



## Service-Oriented Network Resource Orchestration in Space-Air-Ground Integrated Network

Journal:	<i>IEEE Transactions on Vehicular Technology</i>
Manuscript ID	VT-2022-04874
Suggested Category:	Regular Paper
Date Submitted by the Author:	12-Dec-2022
Complete List of Authors:	<p>He, Jingchao            Cheng, Nan; Xidian University, School of Telecommunications Engineering            Yin, Zhisheng; Xidian University,            Zhou, Conghao; University of Waterloo, Electrical and Computer Engineering            Zhou, Haibo; Nanjing University, School of Electronic Science and Engineering            Quan, Wei            Lin, Xiao-Hui; Shenzhen University, Information Engineering; College of Information Engineering, Shenzhen University</p>
Keywords:	space-air-ground integrated network (SAGIN), software-defined networking (SDN), network function virtualization (NFV), service function chain (SFC), wireless resource allocation

SCHOLARONE™  
Manuscripts

# Service-Oriented Network Resource Orchestration in Space-Air-Ground Integrated Network

Jingchao He, *Student Member, IEEE*, Nan Cheng, *Member, IEEE*, Zhisheng Yin, *Member, IEEE*, Conghao Zhou, *Member, IEEE*, Haibo Zhou, *Senior Member, IEEE*, Wei Quan, *Member, IEEE*, Xiao-Hui Lin

**Abstract**—Space-air-ground integrated networks (SAGINs) are envisioned to provide seamless coverage and enhanced flexibility compared with traditional terrestrial mobile networks, which has attracted much attention from both industry and academia. However, orchestrating heterogeneous resources in such a large-scale and dynamic network is challenging, especially encountering diverse services with multi-dimensional requirements. In this paper, we first propose a software-defined networking (SDN) and network function virtualization (NFV)-based reconfigurable SAGIN architecture for constructing service function chains (SFCs). Based on that, we investigate the SFC orchestration and wireless resource management where the **rate adaption** is introduced to improve the network resource utilization. Considering the limited physical resource and the heterogeneity in SAGINs, we jointly formulate the virtual network function (VNF) embedding, **transmission rate adjustment**, and wireless resource allocation as a mixed-integer nonlinear programming (MINLP) problem to maximize the network profit. Due to the NP-hardness of the problem, we first transform the problem into a continuous optimization problem by successive convex approximation. By introducing an additional penalty into the objective function, an iterative alternation algorithm is proposed to find a near-optimal solution of the transformed problem. Extensive simulation results show that our proposed approach outperforms the benchmarks in average network revenue, successfully serving probability, and resource consumption.

**Index Terms**—Space-air-ground integrated network (SAGIN), software-defined networking (SDN), network function virtualization (NFV), service function chain (SFC), wireless resource allocation.

## I. INTRODUCTION

**T**RADITIONAL terrestrial mobile networks have achieved high capacity and low latency with advanced wireless communication and antenna technologies, which have enabled

many a large number of applications, such as augment reality (AR)/virtual reality (VR), industry Internet of things (IoT), and connected vehicles [1], [2]. However, due to the fixed deployment of the network infrastructure, the overall network topology is static, and it is difficult to dynamically adjust the network resource distribution according to real-time requirements. Furthermore, dense base station (BS) deployment and terrestrial backhaul network construction is prohibitively expensive, and the deployment of terrestrial networks is limited by terrain such as deserts, and oceans. Thus, conventional ground-based networks are difficult to cope with the extension trend of network coverage [3], [4].

To complement conventional terrestrial networks, both academia and industry pay attention to non-terrestrial networks (NTNs). Low earth orbit (LEO) satellite constellation enables high-capacity and low-cost satellite network to provide global coverage [5]. Particularly, the STARLINK LEO constellation operated by SpaceX has provided network access to 32 countries, and 12,000 LEO/very-low-earth-orbit (VLEO) satellites are planned to provide global coverage, with a possible extension to 42,000 in the future. Besides, aerial networks, mainly containing high-altitude platform stations (HAPs) and unmanned aerial vehicles (UAVs), have been attracting people in recent years [6], [7]. HAPs are defined as the radio stations at 20-50 kilometers above the Earth, which are able to suspend at a fixed position to provide fixed broadband network access in hard-to-reach areas [8]. The UAV is the component of the unmanned aircraft system (UAS), which operates above the Earth without any human control and is agile enough to serve real-time hot spot areas [9].

Comprehensively integrating with the space networks, aerial networks, and ground networks, the space-air-ground integrated networks (SAGINs) are proposed, which take advantage of the complementary benefits of three network segments and offer unprecedented network ability in coverage, flexibility, capacity, and reconfigurability [10], [11]. However, the inherent heterogeneity and dynamics challenge network operators in traffic distribution, routing protocol design, and load balancing. Conventional exclusive network architectures and dedicated hardware are not reliable and cost-effective to orchestrate network resources for services with multi-dimensional requirements in such a large-scale and dynamic network scenario.

Fortunately, software-defined networking (SDN) technology can disassociate the data plane from the control plane and enables a more flexible, dynamic, and programmatically efficient network operation [12]. By network function virtualization

Jingchao He and Nan Cheng are with the State Key Laboratory of ISN and School of Telecommunications Engineering, Xidian University, Xi'an 710071, China (e-mail: jchhe@stu.xidian.edu.cn; dr.nan.cheng@ieee.org).

Zhisheng Yin is with the State Key Laboratory of ISN and School of Cyber Engineering, Xidian University, Xi'an 710071, China (e-mail: zsyin@xidian.edu.cn).

Conghao Zhou is with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, N2L 3G1, Canada (e-mail: c89zhou@uwaterloo.ca).

Haibo Zhou is with the School of Electronic Science and Engineering, Nanjing University, Nanjing 210023, China (e-mail: haibozhou@nju.edu.cn).

Wei Quan is with the School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing 100044, China (e-mail: wei-quan@bjtu.edu.cn).

Xiao-Hui Lin is with the College of Electronics and Information Engineering, Shenzhen University, Shenzhen 518060, China (e-mail: xhlin@szu.edu.cn).

(NFV) technology, resources of underlying heterogeneous physical infrastructures can be abstracted into the virtual resources pool, which provides a more flexible approach in resource management than traditional dedicated hardware. Based on the SDN and NFV, service function chains (SFCs) can be established with a set of sequential virtual network functions (VNFs), which is customizable for users with various quality-of-service (QoS) requirements. To conduct a SFC configuration, the physical resource is first abstracted and incorporated into the virtual resource pool by the Virtualized Infrastructure Manager (VIM). When service requests arrive, they are described as specific VNF chains with resource requirements and delivered to the SFC orchestrator. Based on the current network state, the SFC orchestrator decides whether to accept the service. When the SFC orchestrator accepts a service request, the corresponding network node's resource blocks are reserved. The VNFs are then executed sequentially from the source to the destination, and the service request is successfully fulfilled [13].

Considering the network topology, link state, and channel capability are static in core networks, researchers usually formulate the SFC mapping problem based on the linear programming (LP) model [14]–[20] or Markov decision process (MDP) [21]–[26], and utilize heuristic or reinforcement learning (RL)-based algorithms to solve it. These algorithms are simple but efficient, which are easy to obtain near-optimal solutions under stable network status with adequate iterations or offline training when the action space is small and discrete. Nevertheless, the dynamic network infrastructure in SAGINs alters the channel status and connectivity, rendering previously efficient resource orchestration inefficient over time, thereby degrading algorithm performance and even infringing on the user requirements [27], [28]. To address the dynamic environment of SAGINs, the coordination of SFC mapping and network resource scheduling is vital in model design. Furthermore, current research on SFC orchestration only considers the SFC mapping without taking the transmission rate into account, which significantly reduces service reception and increases network blocking. In order to maximize network performance and improve heterogeneous resource utilization, a joint algorithm for rate-adaptive SFC orchestration and network resource allocation is urgently required.

In this paper, we investigate the rate-adaptive SFC orchestration and wireless resource allocation jointly. A problem is formulated to maximize the network profit by optimizing the SFC orchestration, transmission rate scheduling, spectrum allocation, and power allocation, where the SFC provision and network resources are constrained. Since it is a mixed integer nonlinear programming (MINLP), it is non-convex and NP-hard, indicating that it is intractable. To solve this problem, we first transform it into a continuous optimization problem by successive convex approximation, where the additional penalty is introduced into the objective function to offset the influence of the inconsistency of integers. To address highly coupled variables in constraints, an iterative alternating optimization algorithm is proposed to obtain near-optimal solutions. During the optimization process, SFC mapping and network resource allocation are optimized with transmission rate iteratively. Our

main contributions can be summarized as follows.

- 1) We propose an SDN/NFV-based SAGIN architecture to support multi-dimensional resource orchestration in a large-scale dynamic network environment.
- 2) Based on the proposed architecture, we formulate a joint optimization problem of SFC orchestration and wireless resource scheduling considering service provision constraint, network resource limitation, and long-term aerial network stability preservation. Specifically, the rate adaption is introduced into the optimization model to maximize the network profit.
- 3) To achieve efficient service deployment and network resource utilization, we present an iterative alternating optimization algorithm by convex approximation. Then, we analyze the influence of wireless resources and derive the expectation of network receiving capacity.
- 4) Extensive simulation results are exhibited to evaluate the proposed algorithm and architecture in network profit, service acceptance ratio, average resource costs, etc.

