

# Computing-Aware Routing for LEO Satellite Networks: A Transmission and Computation Integration Approach

Jiaqi Cao, Shengli Zhang, *Senior Member, IEEE*,  
Qingxia Chen, Houtian Wang, Mingzhe Wang, Naijin Liu

**Abstract**—The advancements of remote sensing (RS) pose increasingly high demands on computation and transmission resources. Conventional ground-offloading techniques, which transmit large amounts of raw data to the ground, suffer from poor satellite-to-ground link quality. In addition, existing satellite-offloading techniques, which offload computational tasks to low earth orbit (LEO) satellites located within the visible range of RS satellites for processing, cannot leverage the full computing capability of the network because the computational resources of visible LEO satellites are limited. This situation is even worse in hotspot areas.

In this paper, for efficient offloading via LEO satellite networks, we propose a novel computing-aware routing scheme. It fuses the transmission and computation processes and optimizes the overall delay of both. Specifically, we first model the LEO satellite network as a snapshot-free dynamic network, whose nodes and edges both have time-varying weights. By utilizing time-varying network parameters to characterize the network dynamics, the proposed method establishes a continuous-time model which scales well on large networks and improves the accuracy. Next, we propose a computing-aware routing scheme following the model. It processes tasks during the routing process instead of offloading raw data to ground stations, reducing the overall delay and avoiding network congestion consequently. Finally, we formulate the computing-aware routing problem in the dynamic network as a combination of multiple dynamic single source shortest path (DSSSP) problems and propose a genetic algorithm (GA) based method to approximate the results in a reasonable time. Simulation results show that the computing-aware routing scheme decreases the overall delay by 78.31% compared with offloading raw data to the ground to process when the computing capability is 100 Giga floating-point operations per second (GFLOPS) which is a trivial computing capability supported by most LEO satellites.

**Index Terms**—LEO satellite network, remote sensing, computing-aware routing, dynamic network, genetic algorithm.

## I. INTRODUCTION

**R**EMOTE sensing (RS) is playing an increasingly important role in Earth science, space science, and exploration science, such as environmental studies, military applications, hazard tracking and monitoring [1]. To complete such space missions, resources in both computation and data transmission are needed. On the one hand, the advancements in image processing and target recognition techniques, especially the

application of machine learning techniques [2], have led to a rapid increase in computational requirements [3], [4]. On the other hand, sensing technology improvements, such as hyperspectral image (HSI) [5], enable increased data precision at the cost of a huge data volume of remote sensing images. Conventionally, these images are offloaded to ground servers for computation.

In the aforementioned ground-offloading approach, satellites act as bent pipes to route massive raw data of RS tasks to ground servers for processing. While the ground servers are powerful in computing capability, the overall delay<sup>1</sup> of the ground-offloading approach is still hard to meet the requirements of most RS applications which require real-time or near real-time processing capabilities [6]. It is because transmitting raw data can be a significant bottleneck: perturbed by the atmosphere frequently [7], the satellite-to-ground link (SGL) can be as low as 20 Mbps in state-of-the-art satellites [8]. To overcome the limitations of the ground offloading scheme, researchers have set their sights on onboard computing.

One promising target to offload computational tasks is low earth orbit (LEO) satellites. Thanks to the development of LEO satellite computation capabilities, high-performance onboard computing provided by a large number of LEO satellites could alleviate these challenges by processing data before transmitting them to ground servers [9]. This scheme can usually achieve impressive performance compared with ground offloading for the following reasons. First, LEO satellites are geographically closer to RS satellites than ground servers, avoiding transmitting large amounts of raw data on SGLs whose data rate are relatively low. Second, onboard processing significantly reduces the resulting data's volume (down to a few bits sometimes), which lowers the transmission delay in turn. Despite the promising future of LEO satellite networks based offloading, major challenges need to be resolved first.

**Challenge 1: Uneven Available-Resource Distribution.** More than three quarters of the Earth's surface is covered by oceans and glaciers with no frequent human activity; instead, a large number of tasks generated by human activity are located on land, especially in hot spots (such as cities and ports). As the distribution of offloaded tasks are unbalanced, available resources including computing and spectrum resources are also unevenly distributed on the LEO satellite network [10], [11]. In this condition, LEO satellites nearby may not have

Naijin Liu is the corresponding author.

Jiaqi Cao and Shengli Zhang are with Shenzhen University, Shenzhen, 518052, P.R. China (e-mail: jiaqicao@szu.edu.cn; zsl@szu.edu.cn).

Qingxia Chen, Houtian Wang and Naijin Liu are with Qian Xuesen Laboratory of Space Technology, China Academy of Space Technology (e-mail: chenqingxia@qxslab.cn; wanghoutian@qxslab.cn; liunaijin@qxslab.cn).

Mingzhe Wang is with Tsinghua University (e-mail: wmzhere@gmail.com).

<sup>1</sup>The overall delay refers to the moment from a task is generated to the moment when the destination obtains the computing results of that task. For a computation task, *both* transmission and computation processes affect the overall delay.

sufficient computational resources. As a result, computational tasks generated by remote sensing satellites need to be routed to farther satellites with sufficient computational resources for processing. Frequently, the distance can be long enough so that multiple hops are required to reach LEO satellites. Therefore, a routing strategy is needed to transmit the tasks generated in hot spots to remote LEO satellites for computing.

**Challenge 2: Inapplicable Terrestrial Routing Strategy.** Existing terrestrial network routing strategies focus on static networks and commonly employ shortest path algorithms such as Dijkstra to find viable paths. However, such algorithms cannot be applied to LEO satellite networks because contemporaneous paths sometimes do not exist between the source and destination [12]. The phenomenon can be explained by the physical nature of LEO satellites: the high-speed relative motions between adjacent-orbiting satellites combined with the limited communication distance cause intermittent connectivity in LEO satellite networks. Consequently, a routing strategy specific to LEO satellite networks is needed to accommodate its innate physical characteristics.

