mathrm{CN}} = \sum_{n\in\mathcal{N}} \mathrm{cost}_{g,n}^{\mathrm{cpu}} \sum_{i\in\mathcal{U}_g} \sum_{j\in\mathcal{S}_i} x_n^{i,j} C_i D_i + \sum_{l\in\mathcal{L}} \mathrm{cost}_{g,l}^{\mathrm{link}} \sum_{i\in\mathcal{U}_g} \sum_{j\in\mathcal{S}_i\setminus\{J_i\}} y_{l_j^{j+1}}^i D_i$. Accordingly, the total cost of each slice g for utilizing resources is expressed as $\mathrm{cost}_g = \mathrm{cost}_g^{\mathrm{RAN}} + \mathrm{cost}_g^{\mathrm{CN}}$.

IV. Problem Formulation

In this section, we formally state the optimization problem of minimizing the E2E energy consumption and utilizing resources cost under the E2E latency constraint for each user. In this optimization problem, a number of constraints need to be satisfied. These constraints can be classified into three categories: resource constraints for RAN, resource constraints for CN, and QoS requirement constraints, which are explained below.

A. Resource Constraints for Radio Access Networks

The maximum transmit power of user i for transmitting data packets to its serving BS b_i is limited to $p_i^{\max}$, so we have

$$\mathrm{C1}: \quad \sum_{k\in\mathcal{C}} a_i^k p_i^k \le p_i^{\max}, \qquad \forall i \in \mathcal{U}. \tag{5}$$

[4]Note that it is assumed that there is no energy consumption on physical links.

The OFDMA technology imposes the following constraint which means that each SC is allocated to at most one user at each cell:

$$\text{C2}: \quad \sum_{i \in \mathcal{U}^m} a_i^k \leq 1, \qquad \forall m \in \mathcal{B}, \quad \forall k \in \mathcal{C}. \tag{6}$$

The maximum capacity of backhaul link, denoted by $Cbh^{\max}$, satisfies the following constraint:

$$\text{C3}: \quad \sum_{i \in \mathcal{U}} \sum_{k \in \mathcal{C}} a_i^k R_i^k \leq Cbh^{\max}. \tag{7}$$

B. Resource Constraints for Core Networks

For embedding the SFCs, only one server should be allocated to each VNF $j \in \mathcal{S}_i$ for SFC of each user i, that is

$$\text{C4}: \sum_{n \in \mathcal{N}} x_n^{i,j} = 1, \quad \forall i \in \mathcal{U}, \quad \forall j \in \mathcal{S}_i. \tag{8}$$

We assume that every VNF of each SFC should be mapped to a different server. Therefore, we have

$$\text{C5}: \sum_{j \in \mathcal{S}_i} x_n^{i,j} \leq 1, \quad \forall i \in \mathcal{U}, \quad \forall n \in \mathcal{N}. \tag{9}$$

Each server has limited processing capacity. The processing capacity limitation of servers, that implement the VNFs, is represented by

$$\text{C6}: \quad \sum_{i \in \mathcal{U}} \sum_{j \in \mathcal{S}_i} x_n^{i,j} C_i \sum_{k \in \mathcal{C}} a_i^k R_i^k \leq C_n^{\max}, \qquad \forall n \in \mathcal{N}. \tag{10}$$

Let $\mathcal{L}_n^{\text{out}}$ and $\mathcal{L}_n^{\text{in}}$ denote the outgoing links and incoming links from/to server n. The following constraint enforces flow conservation, i.e., the sum of all incoming and outgoing traffic in the servers that do not host VNFs should be zero. Therefore, the flow conservation constraint for routing data packets from the access switch to the transport switch should be considered; that is,

$$\begin{aligned} \text{C7}: \quad & \sum_{l \in \mathcal{L}_n^{\text{out}}} y_{l_j^{j+1}}^i - \sum_{l \in \mathcal{L}_n^{\text{in}}} y_{l_j^{j+1}}^i = x_n^{i,j} - x_n^{i,j+1}, \\ & \forall i \in \mathcal{U}, \forall n \in \mathcal{N}, \forall j \in \mathcal{S}_i \backslash \{J_i\}. \end{aligned} \tag{11}$$

The constraint related to the maximum bandwidth of the links carrying the data packets is given by

$$\text{C8}: \quad \sum_{i \in \mathcal{U}} \sum_{j \in \mathcal{S} \backslash \{J_i\}} y_{l_j^{j+1}}^i \sum_{k \in \mathcal{C}} a_i^k R_i^k \leq B_l^{\max}, \quad \forall l \in \mathcal{L}. \tag{12}$$

C. Constraint on QoS Requirement

The QoS requirement for URLLC user $i \in \mathcal{U}_u$ is defined as the maximum E2E tolerable latency. On the other hand, the QoS requirement for eMBB users is described by the minimum data rate in RAN and the maximum E2E tolerable latency. Thus, to ensure a minimum data rate for eMBB users in RAN, we define the minimum data rate (denoted by $R_i^{\min}$) constraint as follows:

$$\text{C9}: \quad \sum_{k \in \mathcal{C}} a_i^k R_i^k \geq R_i^{\min}, \qquad \forall i \in \mathcal{U}_e. \tag{13}$$

Also, we define the following E2E latency constraint for all eMBB and URLLC users:

$$\text{C10}: \ T_i \leq T_i^{\text{th}}, \ \forall i \in \mathcal{U}. \tag{14}$$

C10 means that the one-way E2E latency for transmitting a data packet from user i to the external data networks should not be larger than a maximum tolerable latency.

For E2E slicing, since minimizing the E2E energy consumption as well as the cost of utilized resources is of interest, the objective function is expressed as

$$\min_{\boldsymbol{A},\boldsymbol{P},\boldsymbol{X},\boldsymbol{Y}} \quad \sum_{i \in \mathcal{U}} E_i, \min_{\boldsymbol{A},\boldsymbol{P},\boldsymbol{X},\boldsymbol{Y}} \quad \sum_{g \in \mathcal{G}} \text{cost}_g. \tag{15}$$

Since energy consumption and utilized resources cost have different dimensions, based on (15) and using a weighted sum [44], we can have a normalized single objective function as follows:

$$F(\boldsymbol{A}, \boldsymbol{P}, \boldsymbol{X}, \boldsymbol{Y}) = \alpha \left(\frac{\sum_{i \in \mathcal{U}} E_i}{E^{\max}} \right) + (1 - \alpha) \left(\frac{\sum_{g \in \mathcal{G}} \text{cost}_g}{\text{cost}^{\max}} \right), \tag{16}$$

where $0 \leq \alpha \leq 1$ is a weighting factor that reflects the relative importance of energy consumption and utilized-resources cost [44]. Besides, $E^{\max}$ and $\text{cost}^{\max}$ reflect the maximum value of energy consumption and cost of utilized resources, respectively.

Therefore, the problem of minimizing the E2E energy consumption and utilized-resources cost is formally stated as

$$\begin{aligned} \min_{\boldsymbol{A},\boldsymbol{P},\boldsymbol{X},\boldsymbol{Y}} \quad & F(\boldsymbol{A}, \boldsymbol{P}, \boldsymbol{X}, \boldsymbol{Y}) \\ \text{s.t.} \quad & \text{C1}, \text{C2}, \text{C3}, \text{C4}, \text{C5}, \text{C6}, \text{C7}, \text{C8}, \text{C9}, \text{C10}, \\ & \text{C11}: a_i^k \in \{0,1\}, \quad \forall i \in \mathcal{U}, \ \forall k \in \mathcal{C}, \\ & \text{C12}: p_i^k \geq 0, \qquad \forall i \in \mathcal{U}, \ \forall k \in \mathcal{C}, \\ & \text{C13}: x_n^{i,j} \in \{0,1\}, \quad \forall i \in \mathcal{U}, \\ & \quad \forall j \in \mathcal{S}_i \backslash \{J_i\}, \ \forall n \in \mathcal{N}, \\ & \text{C14}: y_{l_j^{j+1}}^i \in \{0,1\}, \quad \forall l \in \mathcal{L}, \ \forall i \in \mathcal{U}, \\ & \quad \forall j \in \mathcal{S}_i \backslash \{J_i\}, \end{aligned} \tag{17}$$

where C11 shows the binary nature of the SC allocation variable. C12 represents that the transmit power on each SC should be a non-negative value. Finally, C13 and C14 indicate the binary nature of the server and physical links allocation variables, respectively.

V. Proposed Joint RAN and CN Resource Allocation Algorithm (JRCRA)

Problem (17) is complex because i) there are integer and continues variables, ii) the objective function and constraints C3, C6, C8, C9, and C10 are non-convex, and iii) there are many numbers of constraints. For these reasons, we decompose problem (17) into two sub-problems, namely, RAN resource allocation (RRA) and CN resource allocation (CRA) problems. It is worth mentioning that in the RRA problem, radio resources (i.e., power levels and SCs), and in the CRA problem, the resources of core networks (i.e., servers and

physical links) are allocated to users. By doing so, given $\boldsymbol{X}$ and $\boldsymbol{Y}$ the RRA problem is expressed as

$$\begin{aligned} \min_{\boldsymbol{A},\boldsymbol{P}} \quad & F(\boldsymbol{A},\boldsymbol{P},\boldsymbol{X},\boldsymbol{Y}) \\ \text{s.t.} \quad & \text{C1, C2, C3, C6, C8, C9, C10, C11, C12,} \end{aligned} \tag{18}$$

and given $\boldsymbol{A}$ and $\boldsymbol{P}$ the CRA problem is formulated as

$$\begin{aligned} \min_{\boldsymbol{X},\boldsymbol{Y}} \quad & F(\boldsymbol{A},\boldsymbol{P},\boldsymbol{X},\boldsymbol{Y}) \\ \text{s.t.} \quad & \text{C4, C5, C6, C7, C8, C10, C13, C14.} \end{aligned} \tag{19}$$

Although (18) and (19) are decomposed sub-problems of (17), there are constraints C6, C8, and C10 which are common in both sub-problems, and these constraints make (18) and (19) to be coupled. Iteratively solving the sub-problems (18) and (19) gives a sub-optimal solution to the main problem (17). In fact, the existence of constraint C10 in both sub-problems (18) and (19) gives us a degree of freedom to have more choices for the amount of tolerable latency in RAN and CN. By doing so, the value of T_i^{th} is flexibly divided between RAN and CN based on the resource allocation in them. However, if we want to solve sub-problems (18) and (19) disjointly, we have to consider part of the value of T_i^{th} (for example, $\frac{1}{2}T_i^{\text{th}}$) as a tolerable latency in RAN and part of it in CN. Furthermore, constraints C6 and C8 make CN resources more efficiently allocated to users, because if sub-problems (18) and (19) are solved disjointly, we have to set the value of the data rate in constraints C6 and C8 in sub-problem (19) equal to a predetermined value. While this value may be very far from the value obtained from power control and SC allocation in RAN.

