Energy Cost Minimization by Joint Radio and NFV Resource Allocation: E2E QoS Framework

Abulfazl Zakeri, Narges Gholipoor, Mohammad Reza Javan, Senior Member, IEEE, Nader Mokari, Senior Member, IEEE, and Eduard A. Jorswieck, Senior Member, IEEE

Abstract

In this paper, we propose an end to end radio and network function virtualization (NFV) resource allocation for next generation networks providing different types of services with different requirements in terms of latency and data rate. We consider both the access and core parts of the network, and formulate a novel optimization problem whose aim is to perform the radio resource allocation jointly with virtual network function (VNF) embedding, scheduling, and resource allocation such that the network cost, defined as the consumed energy and the number of utilized network servers, is minimized. The proposed optimization problem is non-convex, NP-hard, and mathematically intractable, and hence, we adopt the alternative search method (ASM) to decouple the main problem into low complex sub-problems. Moreover, we propose a novel heuristic algorithm for NFV resource allocation by proposing a novel admission control algorithm. Then, we compare the performance of the proposed algorithm with a greedy-based solution in terms of the acceptance ratio and the number of active servers. Our simulation...
results show that the proposed heuristic algorithm outperforms the conventional ones by approximately 8%.

Index Terms— Resource allocation, optimization, network function virtualization (NFV), virtual network function (VNF), scheduling, 5G.

1. INTRODUCTION

A. Motivations and Related Works

The fifth generation of wireless cellular networks (5G) provides a wide range of services with various requirements that should be guaranteed in the network [1], [2]. In the traditional network, in order to provide a new service in these networks, it is necessary for each operator to purchase the dedicated hardware and deploy it on the network [3]. On the other hand, as operators are under pressure to keep up with capacity demands and launch differentiated offerings in a highly competitive service and market. Recently, network function virtualization (NFV) is introduced as an interesting technology to reduce network cost and time to market by relying on the virtualization of the network equipment such as servers, routers, storage, and switches [1], [4]–[6]. In NFV, general purpose servers are used on which virtual network functions (VNFs) are run which reduces the capital expenditures (CAPEX) and operational expenditures (OPEX) [7].

In NFV, each network service (NS) consists of a set of VNFs\(^1\) that are executed in a particular order on the network servers. Deploying VNFs on network servers can be performed dynamically depending on the user’s geographical location and network conditions [1], [5], [8]. To implement an NS with acceptable performance, it is necessary to map its VNFs to appropriate servers and schedule them in an efficient manner [5], [7]. The resource allocation (RA) in NFV-based network consists of three stages: 1) VNF chain composition in which the execution order of the VNFs is determined for an NS, 2) VNFs embedding (placement) in which VNFs are mapped to the network servers, and 3) VNFs scheduling in which the execution order of the VNFs embedded

\(^1\)Examples of VNFs includes firewall, deep packet inspection, transcoding, and load balancing.
in network servers is determined [5], [7]. In addition to the core network, the access network plays a key role in the quality of service (QoS) provisioning [9]–[11]. The radio resources in the access network, such as transmit power and bandwidth, should be efficiently allocated to users in order to meet the service requirements of the users. This issue is more important in the network when the requested services are sensitive to latency [10].

A dynamic service function chain deployment is proposed in [12] in which the authors consider a trade-off between resource consumption and operational overhead. In [13], NF placement in the network is studied. Moreover, its impact on network performance with the aim of minimizing the cost of having virtual machines (VMs)\(^2\) and the cost of steering traffic into servers are investigated. Service function chain (SFC) placement in the cloud-based network with the aim of minimizing end-to-end (E2E) latency of SFCs and enhancing QoS is investigated in [14]. An automated decentralized method for online placement and optimization of VMs in NFV-based network is proposed in [15]. In [16], VNF embedding with the aim of minimizing physical machine and taking into consideration users’ SFC requests and factors such as basic resource consumption and time-varying workload are studied.

In [1], an online scheduling and embedding algorithm by considering the capacity of available buffers and the processing time of each VNF is proposed for NFV. The authors propose a set of greedy algorithms and tabu search algorithm for mapping and scheduling. Moreover, cost, revenue, and acceptance ratio of these algorithms are compared together. VNF placement in a network with several mobile virtual network operators (MVNOs) are investigated in [17] in which the slice scheduling mechanism is introduced in order to isolate the traffic flow of MVNOs with optimizing VNF placement based on the available radio resources. Joint VNF placement and admission control (AC) with maximizing networks provider revenue in terms of bandwidth and capacity are studied in [18]. The authors in [19], propose an RA algorithm which integrates the placement and scheduling of VNFs together. In [20], a VNF scheduling problem is investigated and joint VNF scheduling and traffic steering is formulated as a mixed integer linear program.

\(^2\)In this paper, VM, node, and server have the same meaning.
As be concluded the current works only focus on NFV RA without considering radio RA effect on the QoS and performance of the network. Moreover, no existed studies work to introduce a unified framework that comprises all mentioned NFV RA stages in a closed form by adopting AC. On the other hand, resource management in both core and radio has pivotal role on E2E QoS guaranteeing and services acceptance ratio which is verified by the AC algorithms.

B. Our Main Contributions

Obviously, in a real network and practical scenarios, QoS is the E2E concept and depends on radio access and core networks. In fact, guaranteeing QoS for different applications refers to ensure all of requirements, such as data rate by all parts of the network. These reasons, motivate us to propose a framework which radio and NFV RA is considered for E2E service provisioning where a new AC mechanism is devised for the service requests. The main contributions of this paper can be summarized as follows:

- In this paper, in order to guaranteeing E2E QoS by utilizing resources, efficiently, we propose a novel E2E QoS-aware framework by consideration of the radio and NFV RA that has not been considered in the literature.

- More importantly, we introduce a new approach for VNF scheduling with considering the network service latency. We introduce a new scheduling variable using which the latency of each VNF is obtained by calculating the processing and waiting time of all VNFs scheduled before it. This means that the time each NS is finished can be calculated as sum of the waiting time and processing time elapsed from the packet entrance to the packet receiving by the destination [21]. On the other hand, we consider a maximum tolerable latency for each packet of different services which should be ensured by the network. We propose a novel efficient and low complexity algorithm based on minimizing the number of active VMs (servers) and guaranteeing the requested service QoS requirements.

- We formulate a new optimization problem for radio and NFV RA with the aim of minimizing cost function in terms of radio and NFV resources. In the proposed optimization problem, subcarrier assignment, power allocation, VNF embedding, scheduling, ordering, and server
utilization are optimization variables. Our main aim is to minimize the network cost in terms of the transmit power and the number of active nodes while guaranteeing the service QoS metrics.

- To overcome the infeasibility issue in the solution of the proposed optimization problem, we propose a new elastication method and a novel AC method to reject some users and guarantee the other users requested service requirements. Based on the proposed AC, the user which has the most effect on the infeasibility on optimization problem i.e., needs more resources to guarantee its QoS is found and its service is rejected.
- We prove the convergence of the proposed algorithm and analyze its computational complexity.
- We provide numerical results for the performance evaluation of the proposed problem and algorithm for different network configurations and greedy-based algorithm. Our simulation results show our proposed algorithm outperforms greedy-based by approximately 8% for same computational complexity.

C. Paper Organization

The rest of the paper is organized as follows. In Section II, the system model is explained. In Section II-E, the problem formulation is presented. The problem solution is presented in Section III. In Section IV, the computational complexity and convergence of the proposed algorithm are investigated. The simulation results are presented in Section V. Finally, in Section VI, the paper is concluded.