**Challenge 3: Unaccounted Computation Cost.** Previous routing strategies for LEO satellite networks focus on data transmission only. However, the transmission process and the computation process jointly determine the overall delay (i.e., the time from task generation to the arrival of computation results at the destination) of a computational task. Existing routing strategies can only find the shortest path from a remote sensing satellite to a given LEO satellite used for computing, but cannot evaluate and select appropriate LEO satellites to conduct computation automatically. Therefore, the satellite to be used for computation is still unanswered.

**Our Solution.** For efficient task offloading via LEO satellite networks, we propose a novel computing-aware routing scheme to minimize the overall delay. As shown in Fig. 1, the computing-aware routing problem contains three subprocesses: route raw data of tasks from the source to the selected computing node; process tasks and generate computing results at the computing node; route the computing results from the computing node to the destination. Since the above transmission and computation processes affect the overall delay collaboratively, these processes are jointly optimized in the proposed computing-aware routing scheme. For challenge 1, the introduction of routing extends the task offloading targets, so that satellites beyond the visible range can be scheduled for computation. For challenge 2, the scheme models the network as a dynamic system, thus the proposed routing algorithm can tolerate the fast change of available resources and network topology by design. For challenge 3, the joint optimization algorithm minimizes the overall delay of both the transmission and computation stages; in other words, it additionally considers the computation delay compared to existing routing algorithms.

#### A. Main Contributions

1) **A Snapshot-Free Dynamic Network Model:** We propose a snapshot-free dynamic network modeling method for LEO satellite networks for cross-time pathfinding with low memory

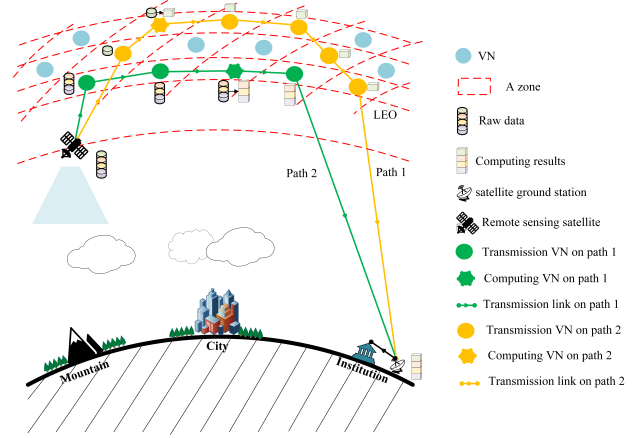


Fig. 1. Example application of the computing-aware routing problem. Remote sensing satellites continuously capture hyperspectral images of the surveillance area. The institution on the ground needs to obtain the analysis results of these images. scheme.

consumption. It can represent both resource dynamics and topology dynamics. It utilizes time-varying edge weights and node weights to represent the *resource dynamics* related to the transmission and computation processes, respectively. In addition, instead of shielding the dynamics, the proposed model converts the *topology dynamics* into the association dynamics between satellites and virtual nodes (VNs), which represents self-loops, special edge and node weights.

2) **A Computing-Aware Routing Scheme:** We propose a novel routing scheme for LEO satellite networks, which processes tasks during the routing process. By performing onboard computing, the proposed scheme can achieve significant bandwidth savings especially for satellite-ground links. Because the slow process of offloading raw data to ground servers is avoided, the proposed routing scheme can reduce the overall delay. This scheme optimizes the computation and transmission processes jointly, therefore, tasks can be offloaded to the optimal satellite via the optimal path. Furthermore, any satellite with sufficient resources can be selected as the offloading target. Since LEO constellations, especially for giant LEO constellations, usually consist of a large number of satellites, a large amount of computing resources could be utilized for computing-aware routing.

3) **A Genetic Algorithm Based Approximation Method:** Since the LEO satellite network is highly dynamic and the on-board resources (related to node weights) need to be considered in the computing-aware routing, we formulate the LEO satellite network computing-aware routing problem based on the proposed edge-weighted and node-weighted dynamic network and convert it to a set of dynamic single source shortest path (DSSSP) problems. Due to the dynamics in graphs, conventional shortest path algorithms such as the static Dijkstra algorithm cannot be used to solve the problem; thus, we propose a genetic algorithm (GA) based method to approximate the results in reasonable time.

The rest of this paper is organized as follows. In Section II, the related studies are summarized. Section III investigates the state-of-the-art computing and transmission capabilities of LEO satellites. Section IV presents the network model, traffic

model, and delay model adopted in this paper. Section III proposed a snapshot-free dynamic network modeling method. The computing-aware routing problem is formulated in Section VI. In Section VII, a GA-based path finding algorithm is proposed. Simulation results and analyses are given in Section VIII. Conclusions are drawn in Section IX.

## II. RELATED WORK

Due to the superiority in latency, cost, development cycle, etc., the LEO satellite network is deemed as the most prospective satellite mobile communication system. Therefore, a lot of studies on LEO satellites and LEO satellite networks have been conducted in academia and industry. Several aspects relevant to this paper are introduced below.

### A. Routing Strategies for LEO Satellite Networks

The ever changing relative positions of satellites bring constant changes in network topology; even worse, for specific instants, no contemporaneous path exists between the source and destination nodes [12]. Therefore, routing is a challenging issue in LEO satellite networks.