Moreover, via simulation results, we show that iterative solving of sub-problems (18) and (19) improves energy consumption by 34% and the cost by 24% compared to the disjoint case. In the disjoint case, sub-problems (18) and (19) are solved disjointly by placing fixed values for the data rate in constraints C6 and C8 and placing half of T_i^{th} in the RAN and half in the CN as a tolerable latency in constraint C10.

A. Solving the RRA Sub-Problem

It is proved in [45] that the power control and sub-channel allocation problems to minimize the total transmit power and maximize the total data rate are NP-hard. Ignoring the cost objective function in (15), the energy consumption problem can be regarded as a multi-objective problem to jointly minimize total transmit power and maximize total data rate. According to [45], both of these problems are NP-hard. Therefore, it can be concluded that problem (18) is also NP-hard.

To tackle this difficulty, we decompose the RRA problem (18) into two SC allocation and power control sub-problems. In other words, we solve this problem in two steps. In the first step, for a given power control, the SC allocation is performed, and in the second step, for a given SC allocation, the power control is performed. The output of each step is the input of the other step, i.e., $\boldsymbol{A}(0) \rightarrow \boldsymbol{P}(0) \rightarrow \cdots \rightarrow \boldsymbol{A}(t-1) \rightarrow \boldsymbol{P}(t-1) \rightarrow \boldsymbol{A}(t) \rightarrow \boldsymbol{P}(t)$, where $t \geq 0$ is the iteration number, $\boldsymbol{A}(t)$ and $\boldsymbol{P}(t)$ are the optimal values at iteration t. $\boldsymbol{A}(t)$ and $\boldsymbol{P}(t)$ are obtained by solving the convex transformation of corresponding optimization problems which is explained below. The iterative procedure stops when $\mid \boldsymbol{A}(t) - \boldsymbol{A}(t-1) \mid \leq \epsilon_1$ and $\mid \boldsymbol{P}(t) - \boldsymbol{P}(t-1) \mid \leq \epsilon_2$, where $0 < \epsilon_1 \ll 1$ and $0 < \epsilon_2 \ll 1$.

1) Sub-Channel Allocation Sub-Problem: The SC allocation sub-problem is formulated as

$$\min_{\boldsymbol{A}} \quad F(\boldsymbol{A}) \quad \text{s.t.} \quad \text{C1, C2, C3, C6, C8, C9, C10, C11.} \tag{20}$$

Because of the binary nature of the SC allocation variable (i.e., a_i^k), problem (20) is NP-hard. Therefore, to overcome this difficulty, similar to [47] and [48], we replace the binary variables constraint C11 in (20) by following equivalent constraints:

$$\begin{aligned} &\text{C11.1}: \ \sum_{i\in\mathcal{U}}\sum_{k\in\mathcal{C}}\left(a_i^k - \left(a_i^k\right)^2\right) \leq 0, \\ &\text{C11.2}: \ 0 \leq a_i^k \leq 1, \ \forall i \in \mathcal{U}, \ \forall k \in \mathcal{C}. \end{aligned} \tag{21}$$

By substituting the binary constraints C11 in (20) with constraints C11.1 and C11.2 in (21), problem (20) is transformed into a non-convex problem (due to constraint C11.1). The following theorem is for handling the constraint C11.1.

Theorem 1. *For a sufficiently large value of $\lambda \gg 1$, problem* (20) *is equivalent to*

$$\begin{aligned} \min_{\boldsymbol{A}} \quad & F(\boldsymbol{A}) + \lambda \sum_{i\in\mathcal{U}}\sum_{k\in\mathcal{C}}\left(a_i^k - \left(a_i^k\right)^2\right) \\ \textit{s.t.} \quad & \text{C1, C2, C3, C6, C8, C9, C10, C11.2,} \end{aligned} \tag{22}$$

where λ acts as a penalty factor to penalize the objective function for any a_i^k that is not equal to 0 *or* 1.

Proof. The proof is given in **Appendix A**. ■

Let $f(\boldsymbol{A}) = F(\boldsymbol{A}) + \lambda \sum_{i\in\mathcal{U}}\sum_{k\in\mathcal{C}} a_i^k$ and $g(\boldsymbol{A}) = \lambda \sum_{i\in\mathcal{U}}\sum_{k\in\mathcal{C}} \left(a_i^k\right)^2$. Therefore, the objective function of problem (22) can be written as the difference of two convex functions $f(\boldsymbol{A})$ and $g(\boldsymbol{A})$. Therefore, problem (22) is a DC programming problem. The majorization-minimization approximation method is a well-known method for approximating DC functions [49] as convex ones. One approach to do majorization-minimization approximation is the first-order Taylor approximation method. In the first-order Taylor approximation, the function $f(x)$ is approximated as

$$f(x) \approx f(\overline{x}) + \nabla_x f(\overline{x})(x - \overline{x}), \tag{23}$$

where $\overline{x}$ is a feasible initial point. Using (23), function $f(x)$ becomes a linear function.

To convexify the objective function of (22), we approximate $g(\boldsymbol{A}) = \lambda \sum_{i\in\mathcal{U}}\sum_{k\in\mathcal{C}} \left(a_i^k\right)^2$ by its first-order Taylor approximation as $g(\boldsymbol{A}) = g(\boldsymbol{A}(t-1)) + \nabla_{\boldsymbol{A}} g(\boldsymbol{A}(t-1))(\boldsymbol{A} - \boldsymbol{A}(t-1))$, where $\boldsymbol{A}(t-1)$ is optimal SC allocation of previous iteration. By doing so, problem (22) can be rewritten as

$$\begin{aligned} \min_{\boldsymbol{A}} \quad & F(\boldsymbol{A}) + \lambda \sum_{i\in\mathcal{U}}\sum_{k\in\mathcal{C}} a_i^k \\ & - \lambda \sum_{i\in\mathcal{U}}\sum_{k\in\mathcal{C}} \left[2a_i^k a_i^k(t-1) - (a_i^k(t-1))^2\right] \\ \text{s.t.} \quad & \text{C1, C2, C3, C6, C8, C9, C10, C11.2.} \end{aligned} \tag{24}$$

Problem (24) is a convex optimization problem and its optimal solution can be obtained by interior-point [51] method or using CVX toolbox [52] in polynomial time.

Proposition 1. *The optimal solution of problem* (24) *at each iteration* t *provides a tight upper bound and local optimum for problem* (20).

Proof. The proof is given in **Appendix B**. ■

2) Power Control Sub-Problem: The power control sub-problem is expressed as

$$\min_{\boldsymbol{P}} \quad F(\boldsymbol{P}) \quad \text{s.t.} \quad \text{C1}, \text{C3}, \text{C6}, \text{C8}, \text{C9}, \text{C10}, \text{C12}. \tag{25}$$

The power control sub-problem (25) is a non-convex problem due to i) the non-convexity of data rate functions defined in (1) and (2), ii) the objective function is non-convex, even if the data rate function is a concave function since the objective function is defined as the summation of fractions that there is no solution to solve such problems, and iii) the constraints C3, C6, and C8 which represent, respectively, the backhaul link capacity, the servers processing capacity, and the physical link bandwidth constraints are non-convex, even if the data rate function is a concave function. In what follows, we deal with each of these difficulties.

(a) Non-convexity of data rate functions in (1) and (2): The data rate function of eMBB users, (1), can be written as the difference of two convex functions as $R_i^k = W^k\left[f(\boldsymbol{P}) - g(\boldsymbol{P})\right]$, where $f(\boldsymbol{P})=\log_2\left(p_i^k h_{i,b_i}^k + \sum\limits_{j\notin\mathcal{U}^{b_i}} p_j^k h_{j,b_i}^k + \sigma_{b_i}^2\right)$ and $g(\boldsymbol{P})=\log_2\left(\sum\limits_{j\notin\mathcal{U}^{b_i}} p_j^k h_{j,b_i}^k + \sigma_{b_i}^2\right)$. Also, the data rate function of URLLC users, (2), can be written as $R_i^k = W^k\left[f(\boldsymbol{P}) - g(\boldsymbol{P})\right] - W^k\sqrt{\frac{1}{L}}Q^{-1}(\epsilon)\log_2 e$. To convexify the data rate functions, we approximate function $g(\boldsymbol{P})$ by employing the first-order Taylor approximation in (23).

(b) Non-convexity of the objective function in problem (25): Due to the existence of equation $\frac{D_i}{\sum\limits_{k\in\mathcal{C}} a_i^k R_i^k}\sum\limits_{k\in\mathcal{C}} a_i^k p_i^k$ which defines users' energy consumption for transmission of a data packet in RAN, the objective function is non-convex. It can be easily observed that the derivative of energy consumption with respect to the user's transmit power is positive which implies energy consumption increases with increasing of transmit power. Thus, the minimal energy consumption for sending a data packet can be obtained if the minimal transmission power is applied. Furthermore, according to the constraint C10 of problem (25), which guarantees the maximum tolerable latency of user i, the transmission latency for sending a packet to BS for user i is upper bounded by $T_i^{\text{th}} - \zeta_i$ where $\zeta_i = \tau_i + \frac{D_i}{Cbh^{\max}} + T_i^{\text{CN,pc}} + T_i^{\text{CN,link}} + T_i^{\text{TN}}$ i.e., $\frac{D_i}{\sum\limits_{k\in\mathcal{C}} a_i^k R_i^k} \leq T_i^{\text{th}} - \zeta_i$. Therefore, the energy consumption of user i for transmitting a data packet in RAN, i.e., $\frac{D_i}{\sum\limits_{k\in\mathcal{C}} a_i^k R_i^k}\sum\limits_{k\in\mathcal{C}} a_i^k p_i^k$ is approximated as $(T_i^{\text{th}} - \zeta_i)\sum\limits_{k\in\mathcal{C}} a_i^k p_i^k$.