The remainder of this paper is organized as follows. In Section II, a review on of related work is presented. In Section III, we present the considered system model in detail. An MINLP problem is formulated in Section IV with the consideration of service provision constraint and network resource constraint. A convex optimization-based iterative alternating algorithm is proposed to solve this problem in Section V. In Section VI, simulations are carried out to evaluate the performance of the proposed algorithm. Finally, Section VII concludes this paper.

## II. RELATED WORK

By investigating the state-of-the-art studies, there has been abundant research on the SFC orchestration in terrestrial networks and few related work in SAGINs.

For the SFC orchestration in terrestrial networks, it is usually formulated as an integer linear programming (ILP) [14]–[18] or mixed integer linear programming (MILP) optimization model [19], where the SFC mapping is optimized to maximize the network revenue [29]. Specifically, the carrier level VNFs placement problem in the cloud is studied in [19] where a betweenness-centrality-based algorithm is proposed to minimize the intra- and end-to-end delays of SFC. Considering the energy consumption of SDN switches, the SFC mapping problem is investigated to minimize the reconfiguration overhead [14]. In [15], the SFC embedding with dynamic VNF deployment in a geo-distributed cloud system is formulated as a binary integer programming, and two algorithms are presented to minimize the embedding cost and service latency separately [15]. Similar research on SFC mapping is presented to minimize the energy consumption in [16], [17]. To cope with the service interruptions caused by node failures, some researchers concentrated on promoting SFC reliability by optimizing both the SFC mapping and backup instance deployment [18], [20]. Besides, the SFC mapping problem is formulated as an MDP model, and the deep neural network (DNN)-based [21], [22] or graph-neural network (GNN)-based [23]–[26] approaches are utilized to solve the problem.

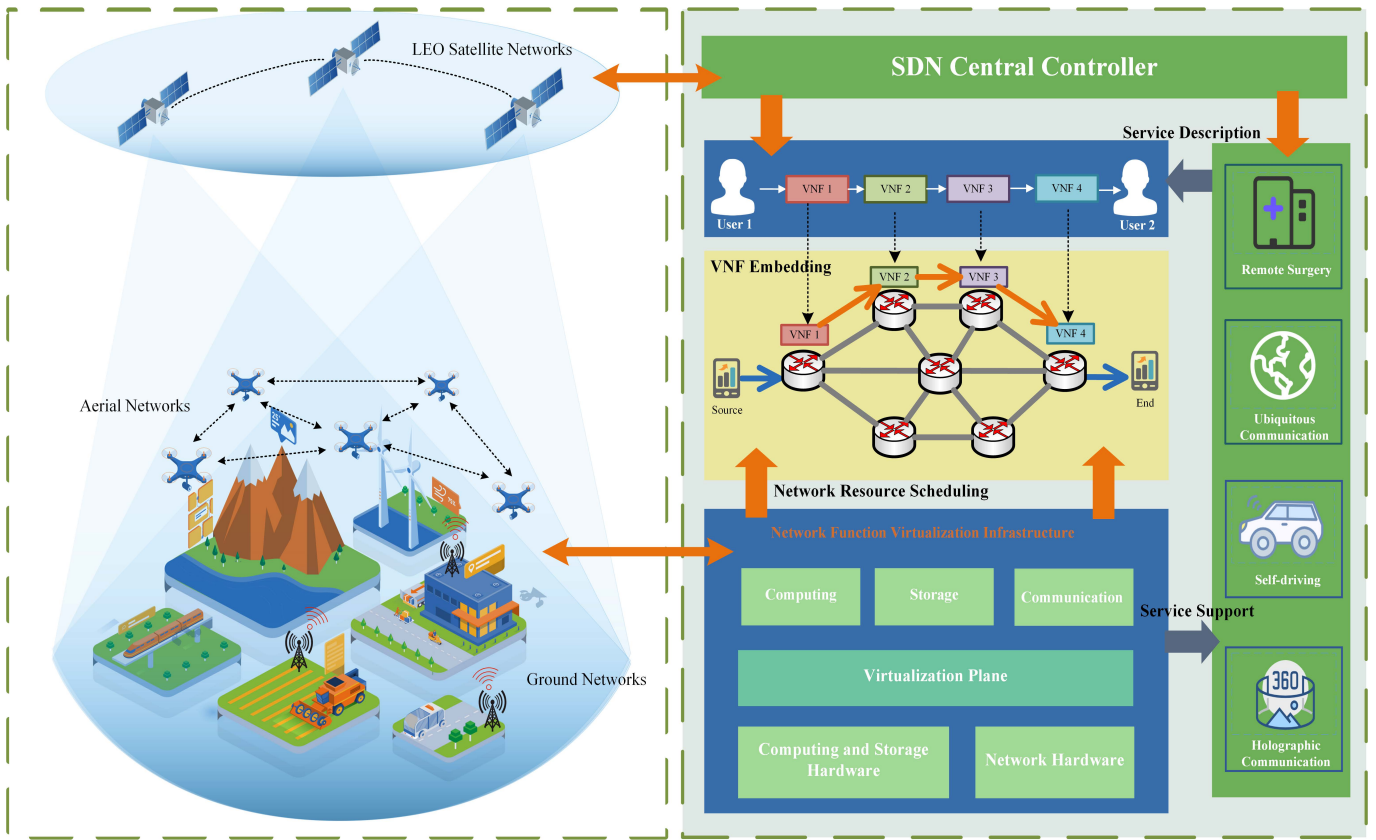


Fig. 1. SDN/NFV-based reconfigurable SAGIN network architecture

There are less research works on SFC orchestration in SAGINs, and most studies introduce the SAGIN to complement conventional terrestrial wireless networks in coverage expansion and performance enhancement [4], [13], [30], [31]. In [30], an SAGIN-based network management and reconfiguration framework is proposed to offload bidirectional missions, which extends the coverage of ground wireless networks and enhances the capability and sustainability of NTN. The simulation results demonstrate that the proposed network architecture achieves a lower blocking rate and the average cost of computation resources compared to ground-based networks, with an acceptable additional bandwidth cost. Based on the proposed architecture in [30], an air-ground integrated architecture composed of a HAP and several ground BSs is proposed in [13], where the node capacity and coverage performance are distinct, and a new metric is defined as *aggregation ratio* to measure the tradeoff between communication costs and computation costs. Similar research is presented in [4] to maximize resource utilization. To adapt to the dynamic environment in SAGINs, an SFC provisioning and reconfiguration mechanism is proposed in [31], which enables the live VNF migration and improves the service acceptance ratio. The above references promote the network performance compared with that solely based on ground networks and provide more possibilities for future network expansion. However, to further develop the performance of SAGINs, the coordination with network resources scheduling and SFC transmission rate adjusting are equally important

[14].

### III. SYSTEM MODEL

To support multi-dimensional resource orchestration in a large-scale dynamic network environment, we propose an SDN/NFV-based reconfigurable SAGIN network architecture, as shown in Fig. 1. The architecture consists of three segments: LEO satellites in the space network, aerial nodes in the aerial network, and ground nodes in the ground network. The satellites configured with the central SDN controller are in charge of SFC orchestration and wireless resource management. Both aerial nodes and ground nodes are equipped with communication units and computation units supporting multi-VNF embedding. VNFs are dedicated and shall not be shared by other services. When a network service (e.g., remote surgery, ubiquitous communication) arrive at the network, it is described as specific sequenced VNFs, and the decision on orchestration policy (if acceptance) or rejection are made by the central controller with the consideration of service requirements and network state. The notations used in this paper are listed in Table I.

The physical network is represented by a graph  $G = (N, E)$ , where  $N$  is the set of network nodes and  $E$  is the set of physical links that interconnect network nodes. In this scenario,  $N = N_A \cup N_G$  where  $N_A$  represents the set of aerial nodes and  $N_G$  is the set of ground nodes. Denote the computation capacity of network node  $n$  by  $C_n$ ;  $E = E_G \cup E_{A1} \cup E_{A2}$ , where  $E_G$  is the set of physical

TABLE I  
NOTATIONS

Notations	Description
$N$	The set of physical nodes
$N_G$	The set of ground nodes
$N_A$	The set of aerial nodes
$E$	The set of physical links
$E_G$	The set of physical links between ground nodes
$E_{A1}$	The set of physical links from ground nodes to aerial nodes
$E_{A2}$	The set of physical links from aerial nodes to other network nodes
$Q$	The set of service requests
$C_n$	The computation capacity of physical node
$s_q$	The source node of service request $q$
$d_q$	The destination node of service request $q$
$\mathbf{f}_q$	The set of VNFs of service request $q$
$E_q$	The set of virtual links of service $q$
$x_{f,n,q}$	The binary variable that indicates whether VNF $f$ of service request $q$ is embedded on node $n$
$y_{(i,j),q}^{(n,m)}$	The binary variable that indicates whether the link between VNF $i$ and VNF $j$ of service $q$ is mapped on physical link $(n,m)$
$z_q$	The binary variable that indicates whether request $q$ is successfully accepted
$t_q$	The required time of service request $q$
$r_q$	The revenue of service request $q$
$l_q^{(i,j)}$	The transmission rate of service $q$ between VNF $i$ to VNF $j$

links between ground stations,  $E_{A1}$  is the set of wireless links from ground nodes to aerial nodes, and  $E_{A2}$  is the set of wireless links from aerial nodes to other network nodes. Ground nodes are interconnected, and the channel capacity is denoted by  $l_{G1}$ . The wireless channel capacity from ground nodes to aerial nodes is denoted by  $l_{G2}$ , which is calculated by the transmission power, path loss, and noise. The channel model from the aerial node to the aerial node or the ground node follows the free-space path loss model [32]. Frequency division multiple access (FDMA) is utilized in aerial networks, and the spectrum is allocated to each aerial node orthogonally. Without loss of generality, network topology and wireless environment are quasi-stationary during each decision-making interval. Nevertheless, it is worth noting that aerial nodes can be moving, and our proposed model is also applicable in such dynamic network scenarios.