Graph theory is an effective mathematical tool to model the network, providing a basis for the routing design [21]. Therefore, graph-based routing strategies for LEO satellite networks have attracted widespread attention, which are summarized in Table I. *Contact graph routing (CGR)* was proposed for dynamic routing over the time-varying topology of satellite networks by NASA [16]. The basic idea of CGR is to utilize a scheduled contact graph for pathfinding. The *contact graph (CG)* records the network dynamics with a “contact plan” which is a time-ordered list of scheduled changes of network topology [22]. However, they could not ensure the mission’s demands and could not fully utilize the resources [20].

To deal with the dynamics in satellite networks, some existing works proposed snapshot-based network modeling methods, such as the *temporal graph (TG)*, the *virtual topology (VT)* model and the *VN* model, and the corresponding routing strategies. The basic idea of TG [13] is to divide the system period into a set of discrete slots. In this way, the dynamic network can be represented as a set of static topology graphs. Similarly, VT-based network models [14], [23], [24], [25] are presented based on the determinism of satellite movements, representing a satellite network as a time-evolving and predictable network [26]. It considers a LEO satellite network as a discrete-time network, and assumes a fixed topology in each time interval [27], which is called a

snapshot. Routes are defined at each snapshot by using this method. VN-based network models [15], [28] are composed of different logical locations, which are static and disjoint zones of Earth (i.e., latitude and longitude), associated with the nearest satellites [29]. The assignment between the logical locations and satellites changes due to satellite movements. With this architecture, each change on the satellite assignment represents a new snapshot [30]. Each snapshot could be considered as a mesh network presenting a static state of the network topology. These methods split the dynamic network into multiple snapshots, where each snapshot corresponded to a constant network topology during a time slot [27]. Obviously, these methods only look for paths within a single snapshot and ignore the relationship between adjacent snapshots [17]. A snapshot is defunct if the duration of the generated route exceeds the snapshot’s valid time. Transmitting tasks based on defunct routes may cause routing failures, especially for highly dynamic LEO satellite networks. Furthermore, as shown in Table I, these snapshot-based models would consume massive memory resources because a prohibitively large number of snapshots would be generated as the time increases or the network expands [31], [30].

To overcome the drawbacks of the snapshot-based routing strategies, the *time-expanded graph (TEG)* [17], [18], [32], [33], [34] was proposed to establish connections between networks of adjacent slots. It duplicates the original network for each time slot and builds edges connecting each node and its copy at the next slot to represent the data storage. Indeed, TEGs are essentially an expansion of static graphs, and hence many standard flow maximization algorithms can be applied to time-expanded graphs [31]. Although the TEG significantly increases the connection of snapshots [35], it also incurs high overhead in storage and algorithm [31].

The *time aggregated graph (TAG)* [19], [36] aggregates the time-dependent attributes over edges and nodes. It represents the time-variance of attributes by modeling them as time series. To consider the buffer size constraint of each relay node in TAG, authors in [20], [31], [37] proposed the *storage time aggregated graph (STAG)*. Although the TAG and STAG models could capture the possibility of edges and nodes being absent during certain instants of time [38], these methods still face the problem of edge explosion when modeling highly dynamic networks, such as giant LEO constellations. More specifically, when the periods (altitudes) of the adjacent orbits of the LEO constellation are different, any two satellites in adjacent orbits may establish inter-satellite link (ISL) within a certain period of time. In this condition, when constructing

TABLE I  
COMPARISON OF GRAPH-BASED SATELLITE NETWORK ROUTING

Reference	Network Modeling	Classification	Dynamics Representation		Graph/Model Size			Main Shortcomings
			Topology	Resource	Snapshot Number	Node Number	Edge Number	
[13]	TG	Snapshot model	✓		$\propto N$	$O( S )$	$O( S )$	Isolated snapshots split the connectivity of the whole network and consume massive memory
[14]	VT		✓		$\propto  S $	$O( S )$	$O( S )$	
[15]	VN		✓		$\propto ( S ,  Z )$	$O( Z )$	$O( Z )$	
[16]	CG	Non-snapshot model	✓		N/A	$O( S )$	$O( S )$	No assurance for task demands and low resource utilization High storage overhead and computational complexity Excessive model size for highly dynamic networks
[17], [18]	TEG		✓	✓	N/A	$O(N \times  S )$	$O(N \times  S )$	
[19], [20]	TAG/STAG		✓	✓	N/A	$O( S )$	$O( S ^2)$	
Our work	SFDNM		✓	✓	N/A	$O( Z )$	$O( Z )$	

$\propto$ : Proportional to,  $\propto$ : related to.  
 $|S|$ : Satellite number,  $|Z|$ : zone (i.e., VN) number,  $N$ : slot number.

TAGs or STAGs, any two satellites in adjacent orbits should be connected with an edge. Consequently, the edge numbers in these graphs are quadratic to satellite numbers.

To overcome the shortcomings of these existing network modeling methods and routing strategies for the LEO satellite network, we propose a novel dynamic network model which can represent both resource dynamics and topology dynamics with low complexity and the corresponding computing-aware routing scheme.

### B. Computing and Transmission Joint Optimization

Some existing works investigated the computation and transmission joint optimization for satellite networks. In these works, satellites are not connected with each other. The work in [39] and [40] investigated the joint computation assignment and resource allocation problem in multi-tier computing architectures composed of mobile devices, LEO satellites, etc. Authors in [41] and [42] proposed hybrid computation offloading architectures to solve the joint computation and resource allocation problem, where computing tasks could be offloaded to both ground servers and visible LEO satellites.