(c) Non-convexity of constraints C3, C6, and C8: To overcome this difficulty, inspired by [50] we introduce auxiliary variable $\nu_i^k,\ \forall i \in \mathcal{U},\ \forall k \in \mathcal{C}$. We substitute R_i^k by ν_i^k in the objective function and constraints C3, C6, and C8, C9, and C10. Also, to handle the auxiliary variable ν_i^k, we add the following constraint to problem (25):

$$\text{C15}: \ \nu_i^k \leq \widetilde{R}_i^k, \tag{26}$$

where $\widetilde{R}_i^k = W^k\left[f(\boldsymbol{P}) - \widetilde{g}(\boldsymbol{P})\right]$, for eMBB user $i \in \mathcal{U}_e$ and $\widetilde{R}_i^k = W^k\left[f(\boldsymbol{P}) - \widetilde{g}(\boldsymbol{P})\right] - W^k\sqrt{\frac{1}{L}}Q^{-1}(\epsilon)\log_2 e$, for URLLC user $i \in \mathcal{U}_u$ in which $\widetilde{g}(\boldsymbol{P})$ is the first-order Taylor approximation of $g(\boldsymbol{P})$.

By taking the steps explained above, problem (25) is transformed into the following problem:

$$\begin{aligned}
\min_{\boldsymbol{P},\boldsymbol{\nu}} \quad & F'(\boldsymbol{P},\boldsymbol{\nu}) \\
\text{s.t.} \quad & \text{C1}: \ \sum_{k\in\mathcal{C}} a_i^k p_i^k \leq p_i^{\max}, \qquad \forall i \in \mathcal{U}, \\
& \text{C2}: \ \sum_{i\in\mathcal{U}^m} a_i^k \leq 1, \qquad \forall m \in \mathcal{B}, \quad \forall k \in \mathcal{C}, \\
& \text{C3}: \ \sum_{i\in\mathcal{U}}\sum_{k\in\mathcal{C}} a_i^k \nu_i^k \leq Cbh^{\max}, \\
& \text{C}'6: \ \sum_{i\in\mathcal{U}}\sum_{j\in\mathcal{S}_i} x_n^{i,j} C_i \sum_{k\in\mathcal{C}} a_i^k \nu_i^k \leq C_n^{\max}, \forall n \in \mathcal{N}, \\
& \text{C}'8: \ \sum_{i\in\mathcal{U}}\sum_{j\in\mathcal{S}\setminus\{J_i\}} y_{l_j^{j+1}}^i \sum_{k\in\mathcal{C}} a_i^k \nu_i^k \leq B_l^{\max}, \quad \forall l \in \mathcal{L}, \\
& \text{C}'9: \ \sum_{k\in\mathcal{C}} a_i^k \nu_i^k \geq R_i^{\min}, \qquad \forall i \in \mathcal{U}_e, \\
& \text{C}'10: \ \frac{D_i}{\sum\limits_{k\in\mathcal{C}} a_i^k \nu_i^k} + \zeta_i \leq T_i^{\text{th}}, \quad \forall i \in \mathcal{U}, \\
& \text{C12}: p_i^k \geq 0, \qquad \forall i \in \mathcal{U},\ \forall k \in \mathcal{C}, \\
& \text{C15}: \ \nu_i^k \leq \widetilde{R}_i^k, \qquad \forall i \in \mathcal{U},\ \forall k \in \mathcal{C},
\end{aligned} \tag{27}$$

where

$$F'(\boldsymbol{P},\boldsymbol{\nu}) = \alpha\big(\sum_{i\in\mathcal{U}} E_i'\big) + (1-\alpha)\sum_{g\in\mathcal{G}} \text{cost}_g, \tag{28}$$

in which

$$E_i' = (T_i^{\text{th}} - \zeta_i)\sum_{k\in\mathcal{C}} a_i^k p_i^k + T_i^{\text{CN,pc}}\widetilde{p}_n. \tag{29}$$

Problem (27) is now a convex optimization problem and its optimal solution can be obtained by interior-point [51] method or using CVX toolbox [52] in polynomial time.

B. Solving the CRA Sub-Problem

It is proved in [46] that because of the binary nature of server allocation (i.e., $x_n^{i,j}$) and physical link allocation (i.e., $y_{l_j^{j+1}}^i$) variables, problem (19) is NP-hard. Therefore,

to overcome this difficulty, similar to the approach used for solving the SC allocation sub-problem in Section V-A, we replace the server and physical links allocation binary variables constraints C13 and C14 in (19), respectively, by the following equivalent constraints:

$$\begin{aligned} &\text{C13.1}: \sum_{n\in\mathcal{N}}\sum_{i\in\mathcal{U}}\sum_{j\in\mathcal{S}_i}(x_n^{i,j}-\left(x_n^{i,j}\right)^2)\leq 0,\\ &\text{C13.2}: \ 0\leq x_n^{i,j}\leq 1, \quad \forall n\in\mathcal{N},\forall i\in\mathcal{U},\forall j\in\mathcal{S}_i, \end{aligned} \tag{30}$$

and

$$\begin{aligned} &\text{C14.1}: \sum_{l\in\mathcal{L}}\sum_{i\in\mathcal{U}}\sum_{j\in\mathcal{S}_i\backslash\{J_i\}}\left(y_{l_j^{j+1}}^i-\left(y_{l_j^{j+1}}^i\right)^2\right)\leq 0,\\ &\text{C14.2}: \ 0\leq y_{l_j^{j+1}}^i\leq 1, \forall l\in\mathcal{L},\forall i\in\mathcal{U},\forall j\in\mathcal{S}_i\backslash\{J_i\}. \end{aligned} \tag{31}$$

By substituting the binary constraint C13 in (19) with constraints C13.1, C13.2 and constraint C14 in (19) with C14.1, and C14.2 in (30) and (31), problem (19) is transformed into a non-convex problem (due to the constraints C13.1 and C14.1). The following theorem is for handling constraints C13.1 and C14.1.

Theorem 2. *For sufficiently large values of* $\lambda_1 \gg 1$ *and* $\lambda_2 \gg 1$*, problem* (19) *is equivalent to the following problem:*

$$\begin{aligned} \min_{\boldsymbol{X},\boldsymbol{Y}} \quad & F(\boldsymbol{X},\boldsymbol{Y})+\lambda_1\sum_{n\in\mathcal{N}}\sum_{i\in\mathcal{U}}\sum_{j\in\mathcal{S}_i\backslash\{J_i\}}\left(x_n^{i,j}-\left(x_n^{i,j}\right)^2\right)+\\ & \lambda_2\sum_{l\in\mathcal{L}}\sum_{i\in\mathcal{U}}\sum_{j\in\mathcal{S}_i}\left(y_{l_j^{j+1}}^i-\left(y_{l_j^{j+1}}^i\right)^2\right)\\ \textit{s.t.} \quad & \text{C4, C5, C6, C7, C8, C10, C13.2, C14.2,} \end{aligned} \tag{32}$$

where λ_1 *and* λ_2 *act as penalty factors to penalize the objective function for any* $x_n^{i,j}$ *and* $y_{l_j^{j+1}}^i$ *that is not equal to* 0 *or* 1.

Proof. The proof is similar to that of **Theorem 1**. ■

Let $f(\boldsymbol{X},\boldsymbol{Y}) = F(\boldsymbol{X},\boldsymbol{Y}) + \lambda_1\sum_{n\in\mathcal{N}}\sum_{i\in\mathcal{U}}\sum_{j\in\mathcal{S}_i}x_n^{i,j} + \lambda_2\sum_{l\in\mathcal{L}}\sum_{i\in\mathcal{U}}\sum_{j\in\mathcal{S}_i\backslash\{J_i\}}y_{l_j^{j+1}}^i$ and $g(\boldsymbol{X},\boldsymbol{Y}) = \lambda_1\sum_{n\in\mathcal{N}}\sum_{i\in\mathcal{U}}\sum_{j\in\mathcal{S}_i}\left(x_n^{i,j}\right)^2 + \lambda_2\sum_{l\in\mathcal{L}}\sum_{i\in\mathcal{U}}\sum_{j\in\mathcal{S}_i\backslash\{J_i\}}\left(y_{l_j^{j+1}}^i\right)^2$. Therefore, the objective function of problem (32) can be written as the difference of two convex functions $f(\boldsymbol{X},\boldsymbol{Y})$ and $g(\boldsymbol{X},\boldsymbol{Y})$. Therefore, problem (32) is a DC programming problem. To convexify the objective function of (32), we approximate $g(\boldsymbol{X},\boldsymbol{Y}) = \lambda_1\sum_{n\in\mathcal{N}}\sum_{i\in\mathcal{U}}\sum_{j\in\mathcal{S}_i}\left(x_n^{i,j}\right)^2 + \lambda_2\sum_{l\in\mathcal{L}}\sum_{i\in\mathcal{U}}\sum_{j\in\mathcal{S}_i\backslash\{J_i\}}\left(y_{l_j^{j+1}}^i\right)^2$ by its first-order Taylor approximation in (23) as $g(\boldsymbol{X},\boldsymbol{Y})$=$g(\boldsymbol{X}(t-1),\boldsymbol{Y}(t-1))+\nabla_{\boldsymbol{X}}g(\boldsymbol{X}(t-1),\boldsymbol{Y})(\boldsymbol{X}-\boldsymbol{X}(t-1)) + \nabla_{\boldsymbol{Y}}g(\boldsymbol{X},\boldsymbol{Y}(t-1))(\boldsymbol{Y}-\boldsymbol{Y}(t-1))$, where $\boldsymbol{X}(t-1)$ and $\boldsymbol{Y}(t-1)$ is the optimal solution of previous iteration. By doing so, problem (32) can be rewritten as

$$\begin{aligned} \min_{\boldsymbol{X},\boldsymbol{Y}} \quad & F(\boldsymbol{X},\boldsymbol{Y})+\lambda_1\sum_{n\in\mathcal{N}}\sum_{i\in\mathcal{U}}\sum_{j\in\mathcal{S}_i}x_n^{i,j}+\lambda_2\sum_{n\in\mathcal{N}}\sum_{i\in\mathcal{U}}\sum_{j\in\mathcal{S}_i\backslash\{J_i\}}y_{l_j^{j+1}}^i\\ & -\lambda_1\sum_{i\in\mathcal{U}}\sum_{j\in\mathcal{S}_i}\sum_{n\in\mathcal{N}}\left[2x_n^{i,j}x_n^{i,j}(t-1)-(x_n^{i,j}(t-1))^2\right]\\ & -\lambda_2\sum_{l\in\mathcal{L}}\sum_{i\in\mathcal{U}}\sum_{j\in\mathcal{S}_i}\left[2y_{l_j^{j+1}}^i y_{l_j^{j+1}}^i(t-1)-(y_{l_j^{j+1}}^i(t-1))^2\right]\\ \text{s.t.} \quad & \text{C4, C5, C6, C7, C8, C10, C13.2, C14.2.} \end{aligned} \tag{33}$$

Problem (33) is a convex and linear optimization problem and its optimal solution can be obtained by interior-point [51] method or using the CVX toolbox [52].