Notations: Vector and matrices are indicated by bold lower-case and upper-case characters, respectively. $\| \cdot \|$ and $\| \cdot \|_p$ represent the absolute value and $p$-norm, respectively. $\mathcal{A}$ denotes set $\{1, \ldots, A\}$, $\mathcal{A}(i)$ is $i$-th element of set $\mathcal{A}$, and $\mathbb{R}^n$ is the set of $n$ dimension real numbers. Moreover, $U_{[a, b]}$ denotes the uniform distribution in interval $a$ and $b$. 
II. SYSTEM MODEL AND PROBLEM FORMULATION

A. Radio RA Parameters

We consider a single-cell network with $U$ users whose set is $\mathcal{U} = \{1, \ldots, U\}$ and $K$ subcarriers whose set is $\mathcal{K} = \{1, \ldots, K\}$. We define the subcarrier assignment variable $\rho_{ku}^k$ with $\rho_{ku}^k = 1$ if subcarrier $k$ is allocated to user $u$ and otherwise $\rho_{ku}^k = 0$. We assume orthogonal frequency division multiple access (OFDMA) as the transmission technology in which each subcarrier is assigned at most to one user. To consider this, the following constraint is considered:

$$\sum_{u \in \mathcal{U}} \rho_{ku}^k \leq 1, \forall k \in \mathcal{K}. \tag{1}$$

Let $h_{ku}^k$ be the channel coefficient between user $u$ and the BS on subcarrier $k$, $p_{ku}^k$ be the transmit power from the BS to user $u$ on subcarrier $k$, and $\sigma_{u}^k$ be the power of additive white Gaussian noise (AWGN)\(^3\) at user $u$ on subcarrier $k$. The received signal to noise ratio (SNR) of user $u$ on subcarrier $k$ is $\gamma_{u}^k = \frac{p_{ku}^k h_{ku}^k}{\sigma_{u}^k}$, and the achievable rate of user $u$ on subcarrier $k$ is given by

$$r_{u}^k = \rho_{ku}^k \log(1 + \gamma_{u}^k), \forall u \in \mathcal{U}, k \in \mathcal{K}. \tag{2}$$

Hence, the total achievable rate of user $u$ is given by $R_{u} = \sum_{k \in \mathcal{K}} r_{u}^k, \forall u \in \mathcal{U}$. The following constraint states the power limitation of BS:

$$\sum_{k \in \mathcal{K}} \sum_{u \in \mathcal{U}} \rho_{ku}^k p_{ku}^k \leq P_{\text{max}},$$

where $P_{\text{max}}$ is the maximum transmit power of BS.

B. NFV RA Framework

In this subsection, we explain how the generated traffic of each user is handled in the network by performing different NFs in the requested user’s NS\(^4\) on the different servers/physical nodes

\(^3\)In this paper, we assume that an AWGN interfering source (IS) interferes at the BS and all users on each subcarrier. We consider a single cell with a BS, in a scenario with many cells and no coordination between BSs, the inter-cell interference distribution converges to a Gaussian and can be integrated into the interferences of other cells which can be modeled by the IS\(^22\).

\(^4\)Defined by European Telecommunications Standards Institute (ETSI) as the composition of Network Function(s) and/or Network Service(s), defined by its functional and behavioral specification\(^23\).
by leveraging NFV\(^5\). In this regard, we consider NFV RA that consists of a new approach for the embedding and scheduling phases. In the embedding phase, we map each NF on the server that is capable to run that NF. Note that we do not consider mapping virtual links on the physical links and leave it as an interesting future work as [1], [25].

We consider \(S\) communication service (CS)\(^6\) types whose set is \(S = \{1, 2, \ldots, S\}\) and \(M\) NFs whose set is \(\mathcal{F} = \{f_m | m = 1, \ldots, M\}\). The considered parameters of the paper are stated in Table II-B. Each CS \(s\) is defined by the tuple \((\text{ingress node}, \text{egress node}, \Omega_s, D_{\text{max}}^s, R_{\text{min}}^s)\) where \(\Omega_s\) is the set of NFs which constructs NS \(s\) defined by \(\Omega_s = \{f_m^s | m \in \{1, \ldots, M\}\}\), \(D_{\text{max}}^s\) is the latency constraint for each packet of NS \(s\) and \(R_{\text{min}}^s\) is the minimum required data rate of NS \(s\). We assume that each user can request at most one CS at a time.

It is worth noting that some of NFs have some association and precedence over some others, for instance, the NF decryption is performed after encryption. We consider a set of VMs denoted by

---

**TABLE I**

**Configuration parameters**

| Notation | Definition |
|----------|------------|
| \(\mathcal{U}\) | Set of users |
| \(\mathcal{N}\) | Set of VMs |
| \(\mathcal{F}\) | Set of NFs |
| \(\mathcal{S}\) | Set of NSs |
| \(P_{\text{max}}\) | Maximum transmit power of the BS |
| \(\rho_{ku}\) | Assignment of subcarrier \(k\) to user \(u\) |
| \(p_{ku}\) | Transmit power of user \(u\) on subcarrier \(k\) |
| \(h_{ku}\) | Channel coefficient between user \(u\) and the BS on subcarrier \(k\) |
| \(\gamma_{ku}\) | SNR of user \(u\) on subcarrier \(k\) |
| \(r_{ku}\) | Achieved rate of user \(u\) on subcarrier \(k\) |
| \(y_u\) | Packet size of the requested service of user \(u\) |
| \(\alpha^{f_m}_{s}, \psi^{f_m}_{s}\) | Processing and buffering demand of NF \(f_m\) in NS \(s\), respectively |
| \(\tilde{\tau}_{nm}^{f_m}\) | Processing latency of NF \(f_m\) on server \(n\) in NS \(s\) |
| \(L_{n}, \Upsilon_n\) | Processing and buffering capacity of VM \(n\), respectively |
| \(\eta_n\) | Server activation indicator |
| \(\beta_{u,n}^{f_m}\) | Server mapping between NF \(f_m\) of service \(s\), user \(u\) and node \(n\) |
| \(t_{u,n}^{f_m}\) | Starting time of NF \(f_m\) of service \(s\) which is requested by user \(u\) at node \(n\) |
| \(x_{u,u',f_m,f_m'}\) | Ordering indicator between NF \(f_m\) of service \(s\) user \(u\) and \(f_m'\) of service \(s'\) user \(u'\)

---

\(^5\)Standardized by ETSI organization for 5G and beyond in the telecommunication [24].

\(^6\)In this paper, the NS and CS are paired together. That means each CS \(s\) has a NS with corresponding NFs that is denoted by set \(\Omega_s\). Note that CS is defined by the 3rd generation partnership project (3GPP) technical specification 28.530 [26].
\( \mathcal{N} = \{1, \ldots, N\} \) in the network each of which has computing and storage resources. We assume that each server can process at most one function at a time [1]. To improve energy efficiency (EE) in our proposed system, we introduce a new variable \( \eta_n \) to determine the active nodes which is defined as

\[
\eta_n = \begin{cases} 
1, & \text{Node } n \text{ is active}, \\
0, & \text{Otherwise.}
\end{cases}
\]

We consider a generalized model for resource sharing of VMs that is introduced in [1]. Therefore, we introduce a binary variable \( \beta_{f_{sm}u,n} \) which denotes that NF \( f_{sm}^u \) for user \( u \) in NS \( s \) is executed at node \( n \), and is defined as

\[
\beta_{f_{sm}u,n} = \begin{cases} 
1, & \text{NF } f_{sm}^u \text{ for user } u \text{ in NS } s \text{ is executed at server } n. \\
0, & \text{Otherwise.}
\end{cases}
\]

When \( \beta_{f_{sm}u,n} \) is set to 1 i.e., \( f_{sm}^u \) in the requested NS \( s \) for user \( u \) is mapped on server \( n \), and this server should be active, i.e., \( \eta_n = 1 \). Therefore, we have the following constraint:

\[
\beta_{f_{sm}u,n} \leq \eta_n, \forall n \in \mathcal{N}, \forall u \in \mathcal{U}, \forall f_{sm}^u \in \Omega_s, \forall s \in \mathcal{S}. \tag{3}
\]

Each NF of each NS is performed completely at only one server at a time. Therefore, we have

\[
\sum_{n \in \mathcal{N}} \beta_{f_{sm}u,n} \leq 1, \forall u \in \mathcal{U}, f_{sm}^u \in \Omega_s, s \in \mathcal{S}. \tag{4}
\]