A general SAGIN topology is shown in Fig. 2 where node 1 to node 5 are ground nodes, and node 6 to node 8 are aerial nodes that connect to each other and ground nodes via wireless channels. Service 1 is embedded on node 1, node 2, node 6, and node 3; service 2 is embedded on node 1, node 2, and node 5. Physical links have fixed capacity, which means that the redundant transmission resource between node 2 and node 4 or others cannot be used for the crowded link between node 1 and node 2. However, transmission resource in aerial networks are dynamic and reconfigurable, where unoccupied bandwidth can be scheduled to node 6 for the transmission from node 6 to node 3. In an extreme example, the rigid terrestrial network will reject many services, even if some other links are still idle. This happens when the number of the same service 1 and 2 increases, or when service 1 and service 2 themselves require more communication resources. Nonetheless, the reconfigurable network can improve resource

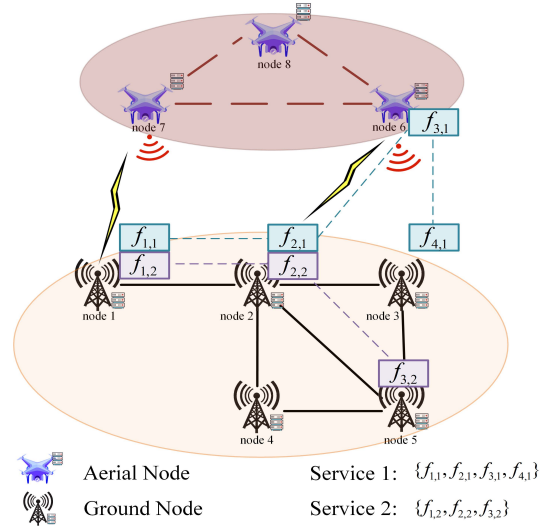


Fig. 2. A general topology of the SAGIN.

utilization and alleviate the congestion. Similarly, the rate adaption mechanism can reduce the network congestion in both terrestrial networks and aerial networks, allowing more services that would otherwise be denied to be received. Thus, our proposed model can promote resource utilization and maximize network profit.

#### A. Service Modeling

Consider that the number and types of services arriving at the network have been determined before each decision-making interval, and the set of services is denoted by  $Q = \{q | q = 1, 2, \dots, |Q|\}$ . Considering transmission requirement and computation requirement, two types of services are studied, which are high-computation low-bandwidth service and low-computation high-bandwidth service, respectively [13]. VNF sequences of each service are predefined, and VNFs are not allowed to be shared by different service requests [31]. Denote the sets of source nodes and destination nodes by  $\{s_q | q \in Q\}$  and  $\{d_q | q \in Q\}$ , respectively.  $\mathbf{f}_q = \{f_i | i = 1, 2, \dots, |\mathbf{f}_q|\}$  denotes the VNF sequence of service  $q$ . A service is completed successfully only if each VNF is executed in order from its source to destination within the required time.

Let binary variable  $x_{f,n,q} = 1$  if VNF  $f$  of service  $q$  is embedded on node  $n$  and  $x_{f,n,q} = 0$  otherwise. The solution vector is denoted by  $\mathbf{x} = \{x_{f,n,q} | \forall f \in \mathbf{f}_q, \forall n \in N, \forall q \in Q\}$ . Similarly, binary variable  $y_{(i,j),q}^{(n,m)} = 1$  when virtual link between VNF  $i$  and VNF  $j$  of service  $q$  is mapped on physical link  $(n,m)$ , and  $y_{(i,j),q}^{(n,m)} = 0$  otherwise. The solution vector is denoted by  $\mathbf{y} = \{y_{(i,j),q}^{(n,m)} | \forall (i,j) \in E_q, \forall (n,m) \in E, \forall q \in Q\}$ , where  $E_q$  denotes the virtual links from VNF  $i$  to VNF  $j$ , and  $E_q = \{(i,j) | \forall i, j \in \mathbf{f}_q, q \in Q\}$ . Another binary variable  $z_q$  is defined to indicate whether service  $q$  is received, as

$$z_q = \begin{cases} 1, & \text{service } q \text{ is successfully received,} \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

The solution vector is denoted by  $\mathbf{z} = \{z_q | q \in Q\}$ .



## B. Delay Modeling

For each service, the system earns a revenue if it is completed in time, which means that a service will be rejected if its delay requirement is violated. The delay of each service is denoted by  $t_q$ , which consists of the communication delay and computation delay, i.e.,

$$t_q = t_{\text{comm},q} + t_{\text{comp},q}, \quad (2)$$

where  $t_{\text{comm},q}$  denotes the transmission delay of service  $q$  over physical links, and  $t_{\text{comp},q}$  is computation delay at network nodes.  $t_{\text{comm},q}$  is expressed as

$$t_{\text{comm},q} = \sum_{(i,j) \in E_q} \frac{l_q}{l_q^{(i,j)}}, \quad (3)$$

where  $l_q$  represents the data volume of service  $q$  needed to be transmitted and  $l_q^{(i,j)}$  represents the allocated transmission rate of service  $q$  between VNF  $i$  and VNF  $j$ . The solution vector is denoted by  $\mathbf{l} = \{l_q^{(i,j)} \mid \forall (i,j) \in E_q, \forall q \in Q\}$ .

On the other hand, the computation delay is caused by VNF execution. Let  $c_{f,n,q}$  denote the allocated computation resource for serving the VNF  $f$  of service  $q$  at node  $n$ , and  $c_{f,q}$  represents the volume of computation data in VNF  $f$  of service  $q$ . The transmission delay of service  $q$  is expressed as

$$t_{\text{comp},q} = \sum_{f \in \mathbf{f}_q} \frac{c_{f,q}}{c_{f,n,q}}. \quad (4)$$

## C. Cost Modeling

In this subsection, we model the cost of each network node from the perspectives of resource utilization and energy consumption. The cost is mainly caused by computation and communication. The computation cost is the ratio of allocated computation resources to computation capacity [33]. For ground nodes, the communication cost depends on link utilization [13]. The total cost of ground nodes  $n$  is expressed as

$$c_n^{BS} = \alpha_{cm,N_G} \sum_{m \neq n} l_{(n,m)} + \alpha_{cp,N_G} \sum_{q \in Q} \sum_{f \in \mathbf{f}_q} \frac{c_{f,n,q} x_{f,n,q}}{C_n}, \quad \forall n \in N_G, \quad (5)$$

where  $\alpha_{cm,N_G}$  and  $\alpha_{cp,N_G}$  are the weight of communication cost and computation cost of ground nodes, respectively.  $l_{(n,m)} = \sum_{q \in Q} \sum_{\forall (i,j) \in E_q} y_{(i,j),q}^{(n,m)} l_q^{(i,j)}$  is the used channel capacity between network node  $n$  and network node  $m$ . As mentioned before, FDMA is employed in aerial nodes.  $b_{(n,m)}$  and  $p_{(n,m)}$  represent the spectrum and transmission power from aerial node  $n$  to network node  $m$ , respectively. The solution vector is denoted by  $\mathbf{b} = \{b_{(n,m)} \mid \forall n, m \in N_A\}$  and  $\mathbf{p} = \{p_{(n,m)} \mid \forall (n,m) \in E_{A1}\}$  parallelly. For the cost model of aerial nodes, the transmission power is considered. The cost of aerial nodes is expressed as

$$c_n^{UAV} = \alpha_{cm,N_A} \sum_{m \neq n} l_{(n,m)} + \alpha_p \sum_{m \neq n} p_{(n,m)} + \alpha_{cp,n_A} \sum_{q \in Q} \sum_{f \in \mathbf{f}_q} \frac{c_{f,n,q} x_{f,n,q}}{C_n}, \quad \forall n \in N_A, \quad (6)$$

where  $\alpha_{cm,N_A}$  and  $\alpha_{cp,N_A}$  denote the weight of communication cost and computation cost of aerial nodes.  $\alpha_p$  denotes the weight of the cost of transmission power, respectively. Compared with ground nodes, aerial nodes lack a continuous and sufficient source of power even equipped with solar battery. Therefore, the energy consumption of aerial nodes is expressed as

$$\Omega_n = \beta_1 \sum_{m \neq n} p_{(n,m)} + \beta_2 \sum_{q \in Q} \sum_{f \in \mathbf{f}_q} c_{f,n,q} x_{f,n,q}, \quad \forall n \in N_A, \quad (7)$$

where  $\Omega_n$  represent the consumed energy of aerial node  $n$ ,  $\beta_1$  and  $\beta_2$  are the weight of power and computation, respectively.

## D. Profit Modeling

The network's profit mainly depends on the completion of each service. When a service is completed within the required delay, the system will earn a certain revenue. The total revenue is expressed as

$$R = \sum_{q \in Q} r_q z_q, \quad (8)$$

where  $r_q$  is the revenue generates from the service  $q$ . The total cost is expressed as

$$C = \sum_{n \in N_A} c_n^{UAV} + \sum_{n \in N_G} c_n^{BS}. \quad (9)$$

The network profit is expressed as

$$P = R - C, \quad (10)$$

which is the main goal to achieve in the considered system.

## IV. PROBLEM FORMULATION

To maximize the total network profit, we formulate the joint rate-adaptive SFC orchestration and wireless resource allocation as an MINLP problem while considering the constraints of service provision, ground networks, and aerial networks.