Proposition 2. *The optimal solution of problem* (33) *at each iteration* t *provides a tight upper bound and local optimal for problem* (19)*.*

Proof. The proof is similar to the proof of **Proposition 1**. ■

To obtain a local optimum of problem (17), we employ the iterative JRCRA algorithm to tighten the obtained upper bound as summarized in **Algorithm 1**. In each iteration, the convex problems in (24), (27), and (33) are solved efficiently by the interior-point method. By solving the convex problems in (24), (27), and (33), the proposed iterative scheme generates a sequence of feasible solutions $\boldsymbol{A}(t)$, $\boldsymbol{P}(t)$, $\boldsymbol{X}(t)$, and $\boldsymbol{Y}(t)$ successively. The proposed sub-optimal iterative algorithm converges to a locally optimal solution of problem (17) in polynomial time.

Algorithm 1: Our proposed JRCRA algorithm to solve problem (17)

Input: The maximum number of iterations $t^{\max}$, $T^{\max}$, $\lambda \gg 1$, $\lambda_1 \gg 1$, $\lambda_2 \gg 1$, iteration index $t=0$, $T=0$, and a feasible initial point $\boldsymbol{A}(0)$, $\boldsymbol{P}(0)$, $\boldsymbol{X}(0)$, and $\boldsymbol{Y}(0)$.
Output: $\boldsymbol{A}(T)$, $\boldsymbol{P}(T)$, $\boldsymbol{X}(T)$, and $\boldsymbol{Y}(T)$.
Repeat
  Set $T \leftarrow T+1$.
  Step 1: Solving the RRA problem
  Repeat
    Set $t \leftarrow t+1$.
    Solve convex optimization problem (24) to obtain $\boldsymbol{A}(t)$.
    Solve convex optimization problem (27) to obtain $\boldsymbol{P}(t)$.
  Until convergence or $t = t^{\max}$.
  Set $A(T)=A(t)$ and $P(T)=P(t)$.
  Step 2: Solving the CRA problem
    Solve convex optimization problem (33) to obtain $\boldsymbol{X}(T)$ and $\boldsymbol{Y}(T)$.
Until convergence or $T = T^{\max}$.

C. Analysis of Computational Complexity

In this section, the computational complexity of our proposed JRCRA algorithm to solve problem (17) is analyzed. In **Algorithm 1**, for updating SC allocation, power control, and server and link allocation in the inner loop, we use the interior-point method. The complexity of the interior-point method is calculated by $\frac{\log(Q/(\rho\delta))}{\log(\eta)}$, where Q is the number of constraints in problems (24), (27) and (33) given

by $Q = UJ+U|\mathcal{N}|+|3\mathcal{N}|+2UJ|\mathcal{N}|+|3\mathcal{L}|+5U+2U_e+2BC+3UC+2UJ|\mathcal{L}|+2$ in which $J = \max_{i\in\mathcal{U}} J_i$, ρ is the initial point to approximate the accuracy of interior point method, $0 < \delta \ll 1$ is the stopping criterion for the interior point method, and η is used for updating the accuracy of interior point method [51]. Accordingly, the computational complexity of Algorithm 1 is $O\left(t^{\max}\left(\frac{\log((2UJ|\mathcal{N}|+UJ|\mathcal{L}|+3UC))/(\rho\delta))}{\log(\eta)}\right)\right)$.

VI. Performance Evaluation

A. Setup for Simulations and Performance Benchmarking

To evaluate our proposed JRCRA algorithm, we consider a multi-cell network consisting of $B = 2$ BSs with a $500\text{m} \times 500\text{m}$ coverage area, where each BS serves eMBB and URLLC slices. Furthermore, users of different slices are randomly distributed in each cell. Similar to [53], the path-gain from each user to the BSs is modeled by $h^k_{i,m} = \mu^k d^{-\beta}_{i,m}$, where $d_{i,m}$ is the distance between user i and BS m, μ^k is a random value that is generated by the Rayleigh distribution, and $\beta = 3$ is the path-loss exponent. Additionally, for CN, it is assumed that servers are connected with randomly established physical links. Unless stated otherwise, all other simulation parameters are listed in Table II.

Moreover, the International Telecommunication Union (ITU) [56] has established the maximum delay threshold as $T^{\text{th}}_i = 4\text{ms}$ for eMBB users and $T^{\text{th}}_i = 1\text{ms}$ for URLLC users. Accordingly, for generating following simulation figures, we set the maximum threshold delay for eMBB users to $T^{\text{th}}_i = 4\text{ms},\ \forall i \in \mathcal{U}_e$, while employing varying values such as $T^{\text{th}}_i = \{0.5, 1, 2\}\text{ms},\ \forall i \in \mathcal{U}_u$, for URLLC users.

In what follows, we evaluate our proposed JRCRA algorithm in terms of the E2E energy consumption and utilized resources cost with respect to the different criteria, including the maximum tolerable latency of URLLC users, the minimum data rate of eMBB users, and the number of SCs, servers, and users in each slice. Then, to verify the performance of the JRCRA algorithm, we compare it with the disjoint case named JRCRA-DS, where the problems RRA (18) and CRA (19) are solved disjointly. To benchmark our proposed algorithm, we compare it with the JRCRA-SP benchmark in which sub-channel allocation is already allocated to users. Besides, to evaluate the efficiency of the JRCRA algorithm compared to the existing algorithm, we compare the performance of JRCRA with the algorithm proposed in [33]. Finally, we analyze the optimality gap of the JRCRA algorithm by comparing its performance with the optimal solution obtained from the exhaustive search method.

B. Performance of JRCRA in Terms of E2E Energy and Resource Cost

Fig. 2 illustrates the performance of our proposed JRCRA algorithm for different values of tolerable latency for URLLC users and minimum data rate eMBB users, and different number of sub-channels for a network with $U_g = 10$ users per each slice. To generate Fig. 2, the number of servers in the CN is set to $|\mathcal{N}| = 20$. From Fig. 2, it can be observed that when the URLLC users' tolerable latency increases and eMBB users' minimum data rate decreases, energy consumption and cost decrease. For the URLLC users, because of the increase in their tolerable latency, they need less data rate. Therefore, they transmit with lower transmit power. Likewise, the eMBB users transmit with lower transmit power to achieve a lower minimum data rate. The lower transmit power creates less interference to other users; as a result, the users can achieve a higher data rate with a lower transmit power. Furthermore, by increasing the number of sub-channels, the energy consumption and cost decrease due to channel diversity and allocation of more sub-channels to each user. Moreover, thanks to the sub-channel allocation in JRCRA, it obtains lower energy consumption and cost in comparison with JRCRA-SP.

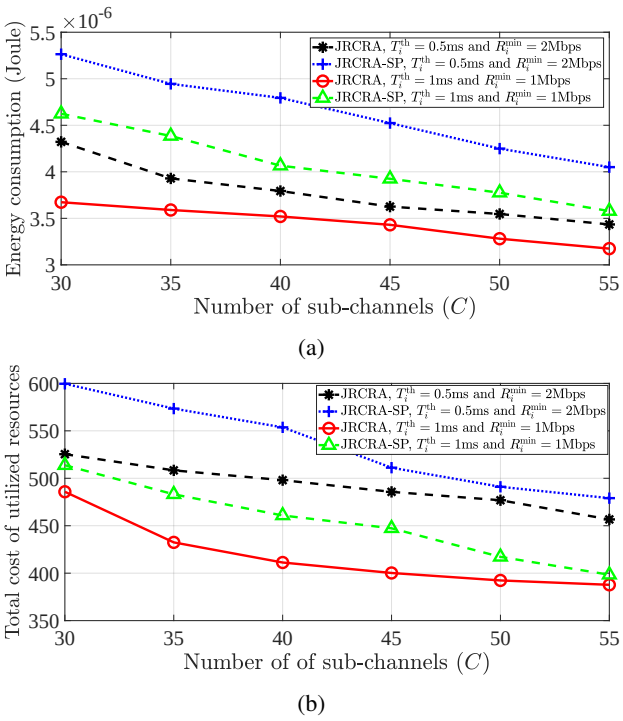

Fig. 2: (a) Energy consumption and (b) total cost of utilized resources versus the number of SCs.

Fig. 3 shows energy consumption and cost for different values of minimum data rate for eMBB users and maximum tolerable latency for URLLC users, and different number of servers. For this figure, we set the number of sub-channels to $C = 30$. As can be observed, the energy consumption and cost decrease with increasing values of maximum tolerable latency for URLLC users and decreasing values of minimum data rate for eMBB users. The reason is that with a higher maximum tolerable latency, the URLLC users can transmit at lower transmit power, decreasing interference to other users. Due to the same reason, energy consumption and cost can be reduced by decreasing the minimum data rate for the eMBB users. In addition, in Fig. 3, the energy consumption and cost decrease as the number of servers increases since, for allocating to each VNF, there are more servers to choose from. Additionally, since the processing capacities of the servers are randomly set, the probability of the existence of high-capacity servers increases, and hence we can run more VNFs on high-

TABLE II: Simulation parameters

Parameter	Value
α	0.5
Each SC bandwidth (W^k)	180KHz
user i's maximum transmit power ($p_i^{\max}$)	100mW
Backhaul link capacity ($Cbh^{\max}$)	1Gbps
Noise power (σ_m^2)	10^{-14}W
Data packet size (D_i)	32byte for URLLC users and 1500byte for eMBB users
Maximum tolerable delay of eMBB users ($T_i^{\text{th}}, \forall i \in \mathcal{U}_e$)	4ms
Link bandwidth ($B_l^{\max}$)	random selection of $[50, 100]$Mbps
Servers CPU capacity ($C_n^{\max}$)	random selection of $[10, 20]$MHz
Constant latency in RAN (τ_i)	0.25ms
Transport network latency (T_i^{TN})	0.1ms
Power consumption of server ($\widetilde{p}_n$)	random selection $[1, 10]$W
unit price of each CPU cycle ($\text{cost}_{g,n}^{\text{cpu}}$)	random selection of $[0.1, 1]$
unit price for transmitted each bit per second ($\text{cost}_{g,l}^{\text{link}}$)	random selection of $[0.1, 1]$
unit price of each SC ($\text{cost}_{g,k}^{\text{sc}}$)	random selection of $[1, 5]$

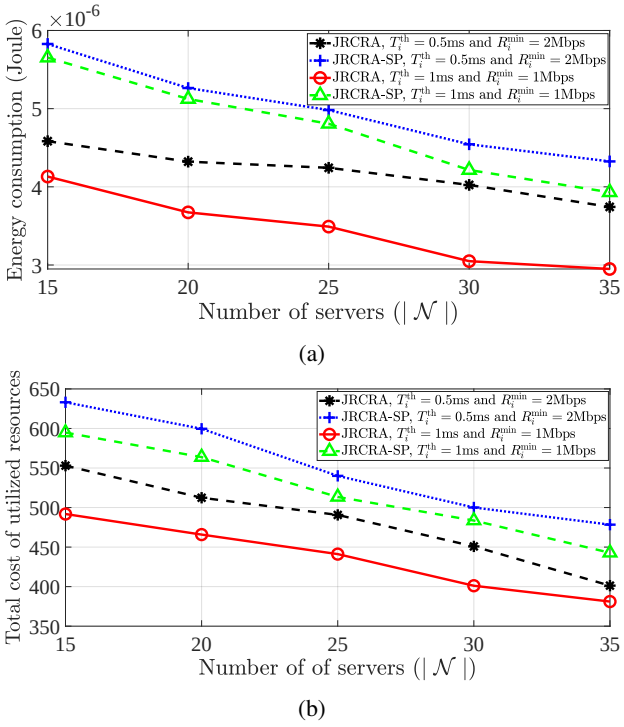

Fig. 3: (a) Energy consumption and (b) total cost of utilized resources versus the number of servers.