Moreover, we assume that each NF needs a specific number of CPU cycles per bit i.e., \( \alpha f_{sm}^u \) to run on a mapped server. From the physical resources perspective, we assume that each server can provide at most \( L_n \) CPU cycles per unit time and hence, we have the following constraint:

\[
\sum_{u \in \mathcal{U}} \sum_{s \in \mathcal{S}} \sum_{f_{sm}^u \in \Omega_s} y_u \alpha f_{sm}^u \beta_{f_{sm}u,n} \leq L_n, \forall n \in \mathcal{N}, \tag{5}
\]

where \( y_u \) is the packet size of service user \( u \) and here is assumed to be equal to the number of bits generated in a time unit, i.e., \( R_{min}^u \). Hence, the processing time of each function \( f_{sm}^u \) for each bit on server \( n \in \mathcal{N} \), i.e., is as follows:

\[
\tilde{\tau}_{f_{sm}^u}^n = \frac{\alpha f_{sm}^u}{L_n}, \forall n \in \mathcal{N}, f_{sm}^u \in \Omega_s. \tag{6}
\]
Therefore, the total processing latency for each packet with packet size $y_u$ is obtained as

$$\tau_{f,m}^{n} = \tau_{f,m}^{n} y_u, \forall n \in N, f^s_m \in \Omega_s. \quad (7)$$

Additionally, we assume that each NF needs specific buffer size, i.e., $\psi_{f,m}^s$, when it is running on the server. Hence, from the storage resource perspective, we consider that each server has the limited buffer size i.e., $\Upsilon_n$, which leads to the following constraint:

$$\sum_{u \in U} \sum_{s \in S} \sum_{f^s_m \in \Omega_s} (\psi_{f,m}^s + y_u) \beta_{u,n}^{f,m} \leq \Upsilon_n, \forall n \in N. \quad (8)$$

C. Latency Model

In NFV RA, our main aim is to guarantee the service requirement includes maximum tolerable latency for each packet with size $y_u$ of the requested services with minimizing consumption of servers. The total latency that we consider in our system model results from executing NFs and queuing (waiting) time. In the following, we calculate the total latency resulting from scheduling.

Remark 1. In this paper, our main aim is to model and investigate the effect of processing and scheduling latency on the service acceptance and the network cost. Hence, we do not consider the other latency factors such as propagation and transmission latencies in our model. In fact,
our proposed scenario is focused on the intra data centers and not appropriate for the national-wide networks. It is worth noting that the aforementioned latency is coming from the high order distance from the source and application servers. Therefore, these concerned can be treated by exploiting the mobile edge computing (MEC) technology to bring the application servers close to clients. In this regard, we generalize this work for the MEC-enabled networks in future works.

1) Scheduling and Chaining: Each NF should wait until its preceding function is processed before its processing can commence. The processing of NS s ends when its last function is processed. Therefore, the total processing time is the summation of the processing times of the NFs at the various servers. For scheduling of each NF on a server, we need to determine the start time of it. Therefore, we define $t_{f_{sm}u,n}$ which is the start time of running NF $f_{sm}$ of the requested service $s$ for user $u$ on server $n$. Furthermore, we introduce a new variable $x_{u,u'}$, in which, if NF $f_{sm}$ of user $u$ is running after NF $f_{sm'}$ of user $u'$, its value is 1, otherwise is 0. By these definitions, the starting time of each NF can be obtained as follows:

$$
t_{f_{sm}u,n} = \max \left\{ \max_{f_{sm'}u' \in \Omega_s, u' \in U} \left\{ f_{sm} \cdot f_{sm'} \cdot f_{sm'} \cdot (t_{f_{sm'}u' \cdot n} + \tau_{f_{sm'}n}) \right\} \right\},$$

$$= \max_{f_{sm}, f_{sm'} \in \{\Omega_s - f_{sm} \}, n' \in \{N - n\}} \left\{ f_{sm} \cdot f_{sm'} \cdot f_{sm'} \cdot (t_{f_{sm'}u' \cdot n} + \tau_{f_{sm'}n}) \right\},$$

$$\forall f_{sm} \in \Omega_s, f_{sm'} \in \Omega_s, s, s' \in S, \forall n \in N, \forall u \in U. \tag{9}$$

To more clarify, we illustrate the proposed scheduling policy in Fig. 2. The total service chain latency for each user $u$ on the requested service is inferred as follows:

$$D_{u}^{\text{Total}} = \max_{n \in N, f_{hm} \in \Omega_s, s \in S} \left\{ t_{f_{hm}u,n} \cdot f_{hm} + \tau_{f_{hm}n} \beta_{f_{hm}u,n} \right\}, \forall u \in U. \tag{10}$$

D. Cost Model: Objective Function

Our aim in this paper is to minimize the total cost of the network. In this regard, we define cost $\Psi$ as the total amount of radio and NFV resources that are utilized in the network to provide services. In particular, the cost function is given as follows:

$$\Psi = \mu \sum_{u \in U, k \in K} p_{u,k}^k + \nu \sum_{n \in N} \eta_n, \tag{11}$$

where $\mu$ and $\nu$ are constants for scaling and balancing the costs of different resource types.
E. Problem Formulation

Based on these definitions, our aim is to solve the following optimization problem:

$$\min_{P, \rho, T, X, \beta, \eta} \Psi(P, \rho, \eta)$$

S.T:

$$R_u \geq R_u^{\text{min}}, \forall u \in U,$$  \hspace{1cm} (12b)

$$\sum_{u \in U} \rho_u^k \leq 1, \forall k \in K,$$  \hspace{1cm} (12c)

$$\sum_{k \in K} \sum_{u \in U} \rho_u^k P_u \leq P_{\text{max}},$$  \hspace{1cm} (12d)

$$\sum_{u \in U} \sum_{s \in S} \sum_{f_{m}^{s} \in \Omega} y_u \alpha_{f_{m}^{s}} \beta_{f_{m}^{s}}^{u,n} \leq L_n, \forall n \in N,$$  \hspace{1cm} (12e)

$$\sum_{u \in U} \sum_{s \in S} \sum_{f_{m}^{s} \in \Omega} \phi_{f_{m}^{s}} \beta_{f_{m}^{s}}^{u,n} + y_u \sum_{f_{m}^{s} \in \Omega} \beta_{f_{m}^{s}}^{u,n} \leq \Upsilon_n, \forall n \in N,$$ \hspace{1cm} (12f)
\[
t_{u,n} \beta_{u,n} \geq \max \left\{ \max_{f_{m,n} \in \Psi_{n}, u' \in U} \left\{ \frac{f_{m,n}}{x_{u,u'} \beta_{u,n} (t_{u,n'} + \tau_{m'})} \right\}, \right. \\
\left. \max_{f_{m,n} \in \{\Omega_{s} - f_{m,n}\}, n' \in \{N - n\}} \left\{ \frac{f_{m,n}}{x_{u,u'} \beta_{u,n'} (t_{u,n''} + \tau_{m''})} \right\} \right\}
\]
\[
\forall f_{m}^{s} \in \Omega_{s}, f_{m'}^{s'} \in \Omega_{s'}, \forall s, s' \in S, \forall n \in N, \forall u \in U,
\]
\[
D_{u}^{\text{Total}} \leq D_{s}^{\max}, \forall u \in U,
\]
\[
0 \leq p_{u}, \forall u \in U, k \in K,
\]
\[
\beta_{u,n} \leq \eta_{n}, \forall n \in N, \forall u \in U, f_{m}^{s} \in \Omega_{s},
\]
\[
\sum_{n \in N} \beta_{u,n} \leq 1, \forall u \in U, f_{m}^{s} \in \Omega_{s}, s \in S,
\]
\[
\rho_{u}^{k} \in \{0, 1\}, \forall u \in U, k \in K,
\]
\[
\beta_{u,n} \in \{0, 1\}, \forall u \in U, \forall f_{m}^{s} \in \Omega_{s}, \forall s \in S,
\]
\[
x_{u,u'}^{f_{m},f_{m'}} \in \{0, 1\}, \forall u, u' \in U, u \neq u', \forall f_{m}, \forall f_{m'} \in \Omega_{s'},
\]
\[
\eta_{n} \in \{0, 1\}, \forall n \in N,
\]
\[
\text{where } \rho = [\rho_{u}^{k}], \beta = [\beta_{u,n}], T = [t_{u,n}], P = [p_{u}], X = [x_{u,u'}^{f_{m},f_{m'}}], \text{ and } \eta = [\eta_{n}]. \text{ In problem (12),}
\]
\[
\text{constraint (12b) ensures the minimum rate requirement, (12c) guarantees that each subcarrier is assigned to at most one user, (12d) is the power limitation constraint. Moreover, constraint (12g) determines the scheduling principle, constrain (12e) guarantees the processing requirement for each NF with the corresponding packet size to run on the server, and (12f) indicates the storage capacity requirement for both buffering and running NFs. Constraints (12l)-(12o) are for binary variables.}
\]

### III. SOLUTION OF THE PROPOSED PROBLEMS

Optimization problem (12) is non-convex including both mixed binary and continues variables with non-linear and non-convex constraints. Hence, it belongs to the NP-hard and mathematically intractable optimization problem and obtaining an optimal solution is not trivial and leads to
high computational complexity and algorithm run time. Therefore, we cannot apply the common convex optimization methods for solving it.