### A. Service Provision Constraints

This subsection presents the constraints of service provision. The sources, destinations, and VNF sequences are predefined before the service arrives at the network. Constraints  $C_1$  and  $C_2$  should be satisfied to ensure that the initial VNF and final VNF are embedded in the source and destination. These two constraints are

$$C1 : x_{f_1, s_q, q} = z_q, \quad \forall q \in Q, \quad (11)$$

$$C2 : x_{f_{|\mathbf{f}_q|}, d_q, q} = z_q, \quad \forall q \in Q, \quad (12)$$

where  $f_1$  and  $f_{|\mathbf{f}_q|}$  are the first and the last VNF of service  $q$ , respectively.

For any arriving service requests that are received, each VNF of which have to be embedded on only one network node, which is expressed as

$$C3 : \sum_{n \in N} x_{f,n,q} = z_q, \quad \forall f \in \mathbf{f}_q, \forall q \in Q. \quad (13)$$

Besides, flow conservation is essential in graph routing, which guarantees the inbound flow units equal outbound flow units [34]. In this paper, the flow conservation ensures the processing sequence of the SFC and is expressed as

$$C4 : \sum_{m \in N} y_{(i,j),q}^{(n,m)} - \sum_{m \in N} y_{(i,j),q}^{(m,n)} = x_{i,n,q} - x_{j,n,q}, \quad \forall n \in N, \forall q \in Q, \forall (i,j) \in E_q. \quad (14)$$

Each service has a strict latency constraints, which if violated no revenue is generated. We assume that the delay requirement of each service cannot be violated, and every received service  $q$  should be completed within the delay constraint  $t_q$ , which is expressed as

$$C5 : \sum_{(i,j) \in E_q} z_q \frac{l_q}{l_q^{(i,j)}} + \sum_{f \in f_q} z_q \frac{c_{f,n,q}}{c_{f,q}} \leq t_q, \quad \forall q \in Q. \quad (15)$$

### B. Ground Network Constraints

Due to the limited resources of ground nodes, the allocated computation resource cannot exceed the capacity, i.e.,

$$C6 : \sum_{q \in Q} \sum_{f \in f_q} x_{f,n,q} c_{f,n,q} \leq C_n, \quad \forall n \in N_G, \quad (16)$$

where  $C_n$  is the computation capacity of network node  $n$ . Moreover, the capacity of the ground network links are fixed and limited, which results in the constraints as follows.

$$C7 : \sum_{q \in Q} \sum_{(i,j) \in E_q} y_{(i,j),q}^{(n,m)} l_q^{(i,j)} \leq l_{G1}, \forall (n,m) \in E_G, \quad (17)$$

$$C8 : \sum_{q \in Q} \sum_{(i,j) \in E_q} y_{(i,j),q}^{(n,m)} l_q^{(i,j)} \leq l_{G2}, \forall (n,m) \in E_{A1}. \quad (18)$$

### C. Aerial Network Constraints

The channel capacity, power, and available spectrum in aerial networks are constrained. The wireless channel follows the path loss model in [35], and channel capacity from aerial node  $n$  to network node  $m$  [32] is expressed as

$$\varphi^{(n,m)} = b_{(n,m)} \log \left( 1 + \frac{h_{(n,m)}^{-\gamma} p^{(n,m)}}{\sigma^2 b_{(n,m)}} \right), \quad (19)$$

where  $h_{(n,m)}$  is the distance from aerial node  $n$  to node  $m$ ,  $\gamma$  represents the constant path loss coefficient, and  $\sigma^2$  indicates the additive white Gaussian white power spectrum density.

The allocated transmission rates cannot exceed the channel capacity, which is expressed as

$$C9 : \sum_{q \in Q} \sum_{(i,j) \in E_q} y_{(i,j),q}^{(n,m)} l_q^{(i,j)} \leq \varphi^{(n,m)}, \forall (n,m) \in E_{A2}, \quad (20)$$

where the left side of the inequality is the rate of all services from aerial node  $n$  to network node  $m$ , and the right side of the inequality is the channel capacity. The transmission power of the aerial node cannot exceed the maximum power, which is expressed as

$$C10 : \sum_{m \in N} p^{(n,m)} \leq P_{max}, \forall n \in N. \quad (21)$$

where  $P_{max}$  is the maximum power of each aerial node. Furthermore, spectrum resource in the aerial network is limited, and the allocated spectrum cannot exceed the total network spectrum, which is expressed as

$$C11 : \sum_{n \in N_A} \sum_{m \in N} b_{(n,m)} \leq B. \quad (22)$$

The energy of aerial nodes is limited, which is determined by the battery capacity. Each individual aerial node that supports excessive service requests will drain power very quickly, resulting in an unstable network topology due to frequent shift work. To stabilize the network topology, load balance constraint is introduced and expressed as C12, where  $|N_A|$  is the number of aerial nodes,  $\bar{\Omega}_n$  is the average load of aerial nodes, and  $\varepsilon^2$  represents the critical value of load variance. The load variance evaluates the difference in battery consumption across aerial nodes, with a smaller value indicating a smaller load differential between aerial nodes.

### D. MINLP Problem

In this model, we optimize the rate-adaptive SFC orchestration and wireless resource to maximize the total network profit from the network operator's perspective. Combining with constraints C1-C12, an MINLP problem is formulated as follows,

$$\begin{aligned} P1 : \quad & \max_{\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{p}, \mathbf{b}, \mathbf{l}} P = R - C \\ & s.t. C1 - C12, \\ & C13 : \mathbf{x}, \mathbf{y}, \mathbf{z} \in \{0, 1\}, \\ & C14 : \mathbf{p}, \mathbf{b}, \mathbf{l} \geq 0. \end{aligned}$$

In this problem, variables  $\mathbf{x}$ ,  $\mathbf{y}$ , and  $\mathbf{z}$  are integers and variables  $\mathbf{p}$ ,  $\mathbf{b}$ , and  $\mathbf{l}$  are continuous variables. The problem is non-convex and NP-hard [36], and its optimal solution cannot be found within polynomial time. To solve this problem, we relax the integer variables and several non-convex constraints, and an iterative alternating optimization algorithm, named optimization of SFC embedding, wireless resources, and transmission rate (EM-WR-TR optimization), is presented in the next section.

## V. EM-WR-TR OPTIMIZATION

### A. Problem Transformation

Both integer variables and continuous variables exist in the proposed problem, and it is an MINLP problem. Traditional MINLP optimization algorithms like spatial branch-and-bound algorithm or Lasserre hierarchy suffers extra complexity for overwhelmed constraints and variables. Therefore, we transform the integer variables at first. Furthermore, the mutual multiplication of  $\mathbf{y}$  and  $\mathbf{l}$  exists in C9 is non-convex, and  $\mathbf{y}$  and  $\mathbf{l}$  cannot be simultaneously optimized. To jointly optimize the rate-adaptive SFC orchestration and wireless resource allocation, an altering optimization approach is proposed where  $\mathbf{x}$ ,  $\mathbf{y}$ ,  $\mathbf{z}$ ,  $\mathbf{p}$ ,  $\mathbf{b}$ , and  $\mathbf{l}$  are optimized iteratively. The details are as follows.

$$C12: \left[ \max_{n \in N_A} \left\{ \beta_1 \sum_{m \neq n} p_{(n,m)} + \beta_2 \sum_{q \in Q} \sum_{f \in f_q} x_{f,n,q} c_{f,q} \right\} - \min_{n \in N_A} \left\{ \beta_1 \sum_{m \neq n} p_{(n,m)} + \beta_2 \sum_{q \in Q} \sum_{f \in f_q} x_{f,n,q} c_{f,q} \right\} \right]^2 \leq \varepsilon^2 \quad (23)$$

Firstly, all integer variables are relaxed to continuous variables. Simply transforming all integer variables into continuous variables produces inaccuracies and errors in SFC orchestration and final results. To make up errors brought from the transformation, we relax  $z_q$  to a continuous variable following a similar approach inspired by [37], and  $z_q$  in C13 can be relaxed as

$$C13a: \sum_{q=1}^Q z_q - \sum_{q=1}^Q z_q^2 \leq 0, \forall q \in Q, \quad (24)$$

$$C13b: 0 \leq z_q \leq 1, \forall q \in Q, \quad (25)$$

where C13a is non-convex and needs to be introduced into the objective function. The other two integer variables  $\mathbf{x}$  and  $\mathbf{y}$  are relaxed as

$$C13c: 0 \leq x_{f,n,q} \leq 1, \forall q \in Q, \forall f \in f_q, \forall n \in N, \quad (26)$$

$$C13d: 0 \leq y_{(i,j),q}^{(n,m)} \leq 1, \forall q \in Q, \forall (n,m) \in E, \forall (i,j) \in E_q. \quad (27)$$

Thus, the problem is transformed into

$$P2: \max_{\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{p}, \mathbf{b}, \mathbf{l}} P = R - C + \kappa \Delta^v \quad (28)$$

$$s.t. \quad C1 - C12, C13(b-d), C14,$$

where  $\Delta^v = \sum_{q=1}^Q z_q + \sum_{q=1}^Q (z_q^v)^2 - 2 \sum_{q=1}^Q z_q^v$  is the Taylor expansion of formula C13a.  $z_q^v$  is the value of  $z_q$  in  $v_{th}$  iteration and  $\kappa$  is the weight of non-integer penalty.