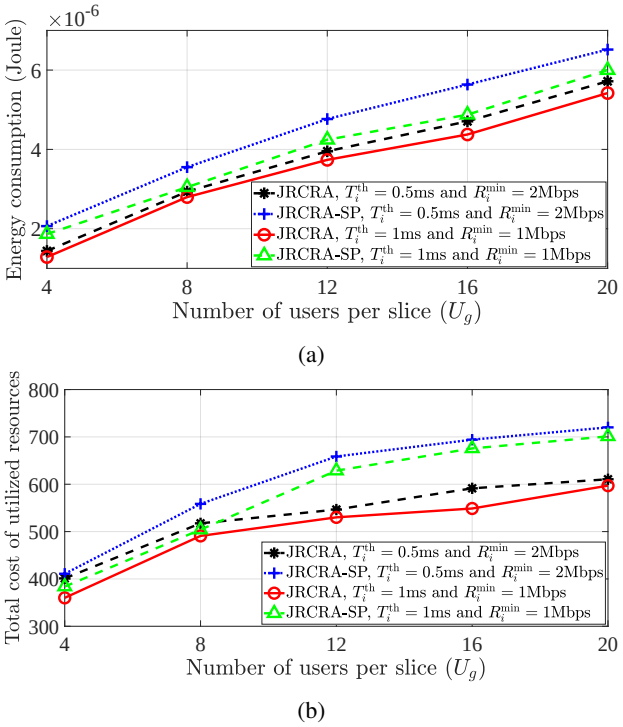

Fig. 4: (a) Energy consumption and (b) total cost of utilized resources versus the number of users per each slice.

capacity servers. Furthermore, since in JRCRA-SP, the sub-channels are already allocated to users, it is outperformed by JRCRA in terms of energy consumption and cost of utilized resources.

Fig. 4 depicts the performance of the JRCRA algorithm for different values of minimum data rate for eMBB users and maximum tolerable latency for URLLC users, and for different number of users. In this case, the number of sub-channels is set to $C = 50$. We can observe that energy consumption and cost increase when the number of users increases. Since more users at each slice need to achieve their QoS, they cause more interference. Consequently, the users cannot achieve a high data rate while transmitting at high transmit power. Moreover, with decreasing values of maximum tolerable latency for URLLC users and increasing values of minimum data rate for eMBB users, the energy consumption and cost increase due to the transmission with higher transmit power. Additionally, owing to the additional degree of freedom in JRCRA, it outperforms JRCRA-SP in terms of energy consumption and cost of utilized resources.

In Fig. 5, we analyze the impact of URLLC users' QoS parameters, namely delay and reliability, on energy consumption and resource utilization costs. To generate Fig. 5, it is assumed that the network comprises $U = 20$ users and $C = 30$ sub-channels. Additionally, the reliability requirement of URLLC users in Fig. 5 is depicted in terms of the decoding error probability, specifically, reliability is defined as $1 - \epsilon$. As depicted in Fig. 5a, the increase in the decoding error probability ϵ results in a reduction in energy consumption. This is because, with a higher decoding error probability,

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2023.3341810

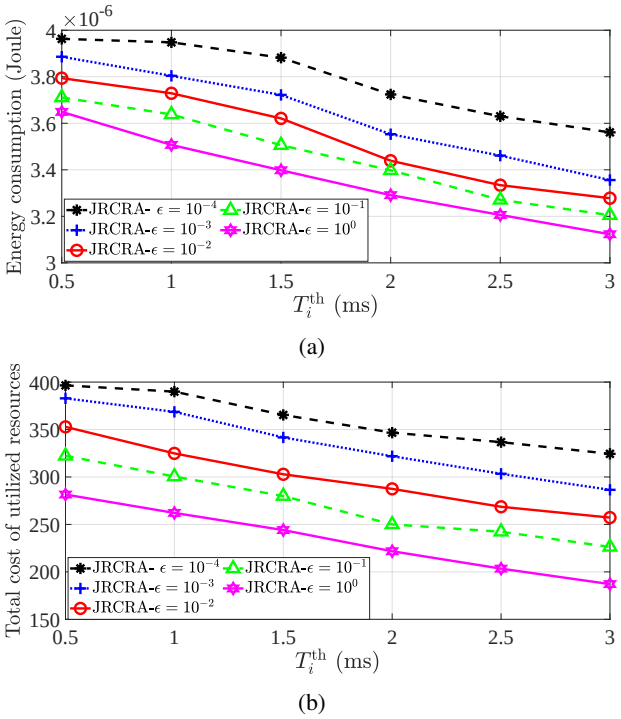

Fig. 5: (a) Energy consumption and (b) total utilized resources' cost versus delay threshold and reliability of URLLC users.

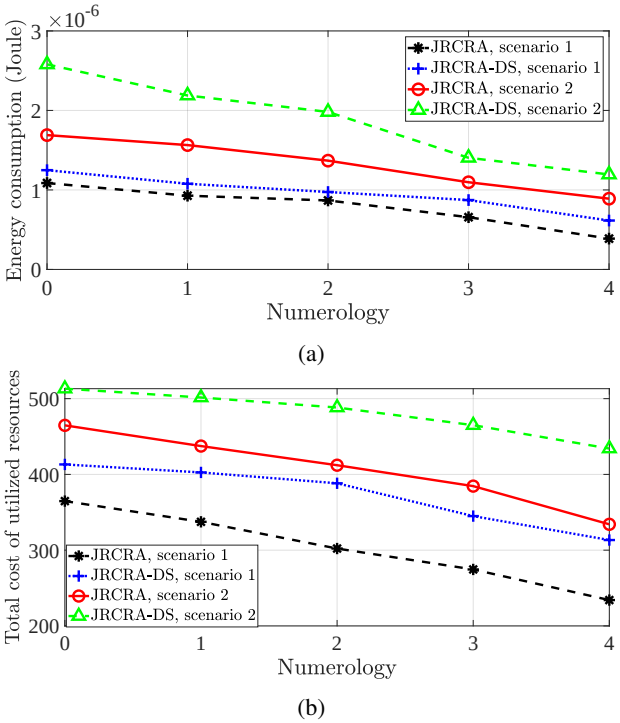

Fig. 6: (a) Energy consumption and (b) total cost of utilized resources versus different numerologies.

URLLC users can attain a higher data rate in accordance with (2). Furthermore, when the E2E delay threshold T_i^{th} is decreased, URLLC users are forced to increase their data rate to meet their E2E delay constraint leading to increasing energy consumption. In Fig. 5b, we illustrate the correlation between the cost of utilized resources and the parameters of the E2E delay threshold and decoding error probability. As the data rate of URLLC users is reduced due to a decrease in decoding error probability, they are forced to occupy a larger number of sub-channels to improve their data rate. This leads to an increase in the cost of utilized resources. Additionally, when the E2E delay threshold diminishes, URLLC users necessitate a greater allocation of resources in both RAN and CN to fulfill the E2E delay constraint. Consequently, the cost of utilized resources increases proportionally as the threshold delay decreases.

C. Applicability of JRCRA to Different Numerologies

In 5G New Radio (NR), various numerologies are supported, representing different subcarrier spacings such as 15, 30, 60, 120, and 240 KHz. Additionally, each resource block also referred to as sub-channel is fixed at 12 subcarriers. Considering these factors, the bandwidth of each SC can indeed vary and be equal to 180, 360, 720, 1440, or 2880 KHz [54]. To illustrate the applicability of our proposed JRCRA algorithm in such scenarios, we have conducted simulations considering two different scenarios with distinct numerologies and varying bandwidths for BSs, which is illustrated in what follows.

In Fig. 6, we present the energy consumption and cost of utilized resources for two distinct scenarios. In the first scenario, we assume that both BSs utilize the entire network bandwidth, which is set at 100MHz. In the second scenario, each BS occupies half of the network bandwidth, resulting in 50MHz of bandwidth allocation per BS. It is worth noting that in both scenarios, we consider different numerologies, represented by values ranging from 0 to 4. The bandwidth of each SC is determined by the formula $W^k = 180 \times 2^{\text{numerology}}(\text{KHz})$, where W^k represents the SC bandwidth.

Fig. 6a illustrates the impact of different numerologies on energy consumption. As can be seen, increasing the numerology results in improved energy consumption. This improvement stems from the increased bandwidth of each SC, which subsequently enhances the data rate for users. Additionally, Fig. 6a demonstrates that allocating orthogonal bandwidth for each BS leads to increased energy consumption. In this case, the available bandwidth for each BS is halved, leading to higher energy consumption compared to scenarios where each BS utilizes the entire network bandwidth.

Moreover, the cost of utilized resources versus different numerologies is shown in Fig. 6b. From Fig. 6b, we can observe that with increasing numerology, the total cost is reduced. The reason is that when the bandwidth of each SC increases, users can reach their QoS requirement, occupying less number of SCs.

D. Impact of Users' Mobility

To investigate the impact of users' mobility on the performance of JRCRA, we consider two distinct scenarios wherein the users move at a constant speed of 5m/s. In the first scenario, it is assumed that the users approach the BS assigned to them, and in the second scenario, the users move away from the BS. To model the mobility of users, we employ the

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2023.3341810

Random Waypoint model, which allows each user to move independently from others, selecting a constant speed and movement direction.