Without considering NFV RA, the radio RA problem, separately, on the power and subcarrier allocation variables is convex optimization problem, and hence, each of them can be solved efficiently. While NFV RA is non-linear mixed integer programming with large number of variables, i.e., \( T, X, \eta, \beta \). These motivate us to develop a new low complexity heuristic algorithm to solve NFV RA sub-problem that is stated with details in Algorithm 2. However, we investigate our proposed algorithm from the different aspects and compare it with other methods.

To solve the optimization problem (12) in an efficient manner, we utilize alternate search method (ASM). To use ASM, we need to calculate initial values of the optimization variables which should satisfy the corresponding constraints of (12). Since the optimization problem (12) would be infeasible, we propose a novel elasticizing approach by introducing a new elastic variable. Based on this method, the constraints that would make the optimization problem infeasible are changed as follows. Assume that we have constraint \( g(y) \leq 0 \), where \( y \in \mathbb{R}^n \) is the objective variable. We elasticize it by \( g(y) \leq A \), where \( A \geq 0 \) is the objective variable. By applying this method, we solve the following optimization problem:

\[
\min_{P, \rho, T, X, \eta, \beta, A} \Psi(P, \rho, \eta) + WA \tag{13a}
\]

S.T:

\[
R_{u}^{\min} - R_u \leq A, \forall u \in \mathcal{U}, \tag{13b}
\]

\[
\sum_{u \in \mathcal{U}} \sum_{s \in \mathcal{S}} \sum_{f_{m}^{s} \in \Omega_{s}} \psi f_{m}^{s} \beta f_{u,n}^{s} + y_u \sum_{f_{m}^{s} \in \Omega_{s}} \beta f_{u,n}^{s} - \Upsilon_n \leq A, \forall n \in \mathcal{N}, \tag{13c}
\]

\[
\sum_{u \in \mathcal{U}} \sum_{s \in \mathcal{S}} \sum_{f_{m}^{s} \in \Omega_{s}} y_u \alpha f_{m}^{s} \beta f_{u,n}^{s} - L_n \leq A, \forall n \in \mathcal{N}, \tag{13d}
\]

\[
D_{u}^{\text{Total}} - D_{s}^{\text{max}} \leq A, \forall u \in \mathcal{U}, \tag{13e}
\]

\[
A \geq 0, \tag{13f}
\]

(12c), (12d), (12j), (12k), (12g), (12i) – (12o),
where $A$ is the elastic optimization variable and $W$ is a large number, i.e., $W \gg 1$. Note that since $A$ can be any non-negative value, the optimization problem (13) is feasible. By solving the optimization problem (13), the infeasibility of the main optimization problem (12) is determined. Therefore, if the elastic variable $A$ is positive, problem (12) is infeasible. To overcome the infeasibility of problem (12), we introduce a new AC method to reject some services providing rooms for the remaining ones. The block diagram illustrating the main steps of the proposed method to solve the optimization problem (12) which is based on solving the optimization problem (13) and AC is shown in Fig. 3.

**Proposition 1.** Problem (13) is equivalent to problem (12), if we have $A = 0$. That means problem (12) is feasible and based on the proposed AC method, all the requested services are accepted.

The elasticated problem (13) is also non-convex and NP-hard. In this regard, we solve it by dividing it into three sub-problems by utilizing ASM. The first sub-problem is power allocation and elastication, the second one is subcarrier allocation, and the last one is NFV RA. In fact, the first and second sub-problems are the radio RA sub-problem and it is stated in Section III-A. In the NFV RA sub-problem, all the optimization variables are integer and the problem formulation and solution are presented in Section III-B. More details of the proposed iterative solution of optimization problem (13) is stated in Algorithm 6. In the next subsection, we explain the solution of the aforementioned sub-problems.

**A. Radio RA**

The radio RA problem is divided into two sub-problems as follows.

1) **Power Allocation and Elasticated Subproblem:** The power allocation and elasticated sub-problem is presented as follows:

$$
\min_{P, A} \sum_{u \in U} \sum_{k \in K} \rho_u^k p_u^k + WA
$$

S.T.: $(13b) – (13f)$, $(12d)$, $(12i)$.  

(14a)

(14b)
Fig. 3. Flowchart of the proposed solution of main-problem (12).

**Algorithm 1:** Iterative RA for problem (13)

**Input:** $\epsilon_{TH} = 10^{-4}$, $Z_{TH} = 100$ and $z = 0$, $P^{(z)} = \frac{P_{max}}{U \times K}$, $A \gg 1$, $T^{(z)} = t^0$, $X^{(z)} = X^0$, $\beta^{(z)} = \beta^0$, and $\eta^{(z)} = \eta^0$

1. repeat
   2. Obtain subcarrier assignment variable, i.e., $\rho$ by solving (15)
   3. Obtain power allocation variable, i.e., $P$, and elastic variable i.e., $A$, by solving sub-problem (14)
   4. Obtain NFV RA variables, i.e., $T$, $X$, $\eta$, and $\beta$ by solving sub-problem (16) by Algorithm 2
   5. $z = z + 1$
5. until $|(\Psi + WA)^{(z)} - (\Psi + WA)^{(z-1)}| \leq \epsilon_{TH}$ or $z \approx Z_{TH}$

**Output:** $\rho$, $P$, $A$, $T$, $X$, $\eta$, and $\beta$

Sub-problem (14) is convex. Hence, it can be solved efficiently by using interior point method (IPM) with CVX toolbox in MATLAB [27].
2) **Subcarrier Allocation Subproblem:** The subcarrier allocation sub-problem is as follows:

\[
\begin{align*}
\min_{\rho} & \quad \sum_{u \in U} \sum_{k \in K} \rho_u^k P_{u,k}, \\
\text{S.T.} & \quad (13b), (12c), (12d), (12l).
\end{align*}
\]  

We solve sub-problem (15) by using MOSEK in MATLAB toolbox [28].

**B. NFV RA**

The NFV RA sub-problem is as follows:

\[
\begin{align*}
\min_{T,X,\eta,\beta} & \quad \sum_{n \in N} \eta_n, \\
\text{S.T.} & \quad (13c) - (13e), (12g), (12j) - (12o).
\end{align*}
\]

To solve problem (16), we propose a new heuristic algorithm where the functions are mapped and scheduled on the servers whose have the minimum processing latency. Moreover, our proposed algorithm is based on minimizing the number of active servers. To this end, we ascendingly sort the servers by the total processing latency metric. After that, the server with the best rank, i.e., the high available capacity in the sorted list is turned on. Then, we activate another server, if the previously activated servers cannot satisfy the resource demands by NFs or QoS of users is degraded. Based on the algorithm, we ascendingly sort users according to latency requirements and then we start to map and schedule each of NFs on the servers. The details of the proposed NFV RA are stated in Algorithm 2.
Algorithm 2: Proposed heuristic NFV RA algorithm for solving sub-problem (16)

Input: Set system configuration parameters: $S, F, \Omega_s, \alpha^{f_m}, \rho^{f_m}$

1. Sorted in ascending order servers according to the total latency of NFs on these servers and write server’s index in $\tilde{N}$ (e.g., $\tilde{N} = \{1, \ldots, 4\} \rightarrow \tilde{N} = \{3, 4, 1, 2\}$ for $N = 4$

2. Ascending sort all users according to the value of $D_u^{\max}$ and write users’s index in $\tilde{U}$ (e.g., $\tilde{U} = \{2, 4, 1, 3, 5\}$ for $U = 5$)