Some studies discussed the computation and transmission integration problem in terrestrial networks. The work in [43] discussed the joint communication and computing resource allocation in a two-tier device–cloud network, where tasks could be processed locally, in the edge cloud, or both. Authors in [44] and [45] investigated the task offloading problem in fog-enabled cellular networks where radio, caching, and computing were jointly optimized. The work in [46] and [47] proposed joint communication and computing resource scheduling approaches for unmanned aerial vehicle (UAV)-assisted local–edge/local–edge–cloud computing systems, where each UAV worked as an edge computing devices to assist devices within its communicable range. Authors in [48] developed a cloud-fog-device computing architecture for internet of things (IoT), where the offloading ratio, transmission power, and local CPU computation speed were jointly optimized.

Although the studies mentioned above jointly optimize the allocation of multiple resources, they still fail to achieve network-wide computation offloading. This is because the networks in the above studies are *tree networks*, where tasks cannot be forwarded to other computing devices in the same network tier. To overcome this limitation, LEO satellites in this paper are connected with ISLs, which form a *mesh network*. In this condition, computing tasks can be offloaded to any satellite via routing, which makes it possible to extend the offloading targets to the entire network. However, the network-wide computation offloading for LEO satellites brings a novel challenge: the transmission path needs to be optimized as well. To address this challenge, we propose a computing-aware routing scheme to jointly optimize the resources and transmission paths.

## III. STATE-OF-THE-ART LEO SATELLITE CAPABILITIES

As the computational requirements and data volume of space missions increase, unprecedented interest and efforts have been devoted to enhancing the computing and transmission capabilities of LEO satellites.

### A. Computation Capability of LEO Satellites

For next-generation science and defense missions, spacecrafts such as LEO satellites must provide advanced processing capability to support a variety of computationally intensive tasks [49]. The desire for even more onboard processing capacity has led to the development of onboard computing systems. The computing capabilities of some typical onboard computing systems are summarized in Table II.

TABLE II  
COMPUTING CAPABILITIES OF ONBOARD COMPUTING SYSTEMS (IN GFLOPS)

Product	Processor	Computing Capability	Reference
Xiphos Q7S	Xilinx Zynq 7020	180	[50]
Xiphos Q8S	Xilinx Ultrascale+	1800	[51]
BAE RAD5545	RAD5545	3.7	[52]
Innoflight CFC-500	Xilinx Kintex Ultrascale+, NVIDIA TK1	1290	[53]
MOOG G-Series Steppe Eagle	AMD G-Series compatible	75	[54]
MOOG V-Series Ryzen	AMD V-Series compatible	1000	[54]
Unibap iX5-100	Microchip SmartFusion2, AMD G-Series SOC	127	[55], [56]
Unibap iX10-100	Microchip PolarFire, AMD V1605b (Ryzen)	3600	[57], [56]
SpaceCube v2.0	Xilinx Virtex 5	200	[58]
SpaceCube v3.0	Xilinx Kintex UltraScale, Xilinx Zynq MPSoC	590	[49], [59]

It can be concluded from Table II that existing onboard computing systems can provide thousands Giga floating-point operations per second (GFLOPS) of computing capability. For example, the national aeronautics and space administration (NASA) Goddard Space Flight Center (GSFC) developed SpaceCube v3.0 in 2019 [60]. It contains a Xilinx Kintex UltraScale with a Xilinx Zynq MPSoC to provide 10–100x or more performance over other flight single-board computers [49]. In specific, the computing capacities of these systems on chips (SoCs) both exceed 100 GFLOPS. Although the computing capability of LEO satellites is not yet comparable to that of geosynchronous equatorial orbit (GEO) satellites and ground servers, the prospect and importance of increasing the computing capability of LEO satellites has been recognized and a great deal of research has been invested, indicating a promising future for onboard computing.

### B. Data Rate of LEO Satellites

ISLs in free space are usually higher in data rate. For instance, the data rate of optical ISLs can achieve 5 Gbps [61]. Mynaric’s laser terminal for LEO constellations is capable of delivering 10 Gbps with a low SWaP unit over a wide range of constellation configurations [62]. It can operate within densely packed constellations with intra/inter-plane link distances up to 7,800 km.

In contrast, the data rate of the satellite-to-ground link cannot keep up with the speed of the inter-satellite link due to the perturbation induced by the atmosphere [7]. The downlink data rate for state-of-the-art satellites ranges from 20 Mbps to 1 Gbps [8]. For example, the CubeSat lasercom module by Hyperion Technologies enables a bidirectional space-to-ground communication link between a CubeSat and an optical ground station, with a downlink speed up to 1 Gbps and an uplink data rate of 200 Kbps [8]. The limited SGL

transmission capability further promotes the application of on-board computing.

#### IV. SYSTEM MODEL

##### A. Network Model

As shown in Fig. 2 (a), the LEO satellites are uniformly distributed over orbits at an altitude of  $h$  kilometers. The satellites on the same orbit are uniformly distributed. The set of satellites is denoted by  $\mathbf{S} = \{S_1, S_2, \dots, S_{|\mathbf{S}|}\}$  and the set of orbits is denoted by  $\mathbf{O} = \{O_1, O_2, \dots, O_{|\mathbf{O}|}\}$ . The orbit inclination  $i_0$  determines the latitude of coverage. The orbital period is  $T_O = 2 \times \pi \times \sqrt{(R_e + h)^3 / (G \times M_e)}$ , where  $R_e$  and  $M_e$  represent the radius and mass of the earth respectively, and  $G$  is the gravitational constant.

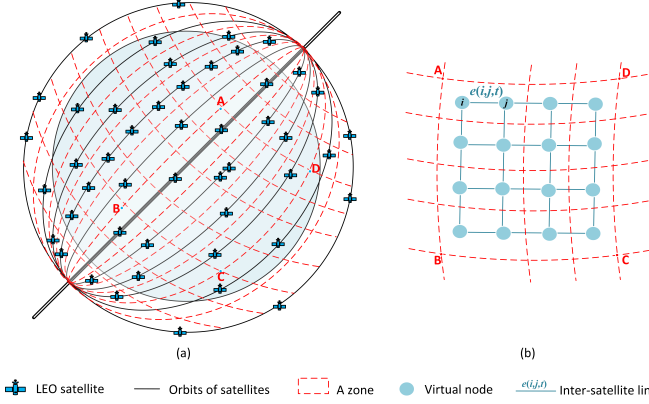


Fig. 2. (a) LEO satellite network with various types of LEO satellites, (b) Snapshot-free dynamic network model.