Specifically, Fig. 7 illustrates the impact of users' mobility on the performance of the JRCRA algorithm. The curve labeled as *JRCRA-moving away users* represents the scenario where all users move away from their assigned BS at a speed of $5\mathrm{m/s}$. On the other hand, the curve labeled as *JRCRA-approaching users* represents the scenario where all users approach their assigned BS at the same speed.

In Fig. 7, we consider a time slot of 10 seconds, where each user moves 5 meters per second. As depicted in Fig. 7a, there is an observable increase in energy consumption when users distance themselves from their associated BS. This can be attributed to a decline in path-gain as users move away from their serving BS. Furthermore, as users move further from the BS, they approach the edge of the cell, which leads to increased interference from neighboring cell users. Consequently, it becomes obvious that energy consumption increases with the distance from the base station. Conversely, as users approach their BS, energy consumption shows an improvement due to enhanced path-gain and reduced interference from adjacent cell users.

Moreover, as illustrated in Fig. 7b, the degradation of path-gain and the increasing interference necessitate users who move away from the BS to allocate additional resources to satisfy their QoS demands, resulting in an increased cost of utilized resources. Conversely, as users approach their assigned BS, the enhancement of path-gain and the mitigation of interference cause a reduction in the resources required to meet their QoS, consequently leading to a decreased cost of utilized resources.

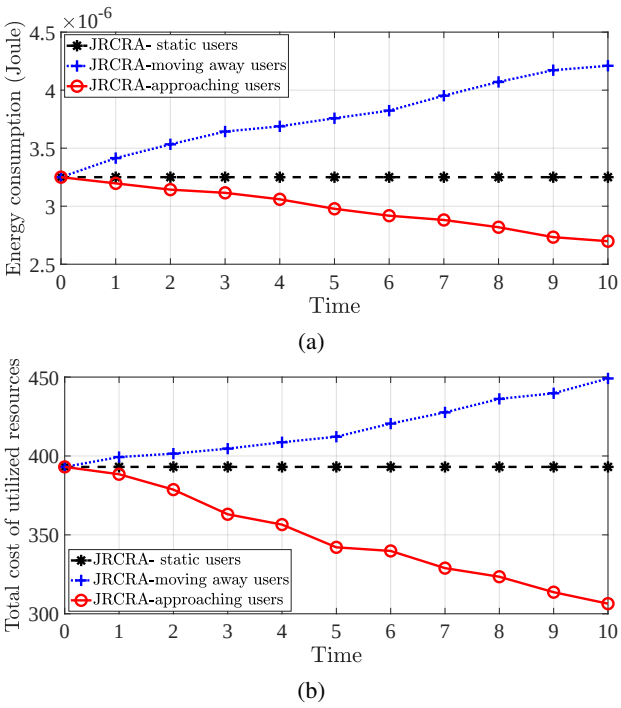

Fig. 7: The impact of users' mobility in terms of (a) energy consumption and (b) total utilized resources' cost.

E. Superiority of JRCRA Algorithm Over Disjoint Case

In Fig. 8, the performance of the JRCRA algorithm is compared with the disjoint case (JRCRA-DS) where problems RRA (18) and CRA (19) are solved disjointly. In Fig. 8, the JRCRA algorithm and JRCRA-DS are compared when the minimum data rate of eMBB users is set to $R_i^{\min} = 1\mathrm{Mbps}$ and $R_i^{\min} = 2\mathrm{Mbps}$. And the E2E tolerable latency for URLLC users is set to $T_i^{\mathrm{th}} = 1\mathrm{ms}$ and $T_i^{\mathrm{th}} = 2\mathrm{ms}$. In JRCRA-DS, inspired by [35] and [36], we consider half of the E2E tolerable latency in RAN and half of the E2E tolerable latency in CN. To clarify, when $T_i^{\mathrm{th}} = 1\mathrm{ms}$, in JRCRA-DS, the tolerable latency of each URLLC user in the RAN and CN is $0.5\mathrm{ms}$. For JRCRA-DS, we first solve RRA problem (18), taking into account the tolerable latency in RAN. Then, we solve problem CRA (19) considering the fixed transmit power and sub-channel allocation. Finally, we sum the value of the objective function obtained by solving two problems RRA (18) and CRA (19). For Fig. 8, we set the number of servers to $|\mathcal{N}| = 20$ and the number of users per each slice to $U_g = 5$.

In the JRCRA algorithm, all the decision variables, including power control and sub-channel allocation in RAN and server and physical links bandwidth allocation in CN, are jointly considered. On the other hand, in the JRCRA-DS case, to solve the CRA problem, fixed power control

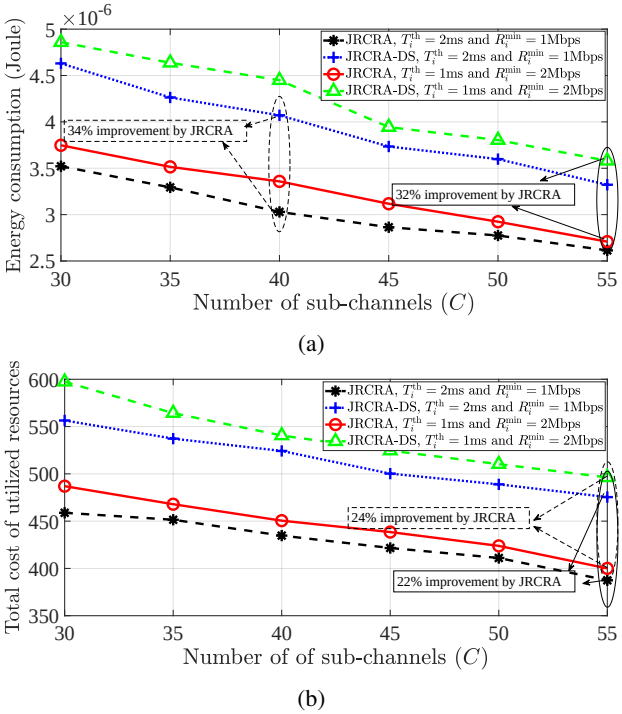

Fig. 8: Comparison of JRCRA and JRCRA-DS in terms of (a) energy consumption and (b) total cost of utilized resources versus the number of SCs.

and sub-channel allocation are considered in the RAN. The performance of JRCRA improves in comparison with the JRCRA-DS due to the joint allocation of decision variables. Furthermore, half of the E2E latency is considered for both RAN and CN in the disjoint case. Therefore, in the disjoint case, the latency constraint may be more strict than that in the joint case. This leads to more resource requirements which results in more energy consumption and cost. Moreover, fixing the power control and sub-channel allocation for solving the CRA problem leads to more strictness for the constraints C6 and C8 in (10) and (12). Consequently, to meet these constraints, users need more resources which results in more cost and energy consumption.

Proposition 3. *The solution space of the disjoint case is a subset of that of joint problem* (17).

Proof. The proof is given in **Appendix C**. ■

As shown in Fig. 8a, by employing the JRCRA algorithm, when $T_i^{\text{th}} = 1\text{ms}$ and $T_i^{\text{th}} = 2\text{ms}$, the total value of energy consumption decrease by 32% and 34%, respectively. Additionally, Fig. 8b demonstrates that the JRCRA algorithm has 24% and 22% improvement on cost of utilized resources, respectively, when $T_i^{\text{th}} = 1\text{ms}$ and $T_i^{\text{th}} = 2\text{ms}$.

F. Comparison of JRCRA Algorithm With an Existing Algorithm

In order to demonstrate the efficacy of the JRCRA algorithm in comparison to existing algorithms, we compare the performance of JRCRA with the algorithm proposed in [33]. To the best of the authors' knowledge and as indicated in Table I, [33] closely aligns with our work.

In [33], the joint slicing of RAN and CN was studied. The objective function of the stated problem in [33] was to maximize the MNO's revenue, which was defined as the difference between the revenue generated from data rates and the cost of resources utilized. In order to ensure the satisfaction of each slice's users, an E2E delay constraint was taken into account, while guaranteeing a reserved data rate constraint for each slice. This was achieved by setting a minimum threshold for the aggregate data rate of the users within each slice. Moreover, the decision variables considered encompassed the sub-channel allocation and power control in RAN as well as the allocation of processing resources and bandwidth of physical links in CN.

In order to ensure a fair comparison between our proposed algorithm and the algorithm introduced in [33], we set the value of α in problem (17) to zero. Consequently, the objective function of problem (17) transforms into minimizing the cost of resources utilized, while the significance of minimizing energy consumption diminishes. Additionally, in order to implement the DRL algorithm proposed by [33], we assign a weight of zero to the data rate revenue. Thus, the objective function defined in [33] is modified to involve the minimization of resource utilization costs. Furthermore, instead of setting a data rate constraint specifically for eMBB users in the stated problem (17), we opt to consider a reserved data rate constraint for both slices, in line with the approach adopted in [33].

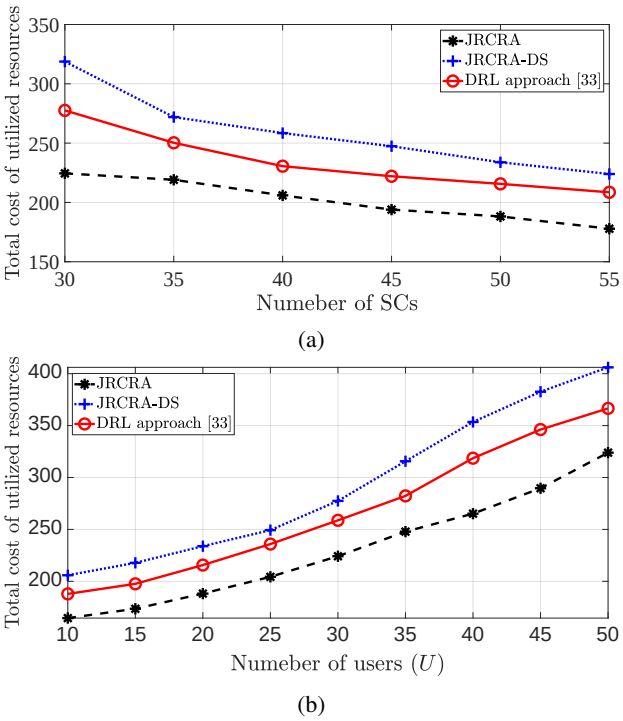

Fig. 9: Comparison of JRCRA with DRL approach proposed in [33] in terms of the total utilized resources' cost versus (a) number of sub-channels and (b) number of total users.