3. for $\tilde{n} = 1$: $\tilde{N}$ do

4. Add $\tilde{N}(\tilde{n})$ to set $\tilde{N}_{\text{used}}$ (Servers are in this set are activated)

5. $t' = [0]_{1, \tilde{n}}$ (Servers are in this set are activated) & $\eta_{\tilde{n}} = 1$

6. for $\tilde{u} = 1$: $\tilde{U}$ do

7. for $m = 1 : |\Omega_s|$ do

8. if $m \geq 2$ then

9. for $\tilde{n} = 1$: $\tilde{N}_{\text{used}}$ do

10. $t_{u,n}^{f_m} = \max\{t_{u,n}^{f_m-1}, t'(\tilde{n})\} + \tau_{\tilde{n}}^{f_m}$

11. else

12. for $\tilde{n} = 1$: $\tilde{N}_{\text{used}}$ do

13. $t_{u,n}^{f_m} = t'(\tilde{n}) + \tau_{\tilde{n}}^{f_m}$

14. Find server $n$ in set $\tilde{N}_{\text{used}}$ which has lowest $t_{u,n}^{f_m}$

15. $\beta^{f_m}_{u,n} = 1$ & Calculate $\phi^{f_m}_{u} = \sum_{n' \in N} \beta^{f_{m'}}_{u,n'} (t_{u,n'}^{f_{m'}} + \tau_{n'}^{f_{m'}})$

16. for $u' \in U, u' \neq u$ & $f_{m'}^{s'} \in \Omega_{u',s'}$ do

17. if $\beta^{f_{m'}}_{u',n'} = 1$ & $\phi^{f_{m'}}_{u'} \geq t_{u,n'}^{f_{m'}}$ then

18. $x_{u,n}^{f_m} = x_{u',n'}^{f_{m'}} = 1$

19. Calculate $t_{u,n}^{f_m}$ according to constraint (9)

20. $t'(n) = t_{u,n}^{f_m}$

21. if (13c)-(13e) satisfy then

22. Break

23. else

24. Return to line 5

Output: $T, X, \eta, \beta$

C. Admission Control

Our proposed AC is based on the value of elastic variable of problem (13). Whereas, if $A$ is non-zero, the original problem (12) is infeasible. This means that one or more elasticated
constraints, i.e., (13b)-(13e), are not satisfied. To ensure these constraints, we can increase network resources (e.g., server’s capacities) or reject some of the users’ service requests. Since the first method is not practical in more cases, we propose to reject some requested services by adopting the proposed AC. In behind of AC, one of the major questions is which one of the requested services should be rejected. In this case, the requested services have diverse characteristics and different effects on the utilization of the network resources, and consequently on the infeasibility of problem (12). To find the user which has the most effects on the infeasibility and reject its service, we do as follows:

$$u^\star = \arg\max_u \varDelta_u \triangleq \kappa_1(R_{s_{\text{min}}} - R_u) + \kappa_2 \sum_{n \in N} \left( \sum_{s \in S} \sum_{f_{n} \in \Omega_s} \psi_{s_{}}^{f_{n}} \beta_{u, n}^{f_{n}} + y_u \sum_{f_{n} \in \Omega_s} \beta_{u, n}^{f_{n}} - \Upsilon_n \right)$$

$$+ \kappa_3 \left( \sum_{n \in N} \sum_{s \in S} \sum_{f_{n} \in \Omega_s} y_u \alpha_{s_{}}^{f_{n}} \beta_{u, n}^{f_{n}} - L_n \right),$$

(17)

where \(\kappa_1, \kappa_2, \text{ and } \kappa_3\) are the fitting parameters to balance \(\varDelta_u\), and we emphasize that in (17) we use the values of the optimization variables of (13) obtained by Algorithm (6). Based on this, we reject user \(u^\star\). Then, solve problem (13) with \(\mathcal{U}' = \mathcal{U} - \{u^\star\}\). We repeat this procedure until, we have \(A = 0\) in the solution of problem (13).

IV. CONVERGENCE AND COMPUTATIONAL COMPLEXITY

A. Convergence of the Solution Algorithm

Based on ASM, after each iteration the objective function in each sub-problem is enhanced and finally it converges. Fig. 4 shows an example about the convergence of our proposed iterative algorithm. Clearly, our proposed solution converges after few iterations.

**Proposition 2.** With a feasible initialization of problem (13), the ASM algorithm converges to a sub-optimal solution.

**Proof.** Remind the objective function of problem (13) as follows:

$$O(\mathcal{P}, \rho, \eta, A) := \sum_{u \in \mathcal{U}} \sum_{k \in K} \mu \rho_{u}^{k} p_{u}^{k} + \sum_{n \in N} \nu_{\eta}^{n} + W A.$$  

(18)
We have the following relations between iterations ($z$ is the iteration number):

$$O(P[z], \rho[z], \eta[z], A[z]) = \min_{P,A} O(P[z], \rho[z], \eta[z], A[z]) \leq O(P[z - 1], \rho[z], \eta[z], A[z - 1])$$

$$= \min_{\rho} O(P[z - 1], \rho[z], \eta[z], A[z - 1]) \leq O(P[z - 1], \rho[z - 1], \eta[z], A[z - 1])$$

$$= \min_{\eta} O(P[z - 1], \rho[z - 1], \eta[z], A[z - 1]) \leq O(P[z - 1], \rho[z - 1], \eta[z - 1], A[z - 1]).$$

This means that the objective function of ASM decreases as the iteration number increases. In addition, with QoS and ensuring the resource demand constraints, i.e., (13b)-(13e), the ASM algorithm converges to a sub-optimal solution which corresponds to the sub-optimal solution of problem (13).

**Proposition 3.** Algorithm 2 is a monotonic algorithm and generates decreased values for the objective function at each iteration $z$.

**Proof.** Algorithm (2) works based on the values of QoS metrics ($D_s^{\max}$ and $R_s$) and capacity requirement of NFs that are in the requested NSs. For the given system parameters such as the number of NFs and $\alpha^f_s$, just the value of $R_u$ is variant and depends on the value of the optimization variables. Therefore the value of it has impact on the value of $\eta = [\eta_n]$ that is output of Algorithm (2). Based on equation (7) and server selection policy of Algorithm (2), $R_u$ is directly proportional to $\eta$. Hence, if the value of $R_u$ is fixed or reduced at each iteration
TABLE II
COMPLEXITY ORDER OF THE PROPOSED SOLUTIONS

| Method                                      | Complexity                        |
|---------------------------------------------|-----------------------------------|
| Heuristic                                   | \(O(U^2 \times F \times N)\)       |
| Greedy-based solution                       | \(O(U^2 \times F \times N)\)       |
| Power Allocation: CVX                       | \(\log \left( \frac{C_1}{\varepsilon} \right) / \log(\varsigma)\) |
| Subcarrier Allocation: CVX-MOSEK            | \(\log \left( \frac{C_2}{\varepsilon} \right) / \log(\varsigma)\) |

\(z\), i.e., if \(R_u^{(z)} \leq R_u^{(z-1)}\), then we have \(\eta^{(z)} \leq \eta^{(z-1)}\). As a result the proposed algorithm is monotonic.

B. Computational Complexity

By utilizing ASM, the overall complexity of the algorithm is a linear combination of the complexity of each sub-problem.

1) Radio RA: For the radio RA sub-problem, we utilize geometric programming (GP) and IPM via CVX toolbox in MATLAB [27]. Based on this method, the computational complexity order of power allocation sub-problem is given by \(\log \left( \frac{C_1}{\varepsilon} \right) / \log(\varsigma)\) where \(C_1 = U + N + N \times U + U + 1\) is the total number of constraints of sub-problem (14), \(\xi\) is the initial point for approximating the accuracy of IPM, \(0 < \varrho \ll 1\) is the stopping criterion for IPM, and \(\varsigma\) is the accuracy of IPM [27]. Similarly, the complexity of sub-problem (15) is given by \(\log \left( \frac{C_2}{\varepsilon} \right) / \log(\varsigma)\) where \(C_2 = U + K + 1\) is the total number of constraints of (15).