In problem P2, fractional and multiplier forms coexist in C5, C7, and C8, which are non-convex and intractable. Similar fractional forms in the objective function can be solved by Dinkelbach's algorithm [38] or Charnes-cooper transformation [39]. However, constraints in this problem are intertwined. Some fractional terms will be turned into sums and others into fractions adversely, which cannot solve the problem and might even make it worse. The main variables in C5, C7, and C8 are  $\mathbf{y}$ ,  $\mathbf{z}$ , and  $\mathbf{l}$ . All fractional and multiplier forms are composed of  $\mathbf{y}$  and  $\mathbf{l}$ , or  $\mathbf{z}$  and  $\mathbf{l}$ . Inspired by [40], we find that P2 can be transformed into a conic optimization problem under fixed  $\mathbf{l}$ , which is expressed as

$$P3: \max_{\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{p}, \mathbf{b}} P = R - C + \kappa \Delta^v \quad (29)$$

$$s.t. \quad C1 - C11, C12a, C13(b-d),$$

$$\mathbf{p}, \mathbf{b} \geq 0.$$

Based on P3, the SFC embedding and wireless resources can be optimized. Then,  $\mathbf{x}^v, \mathbf{y}^v, \mathbf{z}^v, \mathbf{p}^v, \mathbf{b}^v$  are generated and ready to be used latterly. In P3,  $\mathbf{l}$  is served for delay fulfillment and is part of network cost. To further maximize the objective function of P3, the P4 is derived. Based on the results of P3, the transformation rate of each SFC can be optimized by P4, which is expressed as

$$P4: \min \sum_{q \in Q} \sum_{(n,m) \in E_G \cup E_{A1}} \sum_{(i,j) \in E_Q} \alpha_{cm, N_G} l_q^{(n,m)} y_{(i,j),q}^{(n,m)} +$$

$$\sum_{q \in Q} \sum_{(n,m) \in E_{A2}} \sum_{(i,j) \in E_Q} \alpha_{cm, N_A} l_q^{(n,m)} y_{(i,j),q}^{(n,m)},$$

$$s.t. \quad C5, C7 - C9,$$

$$l_{\max} \geq l_q^{(n,m)} \geq 0, \forall (n,m) \in E, \forall q \in Q. \quad (30)$$

In this step, the transmission rates of each service requests, e.g.,  $\mathbf{l}$ , are optimized, and the result of  $v_{th}$  iteration is denoted by  $\mathbf{l}^v$ . Let  $z_q^v = z_q^*$  and input the  $\mathbf{l} = \mathbf{l}^v$  into P3, and execute the conic programming, then,  $\mathbf{x}^{v+1}, \mathbf{y}^{v+1}, \mathbf{z}^{v+1}, \mathbf{p}^{v+1}, \mathbf{b}^{v+1}$  are generated. Then, the iterative alternating optimization can be conducted successively. Finally, the algorithm is terminated when the stopping criteria triggers. The specific process of the algorithm is as follows.

---

**Algorithm 1:** Joint SFC Orchestration and Wireless Resource Allocation Optimization.

---

**Input:** Newly arrived service  $Q$ , current network status, and available wireless resources

**Output:**  $\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{p}, \mathbf{b}, \mathbf{l}$

---

1 **repeat**

2   Begin CP with using MOSEK to solve P3;

**Output:**  $\mathbf{x}^v, \mathbf{y}^v, \mathbf{z}^v, \mathbf{p}^v, \mathbf{b}^v$

3   Set  $\mathbf{x}, \mathbf{y}, \mathbf{p}, \mathbf{b}, \mathbf{z}$  as  $\mathbf{x}^v, \mathbf{y}^v, \mathbf{z}^v, \mathbf{p}^v, \mathbf{b}^v$  parallelly;

4   Begin CP with using MOSEK to solve P4;

**Output:**  $\mathbf{l}^v$

5   Compute the revenue  $P^v$  by  $\mathbf{x}^v, \mathbf{y}^v, \mathbf{p}^v, \mathbf{b}^v, \mathbf{z}^v, \mathbf{l}^v$ ;

6    $v = v + 1$ ;

7 **until**  $P^v - P^{v-1} \leq \delta$ ;

8 Obtain  $\mathbf{x}^v, \mathbf{y}^v, \mathbf{p}^v, \mathbf{b}^v, \mathbf{z}^v, \mathbf{l}^v$  as  $\mathbf{x}, \mathbf{y}, \mathbf{p}, \mathbf{b}, \mathbf{z}$ ;

9 **End**

---

1) Initialization: Before each iteration, the type, source, destination, and the requirements of each service, and current network state have been known to the central controller. Then, we assume that the network topology is quasi-stationary during each policy decision. To optimize SFC embedding and wireless resource, the initial value of  $\mathbf{l}$  is preset. Without loss of generality, the initial value of transmission rate is set as

$$\mathbf{l}^v = \left\{ l_q^{(i,j)} | l_q^{(i,j)} = \frac{2l_q}{t - t_{\text{comp},q}} \forall q \in Q, \forall (i,j) \in E_q \right\}, \quad (31)$$

and the  $v$  is set as 1.

2) Input the initial  $\mathbf{l}^v$  into P3 with MOSEK, then the  $\mathbf{x}^v, \mathbf{y}^v, \mathbf{z}^v, \mathbf{p}^v, \mathbf{b}^v$  are obtained.  $\mathbf{z}^v$  is the decision variables for service request,  $\mathbf{x}^v, \mathbf{y}^v$  are the variables for SFC embedding, and  $\mathbf{p}^v, \mathbf{b}^v$  are the allocated power and spectrum for aerial nodes. Save the value of  $\mathbf{x}^v, \mathbf{y}^v, \mathbf{z}^v, \mathbf{p}^v, \mathbf{b}^v$  for the step-3.



3) In step-2, the SFC embedding and wireless resource are optimized jointly, and  $\mathbf{x}^v, \mathbf{y}^v, \mathbf{z}^v, \mathbf{p}^v, \mathbf{b}^v$  are obtained. Set the  $\mathbf{x}^v, \mathbf{y}^v, \mathbf{z}^v, \mathbf{p}^v, \mathbf{b}^v$  as initial value of P4, execute the interior-point solution by MOSEK, and  $\mathbf{l}$  is optimized as  $\mathbf{l}^v$ .

4) Identify whether the stopping criteria triggers, i.e., the solution of the algorithm remains unchanged within  $\delta$ . If not, turns to step 2, otherwise, turns to step 5.

5) Normalize the  $\mathbf{x}, \mathbf{y}, \mathbf{z}$  to integers and output the optimized  $\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{p}, \mathbf{b}, \mathbf{l}$ .

The details of this algorithm are shown in Algorithm 1. In the proposed algorithm, the SFC embedding and wireless resource allocation are optimized first, then, the transmission rate is optimized. After several iterations, an optimal value is obtained.

### B. Discussion

To find out the influence of wireless resources, we analyze the expectation of maximum receiving services under different bandwidths in this subsection. Without loss of generality, the arrival of service requests is evenly distributed, and a basic SFC form containing three VNFs is considered. The service arrival ratio of ground nodes and aerial nodes are  $|N_G|/|N|$  and  $|N_A|/|N|$  separately. In this subsection, the service is divided into four types by the source-destination (SD) pair, which are air nodes to air nodes (A2A), air nodes to ground nodes (A2G), ground nodes to air nodes (G2A), and ground nodes to ground nodes (G2G). Each SD pair can be divided into two subtypes by the network node that supports the second VNF, i.e., the middle node. It is tedious to discuss the above eight subcases one by one directly. We find that when the middle node is set as an aerial node under A2A and A2G, the SFC's passing channels are all wireless channels that originate from aerial nodes. Similarly, ground middle nodes under A2A and A2G and aerial middle nodes under G2A and G2G can be discussed together where half of the channels are from aerial nodes, and others are from ground nodes to aerial nodes. Additionally, ground middle nodes under G2A and G2G contain half channels from ground nodes to ground nodes. The other channels under G2A are from ground nodes to aerial nodes, and that under G2G are from ground nodes to ground nodes. Therefore, four different cases should be considered. As for the first case, the average bandwidth for aerial networks is  $B$ . Then, the average wireless channel capacity  $\bar{\varphi}$  is expressed as

$$\bar{\varphi}(B) = \frac{B}{2} \log_2 \left( 1 + \frac{2P_{\max} \bar{h}^{-\gamma}}{\sigma^2 B} \right) \quad (32)$$

where  $\bar{h}$  is the statistical distance between aerial nodes and others. Based on the formulated problem, the network is subject to computation resource and channel capacity. Considering the communication capacity constraint only, the expected maximum receiving service number is expressed as

$$\bar{q}_{\text{comm},1}(B) = \frac{\bar{\varphi}}{\bar{l}} = \frac{B}{2\bar{l}} \log_2 \left( 1 + \frac{2P_{\max} \bar{h}^{-\gamma}}{\sigma^2 B} \right), \quad (33)$$

where  $\bar{l}$  is all services' average allocated transmission rate. Considering the computation capacity constraint only, the

maximum receiving service number is  $q_{\text{comp},1} = \bar{C}_n / \bar{c}_q$  where  $\bar{C}_n$  is the average computation capacity and  $\bar{c}_q$  is the average computation requirement of each service. Combining the computation constraint and channel constraint, the expected average maximum service number is

$$q_{\text{max},1}(B) = \min \left[ \frac{\bar{C}_n}{\bar{c}_q}, \frac{B}{2\bar{l}} \log_2 \left( 1 + \frac{2P_{\max} \bar{h}^{-\gamma}}{\sigma^2 B} \right) \right], \quad (34)$$

which is the function of available bandwidth  $B$ . As  $B \rightarrow \infty$ , the channel capacity arrives at its maximum, which is expressed as