In Fig. 9, a comparison is shown between the JRCRA algorithm and the algorithm outlined in [33]. It is noteworthy that in order to implement the algorithm proposed in [33], the codes made available by the authors on [55] were utilized. Fig. 9a illustrates the cost of using resources versus the number of sub-channels. This figure is generated by considering a total of $U = 20$ users allocated to either eMBB or URLLC slices. Each slice is assumed to have a reserved data rate requirement of 1Mbps, and the maximum threshold delay for eMBB users is set at 4ms, while for URLLC users it is set at 1ms. It is evident that the cost of resource utilization decreases as the number of sub-channels increases, owing to the channel diversity. Furthermore, JRCRA demonstrates a higher level of efficiency compared to the algorithm proposed in [33].

Moreover, Fig. 9b illustrates the cost of resource utilization versus the number of users. The generation of this figure is based on the assumption that there are $C = 50$ sub-channels, and the number of users ranges from 10 to 50, with an equal distribution between the two BSs. From Fig. 9b, it can be observed that the cost of utilized resources increases when the number of users increases since a larger quantity of resources is required to satisfy users' QoS requirement constraints. As depicted in Fig. 9, our JRCRA algorithm holds the potential to serve as a benchmark for the DRL algorithm [33].

G. Optimality Gap Analysis of Proposed JRCRA Algorithm

In this subsection, we present a comparison between the performance of the JRCRA algorithm and the optimal solution attained through the exhaustive search method. Specifically, Fig. 10 illustrates the energy consumption and the cost of

This article has been accepted for publication in IEEE Transactions on Mobile Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMC.2023.3341810

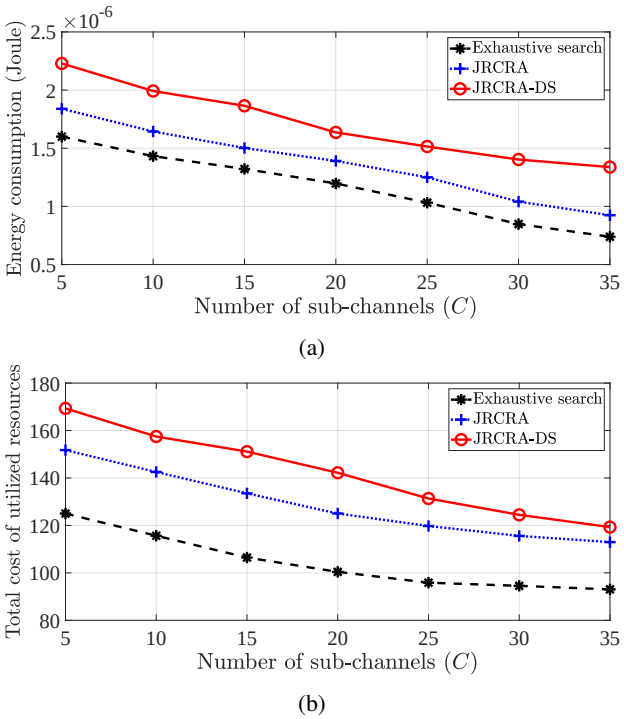

Fig. 10: Comparison of JRCRA and the optimal solution obtained from an exhaustive search in terms of (a) energy consumption and (b) total cost of utilized resources versus the number of SCs.

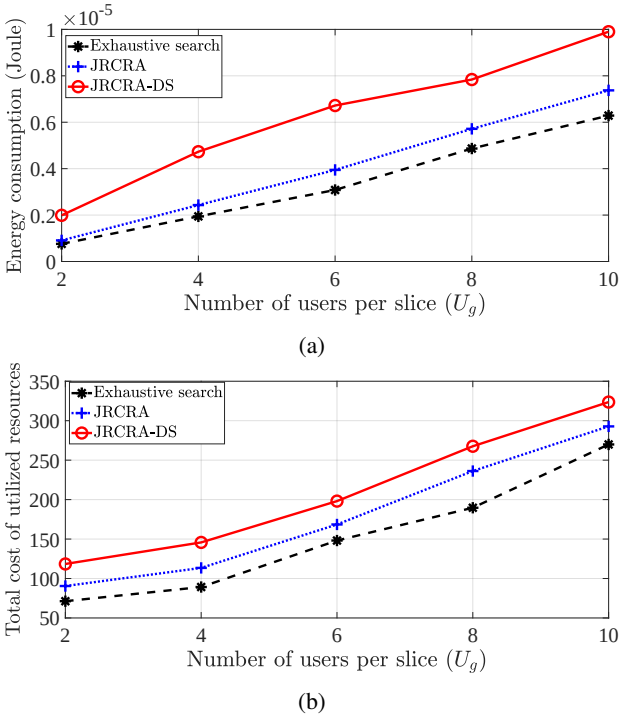

Fig. 11: Comparison of JRCRA and the optimal solution obtained from an exhaustive search in terms of (a) energy consumption and (b) total cost of utilized resources versus the number of users per each slice.

utilized resources versus the number of sub-channels. Also, Fig. 11 shows the energy consumption and the cost of utilized resources with respect to the number of users per slice. To generate Figs. 10 and 11, the QoS requirement for eMBB users is set at $R_i^{\min} = 1\text{Mbps}$, while the QoS requirement for URLLC users is assumed to be $T_i^{\text{th}} = 2\text{ms}$.

In Fig. 10, the number of sub-channels is incremented from 5 to 35, with a step size of 5. As observed in Fig. 10a, the increase in the number of sub-channels leads to a decrease in energy consumption due to enhanced channel diversity. Additionally, the availability of more sub-channels at each BS may reduce interference between users from different BSs thanks to the occupying different SCs in different BSs, resulting in reduced energy consumption. Notably, the JRCRA algorithm performs closely to the optimal response, with an optimality gap ranging from 13% to 25%. Furthermore, Fig. 10b demonstrates that the cost of utilized resources diminishes as the number of sub-channels increases. This is attributed to the possibility of users selecting lower-cost channels. The optimality gap regarding the cost of used resources between the JRCRA algorithm and the optimal solution varies between 21% and 25%.

Fig. 11 depicts the energy consumption and the cost of the utilized resources versus the number of users of each slice, U_g, varying from 2 to 10. From Fig. 11a, it is evident that as the number of users increases, there is a need for more users to meet their QoS requirements. Consequently, this leads to an increase in interference among users, necessitating higher transmission power to achieve their QoS. Consequently, there is a rise in energy consumption. According to Fig. 11a, the JRCRA algorithm exhibits an optimality gap ranging from 17% to 28%. Moreover, Fig. 11b demonstrates that an increase in the number of users results in a higher utilization of resources. This is due to the increased demand for serving a larger number of users, necessitating additional resources in both the RAN and CN. In Fig. 11b, it is observed that the JRCRA algorithm achieves performance close to optimal. In particular, the optimality gap between the JRCRA algorithm and the optimal solution ranges from 8% to 27%.

VII. Conclusion

We have studied the end-to-end (E2E) network slicing problem for the coexistence of eMBB and URLLC services by jointly slicing the radio access network (RAN) and the core network (CN). The objective has been to minimize the E2E energy consumption and the cost of utilized resources while satisfying the minimum data rate and E2E tolerable latency for eMBB users as well as E2E tolerable latency for URLLC users. To address this problem, we have decomposed it into two sub-problems, namely, the RRA and the CRA sub-problems. Due to the binary nature of sub-channel allocation, server, and physical link allocation variables, both of the RRA and CRA problems are non-convex. To deal with this difficulty, we have first relaxed the binary variables by introducing the penalty functions. Also, to handle the non-convexity of the RRA and CRA problems, we have applied the majorization-minimization method. To show the efficacy of our proposed

joint radio and core resource allocation (JRCRA) algorithm, we have evaluated it via simulations. Also, we have compared it with a disjoint solution where the RAN and CN resources are assigned to users separately. The simulation results have shown that our proposed JRCRA algorithm improves energy consumption by 34% and cost by 24% compared to the disjoint case. Such improvements result from the fact that the joint slicing of RAN and CN leads to flexible division of users' E2E tolerable latency between RAN and CN, whereas if resources in RAN and CN are allocated separately, a predefined part of the E2E tolerable latency should be considered for each of the RAN and the CN. Furthermore, simulation results confirm that JRCRA reaches a close performance to the exhaustive search method. A possible future extension of the work would be to consider radio transmissions in the higher frequency bands (e.g., mmWave bands) along with non-orthogonal multiple access in the RAN. Moreover, the optimization framework can be enhanced to include the effect of dynamic network conditions.

REFERENCES

[1] S. K. Taskou, M. Rasti, and P. H. J. Nardelli, "Minimizing Energy Consumption for End-to-End Slicing in 5G Wireless Networks and Beyond," *2022 IEEE Wireless Communications and Networking Conference (WCNC), Austin, TX, USA*, 2022, pp. 860-865.

[2] R. A. Addad, M. Bagaa, T. Taleb, D. L. C. Dutra, and H. Flinck, "Optimization Model for Cross-Domain Network Slices in 5G Networks," *IEEE Transactions on Mobile Computing*, vol. 19, no. 5, pp. 1156-1169, May 2020.

[3] M. Rasti, S. K. Taskou, H. Tabassum and E. Hossain, "Evolution Toward 6G Multi-band Wireless Networks: A Resource Management Perspective," *IEEE Wireless Communications*, early access, 2022.

[4] N. Zhang, Y. Liu, H. Farmanbar, T. Chang, M. Hong, and Z. Luo, "Network Slicing for Service-Oriented Networks Under Resource Constraints," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 11, pp. 2512-2521, 2017.

[5] 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.

[6] "5GMF White Paper- 5G Mobile Communications Systems for 2020 and beyond" V 1.1, July 2016. [Online]. Available: https://5gmf.jp/wp/wp-content/uploads/2017/10/5GMF-White-Paper-v1_1-All.pdf

[7] 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.

[8] 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.

[9] J. Wu, Y. Zhang, M. Zukerman and E. K. Yung, "Energy-Efficient Base-Stations Sleep-Mode Techniques in Green Cellular Networks: A Survey," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 2, pp. 803-826, Secondquarter 2015.

[10] 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," *IEEE Transactions on Wireless Communications*, vol. 20, no. 7, pp. 4585-4600, July 2021.

[11] 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.

[12] M. Almekhlafi, M. A. Arfaoui, M. Elhattab, C. Assi, and A. Ghrayeb, "Joint Resource Allocation and Phase Shift Optimization for RIS-Aided eMBB/URLLC Traffic Multiplexing," *IEEE Transactions on Communications*, vol. 70, no. 2, pp. 1304-1319, Feb. 2022.

[13] A. K. Bairagi *et al*., "Coexistence Mechanism Between eMBB and uRLLC in 5G Wireless Networks," *IEEE Transactions on Communications*, vol. 69, no. 3, pp. 1736-1749, March 2021.