2) NFV RA: Based on the proposed heuristic algorithm in Algorithm 2 for NFV RA, the complexity order of sub-problem (16) is the total number of main calculations that are required for solve it. Hence, the upper bound of complexity of Algorithm 2 is \(O(U^2 \times F \times N)\). The order of computational complexity of all sub-problems are summarized in Table II.

V. EXPERIMENTAL EVALUATION

A. Simulation Environment

In this section, the simulation results are presented to evaluate the performance of the proposed system model. We consider \(U = 50\) users which are randomly distributed in the converge of a
TABLE III

RADIO ACCESS NETWORK CONFIGURATION PARAMETERS VALUES

| Parameters(s)          | Value(s)          |
|-----------------------|-------------------|
| $U$                   | Min: 5 Max: 50    |
| $K$                   | 64                |
| $\sigma$              | $10^{-7}$ Watts   |
| BS radius             | 500 m             |
| $P_{\text{max}}$      | 40 Watts          |
| $\rho$                | 3                 |
| $P_u^{\text{min}}$    | Min: 5 Max: 20 in bps/Hz |

BS with radius 500 m, $\sigma = 10^{-7}$ Watts, $h_u^k = x_{u,k}(d_u)^{-\varphi}$ where $\varphi = 3$ is the path loss exponent, $x_{u,k}$ is the Rayleigh fading, and $d_u$ is the distance between the BS and user $u$. Moreover, we set $K = 64$, $S = 20$, and $N = 25$. We suppose that the users request services with randomly uniform distribution as $RS_u \sim U_d[1, S]$. We define $M = 15$ different NFs with unique labels 1-15, i.e., $\mathcal{F} = \{f_1, \ldots, f_{15}\}$. Each NS $s$ is a combination of several NFs. Each NF in the requested service utilizes existing network resource until its processing time is completed. The other radio network settings are summarized in Table III. For the sake of clarity of network configuration, the main related core network parameters utilized in these simulations for creating the VMs and services are chosen randomly based on the uniform distribution with the minimum and maximum values, i.e., $U_d[\text{minValue}, \text{maxValue}]$ that are shown in Table IV [1].

TABLE IV

CORE NETWORK CONFIGURATION PARAMETERS VALUES

| Parameters(s)                      | Value(s)                      |
|------------------------------------|-------------------------------|
| $N$                                | Min: 15 Max: 40               |
| Server buffer capacity             | Min: 1000 Max: 1500           |
| Function buffer demand             | Min: 5 Max: 15                |
| Number of services                 | Min: 10 Max: 25               |
| Number of functions in each service| Min: 5 Max: 15                |
| Server processing capacity, i.e., $L_n$ | Min: 1500 Max: 3000 Bits per second |
| Processing demand of each NF       | Min: 5 Max: 20 Bits per second|
| Service processing deadline        | Min: 0.3 Max: 7 second        |

B. Simulation Results

The investigation of the proposed algorithm under different network settings and parameters is started by the simulation results that are shown in Figures 5(a)-7(b). These results are obtained by
considering the Monte-Carlo method iterations is 500. We discuss these results in the following.

1) **Acceptance Ratio**: The acceptance ratio is defined by the ratio of the number of accepted services by the network to the total number of the requested services by users and is obtained by $\alpha = 1 - \frac{\hat{U}}{U}$ where $\hat{U}$ is the number of users that their services are rejected based on the proposed AC. It is a criterion to investigate the efficiency of the proposed algorithm in utilizing total network resources to guarantee the requested QoS and accept the service demands.

As can be seen from Figures 5(a), 5(b), and 5(c), the value of the acceptance ratio depends on two main factors, i) the network resources capacity; ii) the number of users (service demands) and service QoS characteristics (latency and data rate). Therefore, we have some challenges to address high data rate and provide low latency services.

Fig. 5(b) illustrates the variation of the value of service acceptance ratio with the number of users (service arrivals) for different service deadlines and data rates. In this figure, we set $K = 64$, $P_{\text{max}} = 80$ Watts, $N = 40$, $F = 15$, $L_n \in [500 \: 2000]$ CPU cycles per unit time, and $S = 15$. Clearly, by increasing the number of users (service requests) the acceptance ratio is decreased, especially for low latency services that have the main contribution on the acceptance ratio. We observe that increasing the number of low latency services leads to reducing the acceptance ratio. For the large number of users, the network guarantees some users’ service requirements and other users are rejected. For this cases, based on $\alpha$ equation, the value of $\hat{U}$ is increased and $U - \hat{U}$ approximately reaches to a fixed value.

Fig. 5(a) shows the variation of the service acceptance ratio with different values of the service latency and server’s processing capacity. This result is obtained for $N = 40$, $F = S = 15$, $U = 50$, $\alpha^{f_m} = 20$, $P_{\text{max}} = 40$ Watts, and $R_{s}^{\text{min}} = 5$ bps/Hz. Form this figure, it can be seen the rejection probability of the low latency services is higher than other types of services. The reason for this is that these services need more servers with high processing capacity to reduce the waiting and processing time. It is clear that by increasing the latency from 0.3 to 1, the acceptance ratio is increased approximately by 50%. Moreover, this figure shows the impact of the minimum data rate on the acceptance of the service request. As can be seen from this figure,
in contrast to the latency requirement, high data rate services are rejected by the network. It would be better that we investigate the effect of latency versus the data rate. Clearly, if the minimum data rate value is doubled, on average approximately \( \frac{0.4292}{0.7372} = 0.58 \approx 58\% \) of users are rejected. Whereas, if latency is halved (0.7 to 0.3), on average approximately \( \frac{0.1155}{0.1555} = 0.74 \approx 74\% \) of users are rejected by the network. Moreover, from this figure, we observe that by increasing the maximum processing capacity form 2500 to 3500, the service acceptance ratio improves by approximately 2 times. In other words, physical resource capacity has a major effect on the acceptance of services by the network, especially for the average latency of about [1 2]. Whereas, high order latency services are not sensitive to the value of the server’s capacity.

Fig. 5(c) shows the variation of the service acceptance ratio with increasing the number of servers for different scenarios. In this figure, we set \( U = [50 \ 70] \), \( \max L_n = 2500 \) bps, \( \alpha_{fs} = 20 \), and \( M = S = 15 \). Clearly, increasing the number of servers in the network improves the service acceptance ratio. Due to the fact that increasing the number of servers reduces the waiting time of NFs to run in the mapped servers. On the other hand, the probability of the large number of mapped NFs on each server is low and hence, the waiting and processing times are reduced. Therefore, the latency and buffering requirements are satisfied and the acceptance ratio of services is improved. From this figure, we conclude that the impact of the number of active servers on the high data rate and low latency services e.g., process automation [29] is more than that of other
services. Furthermore, comparing Fig. 5(c) and Fig 5(a), we obtain that the effect of the server processing capacity is more considerable than the number of active servers on the low latency services. That means low latency services are rejected by the network because their requirements need more resources in the network to reduce waiting and processing times.

2) Network Cost: Fig. 6(a) illustrates the network cost versus the variation of the number of users for $R_s^{\text{min}} = 10 \text{ bps/Hz}$ and service deadline 2 second. The network cost is comprised of both radio and NFV resources costs in terms of power and spectrum consumption and utilizing servers in the network. It can be observed that by increasing the number of users the network cost increases due to increase in both the radio and NFV costs. It is clear that by increasing the number of users the NFV cost increases rapidly than that of the radio cost.

Fig. 7(a) investigates the impact of the total number of users in the network on the utilization of resources with different minimum data rates (as a packet size) and service deadlines. In this experiment, we restrict the number of servers to 80 with maximum processing capacity $L_n = 3000$, $M = 15$, $S = 20$, and $\alpha_f^{\text{min}} = 20$. We define the utilization ratio as $\text{UtiRatio} = \frac{r_U}{r_T}$ where $r_U$ is the amount of the resources utilized by the users and $r_T$ is the total server’s resources. From this figure, we infer that not only the packet size has a direct effect on the utilization ratio, but also the service deadline has a major impact on this. This is due to the fact that a large packet size needs more storage and processing capacity and low service deadline needs minimum waiting and processing times. Therefore, we should make active more servers and exploit their resources for low latency services. Obviously, increasing the number of users increases the utilization ratio approximately in a linear form. From the cost perspective, we can conclude that by increasing the utilization of network resources, the cost network is also increased, especially in terms of power consumption.