$$\begin{aligned} \bar{q}_{\text{max},1} &= \lim_{B \rightarrow \infty} \bar{q}_{\text{max}}(B) \\ &= \lim_{B \rightarrow \infty} \min \left[ \frac{B}{4\bar{l}} \log_2 \left( 1 + \frac{4P_{\max} \bar{h}^{-\gamma}}{\sigma^2 B} \right), \frac{\bar{C}_n}{\bar{c}_q} \right] \\ &= \lim_{B \rightarrow \infty} \min \left[ \frac{\sigma^2 B}{4\bar{l}P_{\max} \bar{h}^{-\gamma}} \log_2 \left( 1 + \frac{4P_{\max} \bar{h}^{-\gamma}}{\sigma^2 B} \right), \frac{P_{\max} \bar{h}^{-\gamma}}{\sigma^2 \bar{l}}, \frac{\bar{C}_n}{\bar{c}_q} \right] \\ &= \min \left[ \lim_{B \rightarrow \infty} \frac{\sigma^2 B}{4\bar{l}P_{\max} \bar{h}^{-\gamma}} \log_2 \left( 1 + \frac{4P_{\max} \bar{h}^{-\gamma}}{\sigma^2 B} \right), \frac{P_{\max} \bar{h}^{-\gamma}}{\sigma^2 \bar{l}}, \frac{\bar{C}_n}{\bar{c}_q} \right] \\ &= \min \left[ \frac{P_{\max} \bar{h}^{-\gamma}}{\sigma^2 \bar{l}} \log_2 e, \frac{\bar{C}_n}{\bar{c}_q} \right]. \end{aligned} \quad (35)$$

Adding the constraint of channel capacity from ground nodes to aerial nodes, and the average maximum receiving service number of the second case is

$$\bar{q}_{\text{max},2} = \min \left[ \frac{\bar{C}_n}{\bar{c}_q}, \frac{P_{\max} \bar{h}^{-\gamma}}{\sigma^2 \bar{l}} \log_2 e, N_G l_{G2} \right]. \quad (36)$$

Similarly, the average maximum receiving service of the third case is

$$\bar{q}_{\text{max},3} = \min \left[ \frac{\bar{C}_n}{\bar{c}_q}, N_G l_{G2}, (N_G - 1) l_{G1} \right], \quad (37)$$

where the  $(|N_G| - 1) l_{G1}$  is the capacity constraint from ground nodes to ground nodes.

The average maximum receiving service of the last case is

$$\bar{q}_{\text{max},4} = \min \left[ \frac{\bar{C}_n}{\bar{c}_q}, (N_G - 1) l_{G1} \right]. \quad (38)$$

In conclusion, the average maximum receiving service is constant when the network setting and service types are predefined. Although we did not do a dedicated simulation for this, validation can be conducted by combining the results in Fig. 4, Fig. 7, and Fig. 8.

## VI. PERFORMANCE EVALUATION

In this section, we exhibit the simulations to evaluate the proposed algorithm in terms of convergence, average network revenue, successfully serving probability, and resource consumption. The main parameters of our scenario are listed in Table II. The simulation is carried out on a computer with 3.0 GHz Intel Core i5-9500 and 16 GB RAM, and we use MATLAB 2019a with the MOSEK of CVX to solve this problem. In our models, the transmission power, channel spectrum, transmission rate, and SFC embedding are jointly optimized to maximize the network revenue. We evaluate the performance and compare it with three benchmarks as follows.

- Optimization of SFC embedding and wireless resources (EM-WR optimization): Based on P3, we optimize the

TABLE II  
SIMULATION SETTINGS

Parameter	Value
$N$	7
$N_G$	4
$N_A$	3
$B$	200 MHz
$\alpha_{int}$	250
$\alpha_p$	100000
$\alpha_{cm,a}$	$7.50 \times 10^{-06}$
$\alpha_{cm,g}$	5
$\alpha_{cp,a}$	6
$\alpha_{cp,g}$	4
$C_n$	450
$P_{max}$	0.1 W
$\sigma^2$	-174 dBm
$T_q$	10-13 evenly distributed
$r_q$	1500-2500 evenly distributed

TABLE III  
COMPUTATION TIME PER REQUESTS (SECONDS)

Number of Service	EM-WR	EM	EM-WR-TR	DE
5	0.594	0.430	1.404	2.634
10	0.367	0.287	0.910	1.339
15	0.297	0.245	0.757	0.907
20	0.244	0.226	0.677	0.691
25	0.225	0.216	0.629	0.563
30	0.213	0.210	0.596	0.476
35	0.195	0.213	0.580	0.414
40	0.222	0.207	0.563	0.367
45	0.246	0.208	0.533	0.332

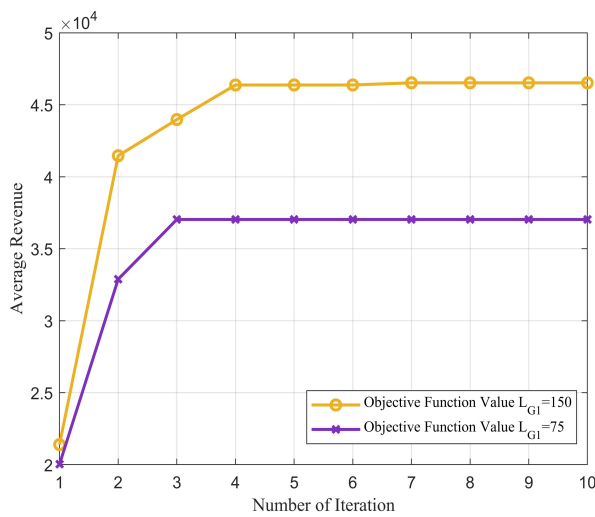


Fig. 3. Iteration and convergence analysis.

transmission power, channel spectrum, and SFC embedding jointly, and which is set as a benchmark to illustrate the advantage brought from transmission rate adaption.

- Optimization of SFC embedding (EM optimization): To illustrate the ascendancy of network reconfigurability, we only optimize the SFC embedding to simulate the traditional terrestrial network with fixed capacity.
- Differential evolution (DE) is a famous heuristic algorithm, and its superiority is demonstrated in complex optimization problems because of its simple computation processes and fewer parameters. In this paper, we employ DE as a benchmark to give an intuitive reference, and its population number is set to 150.

Fig. 3 shows the convergence evaluation of our proposed algorithm, where the service number is set to 40. From Fig. 3, it can be seen our proposed algorithm can converge in three iterations when  $l_{G1}$  equals 75 Mbps, which is faster than that when doubles the  $l_{G1}$ .

Table III shows the average computation time per request for different total numbers. It has been demonstrated that all these algorithms have the same order of magnitude. By comparing the computation time between EM-WR-TR optimization and DE, it can be seen that the proposed algorithm runs faster

than DE before the number of service requests equals 20 and is three times slower than EM-WR optimization and EM optimization. This is due to the fact that the suggested algorithm comprises two loop structures within a single iteration, resulting in a longer convergence time. In DE, the population number is fixed and the constraints are all converted to the numerical penalty in the objective function, which reduces the impact of the total request number and the average computation time.

Fig. 4 shows the comparison of four algorithms in average revenue. From Fig. 4, we can see that our proposed algorithm outperforms the benchmarks by about 10% to 50% in the performance of average revenue. As the number of service requests increases, the average revenues of four cases grow simultaneously. Whereas in a large number of service requests (beyond 30), the growth of average revenue associated with the EM-WR-TR approach the EM-WR approach is slowing down, which means the space for optimization is gradually exhausted and a platform appears. Similarly, the growth of average revenue associated with EM approach slows down earlier as the number of service requests increases to about 25. This is because the EM-WR-TR optimization triggers the allocation of wireless resources compared with EM optimization and unblocks the restriction of transmission rate between each VNF, which brings a significant degree of flexibility for service fulfillment. Unfortunately, as the number of services increases from ten, the result of DE is not as good as the other three algorithms, and its average revenue grows continually but slowly. After all, the problem is complex and comprises a large number of multi-dimensional matrix variables, which easily drives the DE to local maximization. Accordingly, we evaluate another important performance index, i.e., the successfully serving probability. Fig. 5 shows the impact of the number of service requests on the successfully serving probability where all values equal 100% at the beginning. Results of DE and EM optimization keep falling as the number of requests increases, and the latter declines faster. Particularly, our proposed approach has a more robust service reception probability than benchmarks, which verifies the efficiency of our proposed approach.

Fig. 6 shows the average resource costs of each request under different service number. Obviously, resource costs of DE are always the highest, and both DE and EM optimization are in downtrends, which reveals the flaws of DE in multi-dimensional resource allocation that redundant resources are

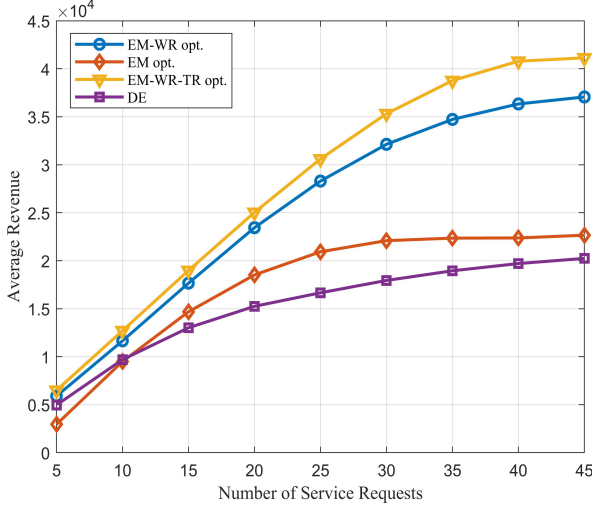


Fig. 4. The average revenue versus the number of service requests.