[14] H. Xiang, S. Yan, and M. Peng, "A Realization of Fog-RAN Slicing via Deep Reinforcement Learning," *IEEE Transactions on Wireless Communications*, vol. 19, no. 4, pp. 2515-2527, April 2020.

[15] A. Filali, Z. Mlika, S. Cherkaoui and A. Kobbane, "Dynamic SDN-based Radio Access Network Slicing with Deep Reinforcement Learning for URLLC and eMBB Services," *IEEE Transactions on Network Science and Engineering*, early access, 2022.

[16] Y. T. Woldeyohannes, A. Mohammadkhan, K. K. Ramakrishnan, and Y. Jiang, "ClusPR: Balancing Multiple Objectives at Scale for NFV Resource Allocation," *IEEE Transactions on Network and Service Management*, vol. 15, no. 4, pp. 1307-1321, 2018.

[17] 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*, vol. 9, no. 4, pp. 1479-1492, Oct.-Dec. 2021.

[18] S. Yang, F. Li, R. Yahyapour and X. Fu, "Delay-Sensitive and Availability-Aware Virtual Network Function Scheduling for NFV," *IEEE Transactions on Services Computing*, vol. 15, no. 1, pp. 188-201, Jan.-Feb. 2022.

[19] H. Hawilo, M. Jammal, and A. Shami, "Network Function Virtualization-Aware Orchestrator for Service Function Chaining Placement in the Cloud," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 3, pp. 643-655, March 2019.

[20] G. Sun, Z. Xu, H. Yu, X. Chen, V. Chang, and A. V. Vasilakos, "Low-Latency and Resource-Efficient Service Function Chaining Orchestration in Network Function Virtualization," *IEEE Internet of Things Journal*, vol. 7, no. 7, pp. 5760-5772, July 2020.

[21] W. -K. Chen, Y. -F. Liu, A. De Domenico, Z. -Q. Luo, and Y. -H. Dai, "Optimal Network Slicing for Service-Oriented Networks With Flexible Routing and Guaranteed E2E Latency," *IEEE Transactions on Network and Service Management*, vol. 18, no. 4, pp. 4337-4352, Dec. 2021.

[22] M. Huang, W. Liang, Y. Ma, and S. Guo, "Maximizing Throughput of Delay-Sensitive NFV-Enabled Request Admissions via Virtualized Network Function Placement," *IEEE Transactions on Cloud Computing*, vol. 9, no. 4, pp. 1535-1548, Oct.-Dec. 2021.

[23] S. K. 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.

[24] M. Masoudi, O. T. Demir, J. Zander, and C. Cavdar, "Energy-Optimal End-to-End Network Slicing in Cloud-Based Architecture," *IEEE Open Journal of the Communications Society*, vol. 3, pp. 574-592, 2022.

[25] A. A. Khan, M. Abolhasan, W. Ni, J. Lipman, and A. Jamalipour, "An End-to-End (E2E) Network Slicing Framework for 5G Vehicular Ad-Hoc Networks," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 7, pp. 7103-7112, July 2021.

[26] P. Doanis, T. Giannakas, and T. Spyropoulos, "Scalable end-to-end slice embedding and reconfiguration based on independent DQN agents," *GLOBECOM 2022 - 2022 IEEE Global Communications Conference, Rio de Janeiro, Brazil*, 2022, pp. 3429-3434.

[27] X. Li, R. Ni, J. Chen, Y. Lyu, Z. Rong, and R. Du, "End-to-End Network Slicing in Radio Access Network, Transport Network and Core Network Domains," *IEEE Access*, vol. 8, pp. 29525-29537, 2020.

[28] Y. -J. Liu, G. Feng, Y. Sun, S. Qin, and Y. -C. Liang, "Device Association for RAN Slicing Based on Hybrid Federated Deep Reinforcement Learning," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 12, pp. 15731-15745, Dec. 2020.

[29] T. Li, X. Zhu, and X. Liu, "An End-to-End Network Slicing Algorithm Based on Deep Q-Learning for 5G Network," *IEEE Access*, vol. 8, pp. 122229-122240, 2020.

[30] P. I. Nikolaidis and J. S. Baras, "A Fast and Scalable Resource Allocation Scheme for End-to-End Network Slices," *2021 IEEE Global Communications Conference (GLOBECOM), Madrid, Spain*, 2021, pp. 1-6.

[31] D. Harutyunyan, R. Fedrizzi, N. Shahriar, R. Boutaba and R. Riggio, "Orchestrating End-to-end Slices in 5G Networks," *2019 15th International Conference on Network and Service Management (CNSM), Halifax, NS, Canada*, 2019, pp. 1-9.

[32] J. Liu, B. Zhao, M. Shao, Q. Yang, and G. Simon, "Provisioning Optimization for Determining and Embedding 5G End-to-End Information Centric Network Slice," *IEEE Transactions on Network and Service Management*, vol. 18, no. 1, pp. 273-285, March 2021.

[33] A. Gharehgoli, A. Nouruzi, N. Mokari, P. Azmi, M. R. Javan, and E. A. Jorswieck, "AI-based Resource Allocation in End-to-End Network Slicing under Demand and CSI Uncertainties," *IEEE Transactions on Network and Service Management*, early access, 2023.

[34] 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.
[35] White Paper, "Cellular IoT in the 5G era," *Ericsson*, 2020.
[36] White Paper, "The impact of latency on application performance," *Nokia Siemens Networks*, 2009.
[37] C. She *et al.*, "A tutorial on ultrareliable and low-latency communications in 6G: Integrating domain knowledge into deep learning," *Proceedings of the IEEE*, vol. 109, no. 3, pp. 204-246, March 2021.
[38] K. Chen, Y. Wang, J. Zhao, X. Wang, and Z. Fei, "URLLC-Oriented Joint Power Control and Resource Allocation in UAV-Assisted Networks," *IEEE Internet of Things Journal*, vol. 8, no. 12, pp. 10103-10116, June 2021.
[39] Y. Ma, W. Liang, Z. Xu, and S. Guo, "Profit Maximization for Admitting Requests with Network Function Services in Distributed Clouds," *IEEE Transactions on Parallel and Distributed Systems*, vol. 30, no. 5, pp. 1143-1157, May 2019.
[40] 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.
[41] 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.
[42] H. A. Alameddine, M. H. K. Tushar, and C. Assi, "Scheduling of Low Latency Services in Softwarized Networks," *IEEE Transactions on Cloud Computing*, vol. 9, no. 3, pp. 1220-1235, July-Sept. 2021.
[43] D. P. Bertsekas, R. G. Gallager, and P. Humblet, "Data networks," *Prentice-Hall International New Jersey*, 1992, vol. 2.
[44] J. Cho, Y. Wang, I. Chen, K. S. Chan, and A. Swami, "A Survey on Modeling and Optimizing Multi-Objective Systems," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 3, pp. 1867-1901, thirdquarter 2017.
[45] Y. -F. Liu and Y. -H. Dai, "On the Complexity of Joint Subcarrier and Power Allocation for Multi-User OFDMA Systems," *IEEE Transactions on Signal Processing*, vol. 62, no. 3, pp. 583-596, Feb, 2014.
[46] F. Bari, S. R. Chowdhury, R. Ahmed, R. Boutaba, and O. C. M. B. Duarte, "Orchestrating Virtualized Network Functions," *IEEE Transactions on Network and Service Management*, vol. 13, no. 4, pp. 725-739, Dec. 2016.
[47] Y. Wang, X. Tao, X. Zhang, P. Zhang, and Y. T. Hou, "Cooperative Task Offloading in Three-tier Mobile Computing Networks: An ADMM Framework," *IEEE Transactions on Vehicular Technology*, 2019.
[48] E. Che, H. D. Tuan, and H. H. Nguyen, "Joint Optimization of Cooperative Beamforming and Relay Assignment in Multi-User Wireless Relay Networks," *IEEE Transactions on Wireless Communications*, vol. 13, no. 10, pp. 5481-5495, 2014.
[49] 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, Feb.1, 2017.
[50] 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.
[51] S. Boyd and L. Vandenberghe, "Convex Optimization," *Cambridge University Press*, 2004.
[52] M. Grant and S. Boyd, "CVX: Matlab Software for Disciplined Convex Programming, version 2.1," [Online] http://cvxr.com/cvx, Mar. 2014.
[53] 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.
[54] https://www.etsi.org/deliver/etsi_ts/138200_138299/138211/16.02.00_60/ts_138211v160200p.pdf
[55] https://ieee-dataport.org/documents/codes-paper-ai-based-resource-allocation-end-end-network-slicing-under-demand-and-csi
[56] https://www.itu.int/en/ITU-R/study-groups/rsg5/rwp5d/imt-2020/Documents/S01-1_Requirements\%20for\%20IMT-2020_Rev.pdf

Shiva Kazemi Taskou is a Research Assistant at the Department of Computer Engineering, Amirkabir University of Technology, Tehran, Iran. She received her PhD and M.Sc. degree from Amirkabir University of Technology, in 2023 and 2017, respectively. Her current research interests include resource management in wireless networks, wireless network virtualization, machine learning, deep learning, and optimization.

Mehdi Rasti (S'08-M'11-SM'21) is currently an Associate Professor with the Centre for Wireless Communications, University of Oulu, Finland. From 2012 to 2022, he was with the Department of Computer Engineering, Amirkabir University of Technology, Tehran. He received his B.Sc. degree from Shiraz University, Shiraz, Iran, and the M.Sc. and Ph.D. degrees both from Tarbiat Modares University, Tehran, Iran, all in Electrical Engineering in 2001, 2003 and 2009, respectively. His current research interests include radio resource allocation in IoT, Beyond 5G and 6G wireless networks.

Ekram Hossain (F'15) is a Professor in the Department of Electrical and Computer Engineering at University of Manitoba, Canada (http://home.cc.umanitoba.ca/∼hossaina). He is a Member (Class of 2016) of the College of the Royal Society of Canada, a Fellow of the Canadian Academy of Engineering, and a Fellow of the Engineering Institute of Canada. Dr. Hossain's current research interests include design, analysis, and optimization of wireless networks with emphasis on next-generation (xG) cellular networks. He was elevated to an IEEE Fellow "for contributions to spectrum management and resource allocation in cognitive and cellular radio networks. He received the 2017 IEEE ComSoc TCGCC (Technical Committee on Green Communications & Computing) Distinguished Technical Achievement Recognition Award "for outstanding technical leadership and achievement in green wireless communications and networking. He was listed as a Clarivate Analytics Highly Cited Researcher in Computer Science in 2017-2022.