3) Service Deadline: Figures 6(b) and 7(b) show the total cost of the network versus different values of the service deadline for various scenarios. Clearly, the requested service deadline has a major effect on the utilization of processing and buffering resources in servers. Form Fig. 7(b), we conclude that for services with lower latency requirements, more servers should be active.
to process the NFVs of corresponding services. That means for providing low latency services, we should pay more costs in terms of radio and NFV resources. By increasing the number of servers, the waiting time for each NF in a NS that it is in queue is minimized, and hence, server availability and probability of QoS guarantee for users are increased. For higher latency values in some cases, one (or two) active server(s) is sufficient. By comparing Fig. 6(b) and Fig. 6(a), we obtain that by reducing the value of the latency, the network cost increases significantly compared to the case where the number of users (the numbers of service requested) increases.

C. Benchmark Algorithms

1) Performance Comparison: To the best of our knowledge, this is the first work (refer to the related works) tackling the E2E RA with proposing new AC and a new closed form formulation of NFV scheduling which is comprised of both waiting time and SFC ordering. Moreover, we propose a new heuristic algorithm to solve the formulated optimization problem. It is worth nothing that in the related works, a greedy-based search is exploited [1], [8], [30]. We compare our proposed algorithm with a modified version of greedy-based algorithm that is proposed in [1]. In greedy-based search, different objectives can be considered, for example, minimizing the total flow time [1]. The greedy-based scheduling and embedding the arrival service requests are performed sequentially based on the greedy criteria. Based on the modified greedy algorithm
(a) The ratio of the utilization of server’s resources versus the total number of users in the network.

(b) The number of active (on) servers versus the service deadline.

Fig. 7. Evaluation of utilization of resources in terms of the number of active servers and ratio of active servers’s resources with variation of different parameters.

to solve sub-problem (16), first, we search servers that are appropriate for embedding and then find the best server by greedy criterion, i.e., the shortest server queues [1]. The steps of the greedy-based algorithm with the minimum queue time criterion is stated in Algorithm 3.
Algorithm 3: Greedy-based NFV RA to solve sub-problem (16)

**Input:** $S,F,\Omega_s,\alpha^{f_m},\rho^{f_m},t'(n)=0,\forall n\in\mathcal{N}$ is the last running time of server $n$

1. Sorted in ascending order all users according to the latency requirement and write user’s index in $\tilde{U}_T$

2. for $u=1: \tilde{U}_T$ do

3. for $m=1:|\Omega_s|$ (s is requested service of user $u$) do

4. Check processing and buffer constraints, i.e., (13c) and (12c), and write servers that satisfy them in set of candidate servers as $\mathcal{N}_{Can}$

5. Sorted in ascending order $\mathcal{N}_{Can}$ according to greedy criterion, i.e., the shortest queuing time for function $f^s_m$

6. Select the first rank server and set $\beta^{f_m}_{u,n}=1$ (index $n$ has first rank in $\mathcal{N}_{Can}$)

7. if $m==1$ then

8. $t^{f_m}_{u,n}=t'(n)$

9. else

10. $t^{f_m}_{u,n} = \max\{t^{f_m-1}_{u,n}, t'(n)\} \& t'(n)=t^{f_m}_{u,n} + \tau^{f_m}_{n}$

11. Update the last released time of server $n$

12. Calculate $\phi_{u,n}^{f_m} = \sum_{n'\in\mathcal{N}} \beta^{f_{m'}}_{u,n'}(t^{f_{m'}}_{u,n'} + \tau^{f_{m'}}_{n'})$

13. for $u'\in\mathcal{U}, u' \neq u \& f^{s'}_{m'}\in\Omega_{u',s'}$ do

14. if $\beta^{f_{m'}}_{u,n}=1 \& \phi_{u,n}^{f_m} \geq t^{f_{m'}}_{u',n}$ then

15. $x^{f_m}_{u',n'=1}$

16. Calculate $t^{f_m}_{u,n}$ according to constraint (12g) \& $t'(n)=t^{f_m}_{u,n}$

**Output:** $T, X, \eta,$ and $\beta$

Fig. 8(a) highlights the comparison of this algorithm with the proposed heuristic algorithm from the acceptance ratio perspective. As seen, the acceptance ratio of the proposed algorithm is better than the greedy algorithm. For a small number of users, the results of both algorithms are the same. As a reason in the case the network resource requirements are low, both algorithms accept almost all users. Moreover, we compare the impact of two mentioned algorithms on the number of active servers versus the requested service deadline in Fig. 8(b). From this figure, we conclude that the number of active servers is greatly lower than that of the greedy algorithm. As a reason, in the greedy algorithm for each NF, the algorithm finds a server with the lowest queuing
time. In some cases, the algorithm adds servers that are release and have more capacities. While it is possible to satisfy the latency of other functions without utilizing this server. In constant, the proposed algorithm activates a server when the previously added servers (activated servers) cannot satisfy the constraints of the problem and users QoS. More importantly, in the greedy algorithm, the number of active servers is fixed after increasing the values of the service deadlines which is the consequent of its server selection policy which is based on the queuing time. While in the proposed algorithm it is reduced.

To more clarify this, assuming that we have five servers in the network with specific capacities as [1000 2000 1500 3000 1800] and two service requests with 2 functions with capacity requirements 20 and 40, \( R_s^{\text{min}} = 10 \text{ bps/Hz} \) and service deadlines 0.3 and 0.7, respectively. Based on Algorithm 2, the service finishing time of user 1 is \( \frac{20 \times 10}{3000} + \frac{40 \times 10}{3000} = 0.2 < 0.3 \) and service finishing time of user 2 is \( \frac{20 \times 10}{3000} + \frac{40 \times 10}{3000} = 0.4 < 0.7 \). That means one active server is sufficient to all users. While based on the greedy algorithm the finishing service time of user 1 is \( \frac{20 \times 10}{3000} + \frac{40 \times 10}{3000} = 0.2 < 0.3 \) and that of user 2 is \( \frac{20 \times 10}{2000} + \frac{40 \times 10}{2000} = 0.3 < 0.7 \), since 0.3 \( < 0.4 \), the greedy algorithm selects a server with capacity 2000 instead of the server with capacity of 3000. As a result, based on the greedy algorithm, two servers are utilized while in the proposed algorithm, only one server in both cases is utilized. Clearly, the greedy algorithm utilizes servers inefficiently, and hence, the acceptance ratio is decreased especially for a large number of users (see Fig. 8(a)).

2) **Optimality Gap**: Another metric for investigation of the performance of the proposed algorithm is optimality gap. In this regard, we adopt the exhaustive search method [31]. Since the complexity of the exhaustive search method is very high and exponentially grows with the size of system parameters, we exploit it for a small scaled network. The considered parameters and the corresponding solution methods values are stated in Table V. The other parameters are based on Tables IV and III.
Fig. 8. Evaluation of different solution methods form different aspects.

**TABLE V**

PERFORMANCE COMPARISON OF DIFFERENT SOLUTION METHODS.

| Scenarios                                                                 | Solution Methods | ASM-Greedy | ASM-Proposed | Optimal |
|---------------------------------------------------------------------------|------------------|------------|--------------|---------|
| Network Cost ($M = S = 5, N = 10, U = K = 10, D_u^{\text{max}} = 1.5, R_s^{\text{min}} = 10$) | ASM-Greedy       | 254        | 232          | 217     |
| Acceptance Ratio ($M = S = 5, U = 40, N = 15, K = 10, D_u^{\text{max}} = 2, R_s^{\text{min}} = 15$) | ASM-Greedy       | 0.59       | 0.64         | 0.7     |
| Number of Activated server ($U = 15, M = S = 5, K = 10, D_u^{\text{max}} = 1, R_s^{\text{min}} = 10$) | ASM-Greedy       | 16         | 9            | 7       |

VI. CONCLUSION

In this paper, we proposed a novel joint radio and NFV RA for heterogeneous services. Our aim is to minimize the utilization of the radio resources and servers. Therefore, we proposed a novel scheduling and energy efficient scheme in terms of minimizing the number of activated servers based on a new heuristic algorithm. More importantly, our scheduling scheme includes queuing effect such as queuing waiting time. To solve the proposed problem, first, we divided it into three sub-problems, and then efficiently solved each of them. To solve NFV RA, we proposed a
novel low complexity heuristic algorithm that is based on minimizing the number of active servers in the network. By this scheme, we significantly reduced the resources cost such as processing, buffering, and power consumption. Moreover, we proposed a novel AC scheme determining which one of the requested services have critical requirements and needs more resources in terms of radio and NFV resources to ensure their QoS requirements and then rejects their services. We evaluated the performance of the proposed scheme with different network parameters and variables such as service demands, service QoS, and network resource capacities. Our simulation results are carried out with different values of the network parameters and service requested with various metrics such as service acceptance ratio, the number of active servers, and network predefined cost. Moreover, to verify the performance of the proposed algorithm, we compared it with the conventional one from the performance perspective. Our simulation results demonstrate that the proposed algorithm outperforms the conventional one by approximately 8%.