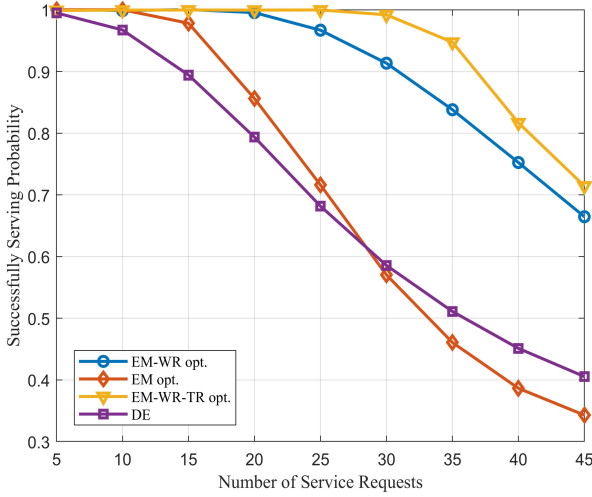


Fig. 5. The successfully serving probability versus the number of service requests.

allocated to avoid constraint violation. The average resource cost of EM-WR optimization is higher than that of EM-WR-TR optimization, and they both continue to increase with the number of service requests until they are nearly identical. Obviously, the average resource costs of EM-WR-TR optimization are lowest when the service number is less than 25, which indicates that EM-WR-TR optimization can receive as many services as possible while minimizing resource consumption. Interestingly, resource costs of EM-WR optimization and EM-WR-TR optimization increase slowly and approach the same value. This is because network operators must utilize expensive resources to obtain more revenues, which reduces the average resource utilization.

Fig. 7 shows the comparison of EM-WR-TR optimization with varied ratio of service types under different available spectrums. We can observe that the average revenue increases with  $\phi$  (before 0.8), which is because that our proposed algorithm is good at scheduling communication resources. As the  $\phi$  increases to 0.8, average revenue begins to fall because of the limitation of transmission resource.

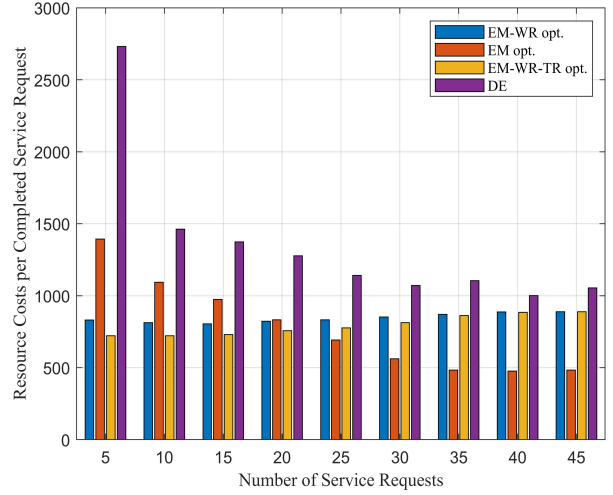


Fig. 6. The resource costs per completed service request versus the number of service requests.

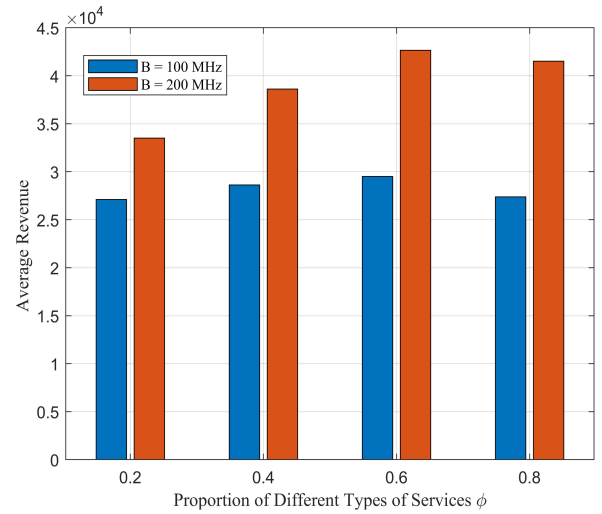


Fig. 7. Comparison of successfully serving probability with the varied ratio of each type of service.

Fig. 8 and Fig. 9 compare four algorithms with varied available spectrums in terms of the average revenue and successfully serving probability, respectively. The number of service requests is set as 40. Fig. 8 and Fig. 9 show that our proposed EM-WR-TR optimization algorithm outperforms the other three algorithms. This is because the proposed algorithm enables both wireless resource allocation and rate-adaptive SFC orchestration. Compared with the EM-WR optimization, the proposed algorithm can minimize resource costs and receive more services. As the available spectrum increases, the average revenue and successfully serving probability increase simultaneously. Two performance measurements in Fig. 8 and Fig. 9 reveal that the DE is only efficient when the action space is small and its performance increases gradually as the available spectrum diminishes.

## VII. CONCLUSION

In this paper, we have proposed an SDN/NFV-based reconfigurable SAGIN network architecture, and based on which,

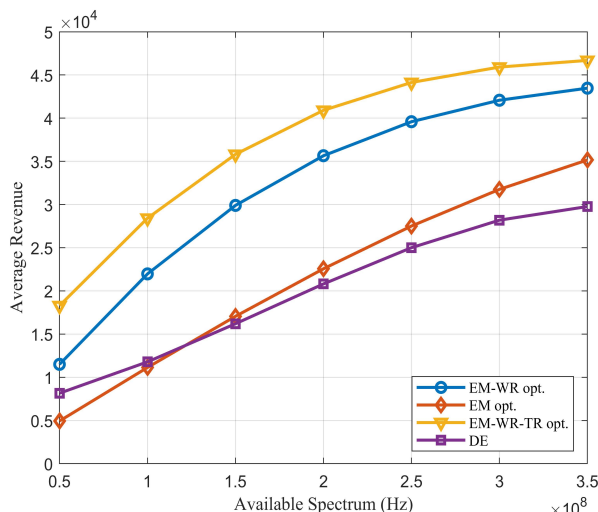


Fig. 8. The average revenue versus the available spectrum.

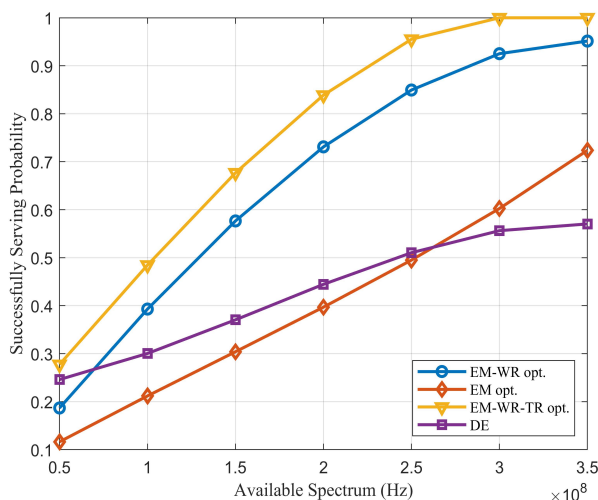


Fig. 9. Comparison of successfully serving probability versus the available spectrum.

rate-adaptive SFC orchestration and wireless resource allocation are investigated comprehensively. Considering the resource limitation of network infrastructures and service requirements, an MINLP problem has been formulated to maximize the network profit. Then, successive convex optimization is utilized to transform the proposed problem into a tractable one, and an iterative altering algorithm is proposed to optimize the SFC embedding, transmission rate, and wireless resource jointly. Extensive simulations have been carried out, and the results have illustrated the effectiveness of the proposed algorithm in SFC orchestration and resource allocation. Specifically, the EM-WR approach achieves a lower average computation time than others and is effective in highly dynamic network scenarios. The proposed architecture and EM-WR-TR approach lay a foundation to future studies related to on-demand service provision and wireless resource scheduling in SAGINs. In future work, we will investigate the service-oriented mobile user access and handover in SAGINs deeply.