In Fig. 2 (a), the Earth is divided into multiple static disjoint zones according to longitude and latitude. These zones are stationary with respect to the ground. Each static zone corresponds to a VN. A device in space (such as satellites) associated with a VN implies that its sub-satellite point (on the ground) is located in the zone corresponding to the VN. The association between satellites and VNs is changing over time. In this way, the *topology dynamics* of the LEO satellite network are converted into the *association dynamics* between satellites and VNs (addressed in Section V-C).

Fig. 2 (b) shows a part of the VN network shown in Fig. 2 (a). The edges between each pair of VNs represent the communication links between the associated satellites. The inter-satellite and satellite-ground connection strategies are stated as follows.

- **Inter-satellite connections:** each satellite has four inter-satellite links with its neighbors where two are intra-plane and two are inter-plane.
- **Satellite-ground connections:** a satellite can communicate with a ground station only when the elevation angle between them is greater than the minimum elevation angle. For simplicity, in this paper we assume that a satellite can communicate with ground stations in a zone when its sub-satellite point is located in that zone.

In this VN network, the transmission resources are related to edges; whereas, the computing resources are related to VNs.

The available resources decrease as they are occupied and increase as they are released, which forms *resource dynamics* (addressed in Section V-B).

We would like to emphasize that, although the proposed dynamic network model generates zones and VNs in the same way as existing VN network models, the abstraction of dynamics is fundamentally different: the proposed model is a continuous-time model rather than a discrete-time model; the proposed model can represent the topology dynamics as well as the resource dynamics.

##### B. Traffic Model

For simplicity, the following traffic model is adopted in this paper without distinguishing applications. The task arrival is assumed to be Poisson stochastic processes with parameter  $\lambda$  because such processes have attractive theoretical properties [63]. The  $k^{\text{th}}$  ( $k \in \mathbb{Z}^+$ ) computational task arrives at VN  $u$  and instant  $t_{u,k}$  is denoted as  $\mathcal{T}_{u,k}$ . It could be divided into multiple independent subtasks, i.e.,  $\mathcal{T}_{u,k} = \{\tau_{u,k}^1, \tau_{u,k}^2, \dots, \tau_{u,k}^l, \dots, \tau_{u,k}^{n_{u,k}}\}$  ( $l, n_{u,k} \in \mathbb{Z}^+$ ,  $l \leq n_{u,k}$ ), where  $\tau_{u,k}^l$  is the  $l^{\text{th}}$  subtask of  $\mathcal{T}_{u,k}$  and  $n_{u,k}$  is the total number of subtasks of  $\mathcal{T}_{u,k}$ . Subtasks are the smallest unit of transmission and computation. Subtasks belonging to the same task are routed to the same destination.

Generally, there are seven items used to depict subtask  $\tau_{u,k}^l$ , i.e.,  $\Lambda_{\tau_{u,k}^l} = (\widetilde{C}_{u,k}^l, \widetilde{N}_{u,k}^l, \widetilde{S}_{u,k}^l, \vartheta_{u,k}^l, \alpha_{u,k}^l, \beta_{u,k}^l, t_{u,k})$ , where  $\widetilde{C}_{u,k}^l$ ,  $\widetilde{N}_{u,k}^l$ ,  $\widetilde{S}_{u,k}^l$  and  $\vartheta_{u,k}^l$  are the computation requirement (i.e., necessary CPU cycles) of accomplishing subtask  $\tau_{u,k}^l$  in GFLOPS, the data volume of subtask  $\tau_{u,k}^l$  in gigabytes (GB), the amount of memory needed to complete the computation of subtask  $\tau_{u,k}^l$  in GB and the required delay threshold to process subtask  $\tau_{u,k}^l$  in seconds.  $\alpha_{u,k}^l$  and  $\beta_{u,k}^l$  represent the longitude and latitude of the destination, respectively.

In addition,  $n_{u,k}$ ,  $\widetilde{C}_{u,k}^l$ ,  $\widetilde{N}_{u,k}^l$  and  $\widetilde{S}_{u,k}^l$  follow the log-normal distribution (i.e.,  $\ln(n_{u,k}) \sim N(\mu_n, \sigma_n^2)$ ,  $\ln(\widetilde{C}_{u,k}^l) \sim N(\mu_C, \sigma_C^2)$ ,  $\ln(\widetilde{N}_{u,k}^l) \sim N(\mu_N, \sigma_N^2)$  and  $\ln(\widetilde{S}_{u,k}^l) \sim N(\mu_S, \sigma_S^2)$ ). The delay threshold of subtask  $\tau_{u,k}^l$  is randomly chosen from  $\{\vartheta_1, \vartheta_2\}$ . The longitude  $\alpha_{u,k}^l$  and latitude  $\beta_{u,k}^l$  of the destination of  $\tau_{u,k}^l$  are randomly generated and subject to uniform distribution.

##### C. Delay Model

In this paper, the overall delay consists of the transmission delay, the propagation delay, the computation delay and the waiting delay.

In dynamic networks, the *transmission* delay  $T_{trans}$  of subtask  $\tau_l$  on edge  $e(i, j)$  starting from instant  $t$  satisfies the following equation:  $\widetilde{N}_l = \int_t^{t+T_{trans}} R_{i,j}^r(t) dt$ , where  $\widetilde{N}_l$  is the data volume of subtask  $\tau_l$  and  $R_{i,j}^r(t)$  is the available transmission rate of edge  $e = (i, j)$  at instant  $t$ .