REFERENCES

[1] R. Mijumbi, J. Serrat, J. Gorricho, N. Bouten, F. De Turck, and S. Davy, “Design and evaluation of algorithms for mapping and scheduling of virtual network functions,” in Proc. 2015 1st IEEE Conference on Network Softwarization (NetSoft), April 2015, pp. 1–9.

[2] Z. Chang, Z. Zhou, S. Zhou, T. Chen, and T. Ristaniemi, “Towards service-oriented 5G: Virtualizing the networks for everything-as-a-service,” IEEE Access, vol. 6, pp. 1480–1489, Dec 2018.

[3] J. F. Riera, E. Escalona, J. Batallé, E. Grasa, and J. A. García-Espín, “Virtual network function scheduling: Concept and challenges,” in Proc. 2014 International Conference on Smart Communications in Network Technologies (SaCoNeT), June 2014, pp. 1–5.

[4] R. Mijumbi, J. Serrat, J. Gorricho, N. Bouten, F. De Turck, and R. Boutaba, “Network function virtualization: State-of-the-art and research challenges,” IEEE Communications Surveys Tutorials, vol. 18, no. 1, pp. 236–262, Sep 2016.

[5] J. Gil Herrera and J. F. Botero, “Resource allocation in NFV: A comprehensive survey,” IEEE Transactions on Network and Service Management, vol. 13, no. 3, pp. 518–532, Sep. 2016.

[6] A. N. Al-Quzweeni, A. Q. Lawey, T. E. H. Elgorashi, and J. M. H. Elmirghani, “Optimized energy aware 5G network function virtualization,” IEEE Access, pp. 1–1, 2019.

[7] R. Mijumbi, J. Serrat, J. Gorricho, N. Bouten, F. De Turck, and R. Boutaba, “Network function virtualization: State-of-the-art and research challenges,” IEEE Communications Surveys Tutorials, vol. 18, no. 1, pp. 236–262, Sep 2016.
[8] D. Zeng, J. Zhang, L. Gu, and S. Guo, “Stochastic scheduling towards cost efficient network function virtualization in edge cloud,” in Proc. 2018 15th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON), June 2018, pp. 1–9.

[9] E. Hossain and M. Hasan, “5G cellular: key enabling technologies and research challenges,” arXiv preprint arXiv:1503.00674, 2015.

[10] H. Dahrouj, A. Douik, O. Dhifallah, T. Y. Al-Naffouri, and M. Alouini, “Resource allocation in heterogeneous cloud radio access networks: advances and challenges,” IEEE Wireless Communications, vol. 22, no. 3, pp. 66–73, June 2015.

[11] H. Zhang, N. Liu, X. Chu, K. Long, A. Aghvami, and V. C. M. Leung, “Network slicing based 5G and future mobile networks: Mobility, resource management, and challenges,” IEEE Communications Magazine, vol. 55, no. 8, pp. 138–145, Aug 2017.

[12] J. Liu, W. Lu, F. Zhou, P. Lu, and Z. Zhu, “On dynamic service function chain deployment and readjustment,” IEEE Transactions on Network and Service Management, vol. 14, no. 3, pp. 543–553, Sep. 2017.

[13] R. Cohen, L. Lewin-Eytan, J. S. Naor, and D. Raz, “Near optimal placement of virtual network functions,” in Proc. 2015 IEEE Conference on Computer Communications (INFOCOM), April 2015, pp. 1346–1354.

[14] 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.

[15] X. Chen, W. Ni, I. B. Collings, X. Wang, and S. Xu, “Automated function placement and online optimization of network functions virtualization,” IEEE Transactions on Communications, vol. 67, no. 2, pp. 1225–1237, Feb 2019.

[16] D. Li, P. Hong, K. Xue, and j. Pei, “Virtual network function placement considering resource optimization and SFC requests in cloud datacenter,” IEEE Transactions on Parallel and Distributed Systems, vol. 29, no. 7, pp. 1664–1677, July 2018.

[17] R. Riggio, A. Bradai, D. Harutyunyan, T. Rasheed, and T. Ahmed, “Scheduling wireless virtual networks functions,” IEEE Transactions on Network and Service Management, vol. 13, no. 2, pp. 240–252, June 2016.

[18] M. A. T. Nejad, S. Parsaeefard, M. A. Maddah-Ali, T. Mahmoodi, and B. H. Khalaj, “vSPACE: VNF simultaneous placement, admission control and embedding,” IEEE Journal on Selected Areas in Communications, vol. 36, no. 3, pp. 542–557, 2018.

[19] H. Kim, “Performance evaluation of revised virtual resources allocation scheme in network function virtualization (NFV) networks,” Cluster Computing, pp. 1–9, 2018.

[20] L. Qu, C. Assi, and K. Shaban, “Delay-aware scheduling and resource optimization with network function virtualization,” IEEE Transactions on Communications, vol. 64, no. 9, pp. 3746–3758, Sep. 2016.

[21] B. Ren, D. Guo, Y. Shen, G. Tang, and X. Lin, “Embedding service function tree with minimum cost for NFV-enabled multicast,” IEEE Journal on Selected Areas in Communications, vol. 37, no. 5, pp. 1085–1097, May 2019.

[22] P. D. Diamantoulakis, K. N. Pappi, G. K. Karagiannidis, H. Xing, and A. Nallanathan, “Joint downlink/uplink design for wireless powered networks with interference,” IEEE Access, vol. 5, pp. 1534–1547, Jan 2017.

[23] G. ETSI, “Network functions virtualisation (NFV); terminology for main concepts in NFV,” ETSI GS NFV, vol. 2, no. 2, p. V1, Aug. 2018.

[24] ——, “Network functions virtualisation (NFV): Architectural framework,” ETSI GS NFV, vol. 2, no. 2, p. V1, 2018.
[25] M. S. Yoon and A. E. Kamal, “NFV resource allocation using mixed queuing network model,” in Proc. IEEE Global Communications Conference (GLOBECOM). IEEE, 2016, pp. Washington, DC, USA, 1–6. Dec. 2016.
[26] G. T. 28.530, “Management and orchestration; concepts, use cases and requirements,” IEEE Systems Journal, Mar. 2019.
[27] M. Grant, S. Boyd, and Y. Y. Cvx, “Matlab software for disciplined convex programming, version 1.0 beta 3,” Recent Advances in Learning and Control, pp. 95–110, 2006.
[28] A. Mosek, “The MOSEK optimization toolbox for MATLAB manual,” 2015.
[29] S. Misra, A. Mondal, and S. Khajjayam, “Dynamic big-data broadcast in fat-tree data center networks with mobile IoT devices,” IEEE Systems Journal, pp. 1–8, March 2019.
[30] S. Mehraghdam, M. Keller, and H. Karl, “Specifying and placing chains of virtual network functions,” in Proc. IEEE 3rd International Conference on Cloud Networking (CloudNet), Luxembourg, Luxembourg, PP. 7-3, Oct. 2014.
[31] A. Zakeri and N. Mokari, “Joint radio resource allocation and 3D beam-forming in MIMO-NOMA-based network with profit maximization for mobile virtual network operators,” arXiv preprint arXiv:1907.05161, 2019.