## REFERENCES

- [1] M. Agiwal, A. Roy, and N. Saxena, "Next generation 5G wireless networks: A comprehensive survey," *IEEE Communications Surveys and Tutorials*, vol. 18, no. 3, pp. 1617–1655, 2016.
- [2] J. Kang, Z. Xiong, D. Niyato, D. Ye, D. I. Kim, and J. Zhao, "Toward secure blockchain-enabled internet of vehicles: Optimizing consensus management using reputation and contract theory," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 3, pp. 2906–2920, 2019.
- [3] T. Ma, H. Zhou, B. Qian, N. Cheng, X. Shen, X. Chen, and B. Bai, "UAV-LEO Integrated Backbone: A Ubiquitous Data Collection Approach for B5G Internet of Remote Things Networks," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 11, pp. 3491–3505, 2021.
- [4] P. Zhang, P. Yang, N. Kumar, and M. Guizani, "Space-Air-Ground Integrated Network Resource Allocation Based on Service Function Chain," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 7, pp. 7730–7738, 2022.
- [5] J. Hu, C. Chen, L. Cai, M. R. Khosravi, Q. Pei, and S. Wan, "UAV-assisted vehicular edge computing for the 6G internet of vehicles: architecture, intelligence, and challenges," *IEEE Communications Standards Magazine*, vol. 5, no. 2, pp. 12–18, 2021.
- [6] L. Shen, N. Wang, Z. Zhu, W. Xu, Y. Li, X. Mu, and L. Cai, "UAV-Enabled Data Collection Over Clustered Machine-Type Communication Networks: AEM Modeling and Trajectory Planning," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 9, pp. 10016–10032, 2022.
- [7] Z. Xiong, Y. Zhang, W. Y. B. Lim, J. Kang, D. Niyato, C. Leung, and C. Miao, "UAV-assisted wireless energy and data transfer with deep reinforcement learning," *IEEE Transactions on Cognitive Communications and Networking*, vol. 7, no. 1, pp. 85–99, 2020.
- [8] ITU, "HAPS – High-altitude platform systems," Tech. Rep., 2019.
- [9] H. A. Hou and L. C. Wang, "Analysis on time-variant air-to-ground radio communication channel for rotary-wing UAVs," in *Proceedings - 2019 IEEE 89th Vehicular Technology Conference (VTC2019-Spring)*, Malaysia, 2019, pp. 1–6.
- [10] Y. Wang, Z. Su, J. Ni, N. Zhang, and X. Shen, "Blockchain-Empowered Space-Air-Ground Integrated Networks: Opportunities, Challenges, and Solutions," *IEEE Communications Surveys and Tutorials*, vol. 24, no. 1, pp. 160–209, 2022.
- [11] A. Liao, Z. Gao, D. Wang, H. Wang, H. Yin, D. W. K. Ng, and M. S. Alouini, "Terahertz Ultra-Massive MIMO-Based Aeronautical Communications in Space-Air-Ground Integrated Networks," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 6, pp. 1741–1767, 2021.
- [12] Z. Zhou, J. Feng, C. Zhang, Z. Chang, Y. Zhang, and K. M. S. Huq, "SAGECELL: Software-defined space-air-ground integrated moving cells," *IEEE Communications Magazine*, vol. 56, no. 8, pp. 92–99, 2018.
- [13] G. Wang, S. Zhou, S. Zhang, Z. Niu, and X. Shen, "SFC-Based service provisioning for reconfigurable space-air-ground integrated networks," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 7, pp. 1478–1489, 2020.
- [14] M. M. Tajiki, S. Salsano, L. Chiaraviglio, M. Shojafar, and B. Akbari, "Joint Energy Efficient and QoS-Aware Path Allocation and VNF Placement for Service Function Chaining," *IEEE Transactions on Network and Service Management*, vol. 16, no. 1, pp. 374–388, 2019.
- [15] J. Pei, P. Hong, K. Xue, and D. Li, "Efficiently Embedding Service Function Chains with Dynamic Virtual Network Function Placement in Geo-Distributed Cloud System," *IEEE Transactions on Parallel and Distributed Systems*, vol. 30, no. 10, pp. 2179–2192, 2019.
- [16] Y. Yue, B. Cheng, X. Liu, M. Wang, S. Member, B. Li, and J. Chen, "Resource Optimization and Delay Guarantee Virtual Network Function Placement for Mapping SFC Requests in Cloud Networks," *IEEE Transactions on Network and Service Management*, vol. 18, no. 2, pp. 1508–1523, 2021.
- [17] A. Varasteh, B. Madiwalar, A. Van Bemten, W. Kellerer, and C. Mas-Machuca, "Holu: Power-Aware and Delay-Constrained VNF Placement and Chaining," *IEEE Transactions on Network and Service Management*, vol. 18, no. 2, pp. 1524–1539, 2021.
- [18] Y. Wang, C. K. Huang, S. H. Shen, and G. M. Chiu, "Adaptive Placement and Routing for Service Function Chains with Service Deadlines," *IEEE Transactions on Network and Service Management*, vol. 18, no. 3, pp. 3021–3036, 2021.
- [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, 2019.

- [20] N. Siasi, M. A. Jasim, A. Yayimli, and N. Ghani, "Service Function Chain Survivability Provisioning in Fog Networks," *IEEE Transactions on Network and Service Management*, vol. 19, no. 2, pp. 1117–1128, 2022.
- [21] R. Shi, J. Zhang, W. Chu, Q. Bao, X. Jin, C. Gong, Q. Zhu, C. Yu, and S. Rosenberg, "MDP and Machine Learning-Based Cost-Optimization of Dynamic Resource Allocation for Network Function Virtualization," in *Proceedings - 2015 IEEE International Conference on Services Computing*, USA, 2015, pp. 65–73.
- [22] F. Wei, G. Feng, Y. Sun, Y. Wang, Q. Shuang, and Y.-C. Liang, "Network Slice Reconfiguration by Exploiting Deep Reinforcement Learning With Large Action Space," *IEEE Transactions on Network and Service Management*, vol. 17, no. 4, pp. 2197–2211, 2020.
- [23] D. N. Heo, S. Lange, H. G. Kim, and H. Choi, "Graph neural network based service function chaining for automatic network control," in *Proceedings - 2020 21st Asia-Pacific Network Operations and Management Symposium: Towards Service and Networking Intelligence for Humanity*, South Korea, 2020, pp. 7–12.
- [24] Y. Liu, Y. Lu, X. Li, Z. Yao, and D. Zhao, "On Dynamic Service Function Chain Reconfiguration in IoT Networks," *IEEE Internet of Things Journal*, vol. 7, no. 11, pp. 10969–10984, 2020.
- [25] H. G. Kim, S. Park, S. Lange, D. Lee, D. Heo, H. Choi, J. H. Yoo, and J. W. K. Hong, "Graph neural network-based virtual network function deployment optimization," *International Journal of Network Management*, vol. 31, no. 6, pp. 2164–2170, 2021.
- [26] B. Li and Z. Zhu, "GNN-Based Hierarchical Deep Reinforcement Learning for NFV-Oriented Online Resource Orchestration in Elastic Optical DCIs," *Journal of Lightwave Technology*, vol. 40, no. 4, pp. 935–946, 2022.
- [27] N. Cheng, F. Lyu, W. Quan, C. Zhou, H. He, W. Shi, and X. Shen, "Space/Aerial-Assisted Computing Offloading for IoT Applications: A Learning-Based Approach," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 5, pp. 1117–1129, 2019.
- [28] X. Shen, J. I. E. Gao, W. E. N. Wu, K. Lyu, M. Li, W. Zhuang, X. U. Li, and J. Rao, "AI-Assisted Network-Slicing Based Next-Generation Wireless Networks," *IEEE Open Journal of Vehicular Technology*, vol. 1, no. December 2019, pp. 45–66, 2020.
- [29] H. Hantouti, N. Benamar, and T. Taleb, "VLAN-based Traffic Steering for Hierarchical Service Function Chaining," in *IEEE Wireless Communications and Networking Conference*, vol. 21, no. 1, Morocco, 2019, pp. 487–507.
- [30] S. Zhou, G. Wang, S. Zhang, Z. Niu, and X. S. Shen, "Bidirectional mission offloading for agile space-air-ground integrated networks," *IEEE Wireless Communications*, vol. 26, no. 2, pp. 38–45, 2019.
- [31] J. Li, W. Shi, H. Wu, S. Zhang, and X. Shen, "Cost-Aware Dynamic SFC Mapping and Scheduling in SDN/NFV-Enabled Space-Air-Ground Integrated Networks for Internet of Vehicles," *IEEE Internet of Things Journal*, vol. 9, no. 8, p. 5824–5838, 2021.
- [32] A. Goldsmith, *Wireless Communication*. Cambridge university press, 2005.
- [33] B. Farkiani, B. Bakhshi, S. A. Mirhassani, T. Wauters, B. Volckaert, and F. De Turck, "Prioritized Deployment of Dynamic Service Function Chains," *IEEE/ACM Transactions on Networking*, vol. 29, no. 3, pp. 979–993, 2021.
- [34] P. D. Thomas and C. K. Lombard, "Geometric Conservation Law and Its Application to Flow Computations on Moving Grids," *Journal of American Institute of Aeronautics and Astronautics*, vol. 17, no. 10, pp. 1030–1037, 2012.
- [35] Y. Zeng and R. Zhang, "Energy-Efficient UAV Communication with Trajectory Optimization," *IEEE Transactions on Wireless Communications*, vol. 16, no. 6, pp. 3747–3760, 2017.
- [36] J. Chen, Q. Wu, Y. Xu, N. Qi, T. Fang, and D. Liu, "Spectrum Allocation for Task-Driven UAV Communication Networks Exploiting Game Theory," *IEEE Wireless Communications*, vol. 28, no. 4, pp. 174–181, 2021.
- [37] Z. Fei, Y. Wang, R. Sun, and Y. Liu, "Delay-oriented task scheduling and bandwidth allocation in fog computing networks," in *Proceedings - 2019 IEEE Global Communications Conference*, USA, 2019, pp. 1–6.
- [38] G. Scutari, F. Facchinei, L. Lampariello, S. Sardellitti, and P. Song, "Parallel and Distributed Methods for Constrained Nonconvex Optimization-Part II: Applications in Communications and Machine Learning," *IEEE Transactions on Signal Processing*, vol. 65, no. 8, pp. 1945–1960, 2017.
- [39] A. Charnes, W. W. Cooper, Q. L. Wei, and Z. M. Huang, "Cone ratio data envelopment analysis and multi-objective programming," *International Journal of Systems Science*, vol. 20, no. 7, pp. 1099–1118, 1989.
- [40] Z. Yin, M. Jia, N. Cheng, W. Wang, F. Lyu, Q. Guo, and X. Shen, "UAV-Assisted Physical Layer Security in Multi-Beam Satellite-Enabled Vehi-

cle Communications," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 3, pp. 2739–2751, 2022.