Similarly, the *computation* delay  $T_{comp}$  of  $\tau_l$  at node  $i$  starting from instant  $t$  satisfies the following equation:  $\widetilde{C}_l = \int_t^{t+T_{comp}} C_i^r(t) dt$ , where  $\widetilde{C}_l$  is the computational requirement of subtask  $\tau_l$ .  $C_i^r(t)$  is the amount of available computing capability that node  $i$  can provide at instant  $t$ .



Ignoring the minor distance changes during transmissions, the *propagation* delay  $T_{prop}$  on edge  $e(i, j)$  starting from instant  $t$  is mathematically defined as  $T_{prop}(t) = D_{i,j}(t)/c$ , where  $D_{i,j}(t)$  represents the distance between node  $i$  and node  $j$  at instant  $t$ .  $c$  is the speed of light.

The *waiting* delay is the duration from the arrival of a subtask to the moment when the subtask starts being transmitted or processed. In the following, the waiting delay before transmission and computation are included in the corresponding transmission delay and computation delay, respectively.

## V. SNAPSHOT-FREE DYNAMIC NETWORK MODELING

To overcome the limitations of existing LEO satellite network models stated in Section II-A, a snapshot-free dynamic network modeling method is proposed in this section. In this section, the resource dynamics are addressed first. Then we propose a graph-based method to address the topology dynamics. Finally, the advantages of the proposed network modeling method are summarized.

### A. Definition of Dynamic Network Model

The proposed dynamic network model generates zones and VNs in the same way as existing VN network models. However, they have some significant differences: the proposed model is a continuous-time model rather than a discrete-time model; the proposed model can represent the topology dynamics as well as the resource dynamics.

The snapshot-free dynamic network model (SFDNM) could be defined as  $G_{SFDNM}(t) = (\mathbf{V}, \mathbf{E}(t), \mathbf{W}_E(t), \mathbf{W}_V(t))$ , where  $\mathbf{V} = \{v_1, v_2, \dots, v_n\} (n = |\mathbf{V}|)$  is the set of VNs,  $\mathbf{E}(t) = \{e_1, e_2, \dots, e_{m(t)}\} (m(t) = |\mathbf{E}(t)|)$  is the set of edges. An edge could be represented as  $e = (i, j)$  if  $i$  is the head of  $e$  and  $j$  is the tail of  $e$ .  $\mathbf{W}_E(t) = \{\omega_{e_1}(t), \omega_{e_2}(t), \dots, \omega_{e_{m(t)}}(t)\}$  is the weight set of edges and  $\mathbf{W}_V(t) = \{\omega_{v_1}(t), \omega_{v_2}(t), \dots, \omega_{v_n}(t)\}$  is the weight set of node (i.e., VNs). Since both transmission and computing resources have impact on computing-aware routing problem stated in Section I, the proposed model have both edge weights and node weights. These weights are task-related, which will be elaborated in Section V-B2.

### B. Time-Varying Resources Modeling

1) *Impact Factors of Edges and VNs*: Due to the high-speed mobility and the limited on-board resources of LEO satellites, the routing process is impacted by many factors, including 1) the intermittent communications between LEO satellites, 2) the bandwidths of ISLs, and 3) the available on-board resources of computing, memory and energy. Among these factors mentioned above, factor 1) and factor 2) are related to the transmission process and form the impact factor set of edges, whereas factor 3) is about the computing process and forms the impact factor set of VNs. The definitions of the impact factor set of VNs edges are given below.

The *impact factor set of edges*  $\Lambda_E = \{\mathbb{D}, \mathbb{B}\}$ . The intermittent communication between satellites is determined by the variation of distance caused by the relative motion between

satellites. Here the inter-satellite distance matrix is represented as  $\mathbb{D} = \{D_{i,j}(t), i, j \in \mathbf{V}, t \in [0, T]\}$ , where  $D_{i,j}(t)$  is the distance between satellites associated with VN  $i$  and VN  $j$  at instant  $t$ ;  $T$  is the duration of simulation.  $\mathbb{B} = \{B_{i,j}^r(t), i, j \in \mathbf{V}, t \in [0, T]\}$  presents the available spectrum bandwidth matrix of inter-satellite links, where  $B_{i,j}^r(t)$  is the available spectrum bandwidth of the communication link between VN  $i$  and VN  $j$  at instant  $t$ . For inter-satellite links,  $R_{i,j}^r(t) = \sigma \times B_{i,j}^r(t)$  indicates that the data transmission rate of edge  $e = (i, j)$  at instant  $t$  is proportional to the spectrum bandwidth  $B_{i,j}^r(t)$ .

The *impact factor set of VNs*  $\Lambda_V = \{\mathbb{C}, \mathbb{S}, \mathbb{E}\}$ . Here the available computing capability matrix is denoted as  $\mathbb{C} = \{C_i^r(t), i \in \mathbf{V}, t \in [0, T]\}$ , the available memory matrix is defined as  $\mathbb{S} = \{S_i^r(t), i \in \mathbf{V}, t \in [0, T]\}$ , and the available energy matrix is represented as  $\mathbb{E} = \{E_i^r(t), i \in \mathbf{V}, t \in [0, T]\}$ .  $C_i^r(t)$ ,  $S_i^r(t)$ , and  $E_i^r(t)$  represent the available computing capability in GFLOPS, memory resources in GB and battery energy in W.h of VN  $i$  at instant  $t$ , respectively.

It is worth noting that the inter-satellite distance matrix  $\mathbb{D}$  is derived from the satellite orbit parameters. In addition, the available bandwidth matrix  $\mathbb{B}$ , the available computing capability matrix  $\mathbb{C}$ , the available memory matrix  $\mathbb{S}$ , and the available energy matrix  $\mathbb{E}$  are updated when the resources of the LEO satellite network change. Therefore, the proposed dynamic network model is update-driven and snapshot-free.

2) *Edge Weights and Node Weights*: For computing-aware routing, the routing process is composed of a transmission process and a computing process. That is, not only the data should be transmitted along the path, but also the tasks should be computed at the selected computing VNs on the path. Specifically, the weight of an edge is related to the transmission process and is denoted as the sum of transmission delay and propagation<sup>3</sup> delay. The weight of a VN is related to the computing process and is defined as its processing delay. The equations for calculating the edge weights and node weights are defined as follows.

The *set of edge weights*.  $\mathbf{W}_E(t) = \{\omega_e^{u,k,l}(t) \mid e \in \mathbf{E}(t), u \in \mathbf{V}, t \in [0, T]\}$  is the set of edge weights.  $\omega_e^{u,k,l}(t)$  is the weight of edge  $e = (i, j)$  corresponding to subtask  $\tau_{u,k}^l$  and instant  $t$ . It is the sum of the transmission delay  $T_{trans}^{e,u,k,l}(t)$  and propagation delay  $T_{prop}^e(t)$  of subtask  $\tau_{u,k}^l$  on edge  $e = (i, j)$  starting from instant  $t$ , which can be mathematically defined as follows.

$$\omega_e^{u,k,l}(t) = \begin{cases} T_{trans}^{e,u,k,l}(t) + T_{prop}^e(t), & \zeta \leq \mathcal{T}_{i,j}^r(t), \\ \infty, & \text{otherwise,} \end{cases} \quad (1)$$

In equation (1),  $\zeta = T_{trans}^{e,u,k,l}(t) + T_{prop}^e(t)$ .  $\mathcal{T}_{i,j}^r(t)$  denotes the visible duration between  $i$  and  $j$  at  $t$  reflecting the intermittent communication of LEO satellite networks. It can be concluded that  $\omega_e^k(t)$  equals 0 when  $\zeta > \mathcal{T}_{i,j}^r(t)$ , which indicates that task  $k$  cannot be transmitted through edge  $e = (i, j)$  successfully at  $t$  because the visible duration is shorter than the time required by the transmission process.

<sup>3</sup>Propagation delays should be accounted for in LEO satellite networks, because the inter-satellite distance ranges from ten to thousands of kilometers. The propagation delays at milliseconds levels are much larger than those in terrestrial mobile networks [64].

*The set of node weights.*  $\mathbf{W}_V(t) = \{\omega_i^{u,k,l}(t) \mid i, u \in \mathbf{V}, t \in [0, T]\}$  is the set of node weights.  $\omega_i^{u,k,l}(t)$  is the weight of VN  $i$  corresponding to subtask  $\tau_{u,k}^l$  and instant  $t$ . It is the processing delay  $T_{proc}^{i,u,k,l}(t)$  of  $\tau_{u,k}^l$  at VN  $i$  start from instant  $t$ , which can be mathematically defined as follows.

$$w_i^{u,k,l}(t) = \begin{cases} T_{proc}^{i,u,k,l}(t), & S_i^r(t) \geq \widetilde{S}_{u,k}^l \text{ \& } E_i^r(t) \geq f(\widetilde{C}_{u,k}^l), \\ \infty, & \text{otherwise,} \end{cases} \quad (2)$$

In equation (2),  $S_i^r(t)$  and  $E_i^r(t)$  are the amount of available memory and energy that VN  $i$  can provide at instant  $t$ , respectively.  $\widetilde{S}_{u,k}^l$  is the amount of memory required to complete subtask  $\tau_{u,k}^l$ .  $\widetilde{C}_{u,k}^l$  is the computation requirement of subtask  $\tau_{u,k}^l$ .  $f(\cdot)$  maps the amount of computation to the amount of energy consumption. It can be concluded that  $w_i^k(t)$  equals 0 when  $S_i^r(t) < \widetilde{S}_{u,k}^l$  or  $E_i^r(t) < f(\widetilde{C}_{u,k}^l)$ , which means that the computation requirement of subtask  $\tau_{u,k}^l$  cannot be completed by VN  $i$  at instant  $t$  if the VN's available memory or energy is insufficient.

Based on the statements above, it can be concluded that the snapshot-free dynamic network model has two main differences from the conventional static network models. First, the edges and nodes of the proposed dynamic network model are both weighted. For computing-aware routing, each subtask should be computed in the computing VN selected on the path. Since the processing delay is related to the available on-board resources of the selected computing VN, the nodes of the proposed dynamic network model should be weighted to assist the computing VN selection. Second, both the edge weights and the node weights are time-varying. Unlike static terrestrial networks, LEO satellites are moving at high speed. Therefore, what is changed is not limited to the network topology; the inter-satellite distances, the available ISL data rates, and the available on-board resources of satellites all change in real-time. The dynamics stated above are closely related to the routing path selection. It is necessary to model the LEO satellite network as a dynamic network with time-varying weights.

### C. Dynamic Topology Modeling

After modeling the LEO satellite network with the VN network model, the topology dynamics are converted into the dynamics of the association between satellites and VNs. The numbers of satellites associated with two adjacent VNs and the positions of satellites located in the zones could lead to different types of edges between these two VNs. Since a complex scenario can be considered as a combination of several basic scenarios, as Fig. 3 shows, we present five basic scenarios observed from the whole dynamic network, where two adjacent VN  $i$  and  $j$  are located in zone 1 and zone 2 respectively.

**Scenario 1:** at instant  $t$ , both VN  $i$  and VN  $j$  are associated with a satellite and neither of the satellites has reached the boundaries of their corresponding zones, then there is an edge from  $j$  to  $i$  with attribute values  $D_{i,j}(t)$  and  $B_{i,j}^r(t)$ .

**Scenario 2:** at instant  $t$ , VN  $i$  is not associated with any satellite. In other words, there is no satellite in zone 1 and

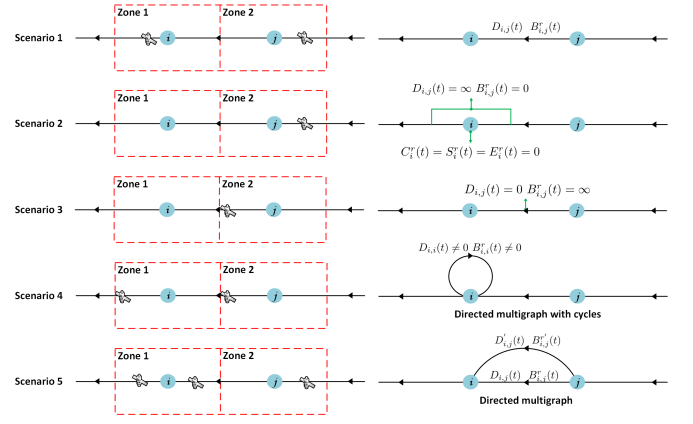


Fig. 3. Association between satellites and VNs. The left side of the figure illustrates how the satellites move in each zone. The right side presents the corresponding changes in node weights and edge weights.

communications cannot be established between  $i$  and  $j$ . In this case, the attribute values of the edge from  $j$  to  $i$  are  $B_{i,j}^r(t) = 0$  and  $D_{i,j}(t) = \infty$ . The available computing capability  $C_i^r(t)$ , memory resources  $S_i^r(t)$  and energy storage  $E_i^r(t)$  of  $i$  all equal to 0.

**Scenario 3:** at instant  $t$ , VN  $i$  is not associated with any satellite; meanwhile, the satellite associated with VN  $j$  just runs to the boundary between  $i$  and  $j$ , and a task needs to be transmitted from  $j$  to  $i$ . Although the task is transmitted from  $j$  to  $i$ , these two VNs are actually associated with the same satellite, thereby the inter-satellite transmission is not required. In this case, it can be considered that  $D_{i,j}(t) = 0$  and  $B_{i,j}^r(t) = \infty$ . The transmission delay and the propagation delay of transmitting the task from  $j$  to  $i$  are both 0.

**Scenario 4:** at instant  $t$ , satellite  $s_1$  is associated with VN  $i$  and about to leave zone 1. Satellite  $s_2$  is about to enter zone 1. If a task computed by  $s_1$  needs to be computed at VN  $i$  after instant  $t$ , it needs to be transmitted from  $s_1$  to  $s_2$  at instant  $t$ . During this process, although the satellites used for computing the task are changed, the VN for task execution remains the same. Therefore, VN  $i$  forms a loop at  $t$ : in this condition, tasks scheduled to be computed by  $i$  after  $t$  must be transmitted from VN  $i$  to itself through the loop to continue the execution.

**Scenario 5:** at instant  $t$ , VN  $i$  is associated with multiple satellites. In this case, VN  $i$  and its neighboring VN  $j$  are connected by multiple links, forming a multi-graph thereby.

Based on the above five basic scenarios, the association between satellites and VNs at any moment and the switching of the association can be accurately represented graphically.

## VI. PROBLEM FORMULATION

In this section, a typical DSSSP problem is introduced first, then the computing-aware routing problem is formulated and converted to a set of multiple DSSSP problems.

### A. Dynamic Single Source Shortest Path Problem

Many problems can be solved by searching the path with the minimum cost from the source node to the destination node.

The cost model varies in different problems, for example, it can be time or the number of hops. The problem becomes a DSSSP problem when the weight of each edge changes with the evolution of time [65], [66], [67].

The DSSSP problems cannot be solved by traditional dynamic programming methods (such as the Dijkstra algorithm). It has aroused wide interests among researchers.

**Problem  $\mathcal{P}0$  [68]:** Let  $G = (\mathbf{V}, \mathbf{E}(t), w(t))$  be a simple directed graph, where  $\mathbf{V} = \{V_1, V_2, \dots, V_n\}$  ( $n = |\mathbf{V}|$ ) and  $\mathbf{E}(t) = \{e_1, e_2, \dots, e_{m(t)}\}$  ( $m(t) = |\mathbf{E}(t)|$ ) are the sets of vertices and edges, respectively. Let  $e = (u, v) \in \mathbf{E}(t)$ ; then  $u$  is the head of  $e$  denoted as  $e_h$ , and  $v$  is the tail of  $e$  denoted as  $e_t$ . The edge weight function  $w(e, t)$  maps  $e \in \mathbf{E}(t)$ ,  $t \in [0, T]$  to non-negative real numbers. It gives the weights of corresponding edges at instant  $t^4$ . In other words, the length of the path depends on time  $t$ : assuming that there is a path  $P_{u,v} = \{(u_1, v_1), (u_2, v_2), \dots, (u_p, v_p)\}$  ( $u_1 = u$ ,  $v_p = v$ ,  $u_p = v_{p-1}$ ,  $p \in \mathbb{Z}^+ + 1$ ) and the start time is  $t_1$ , then the length of path  $P_{u,v}$  is  $L_{t_1}^{P_{u,v}} = \psi(P_{u,v}, t_1) = w(e_1, t_1) + w(e_2, t_2) + \dots + w(e_p, t_p)$ , where  $t_p = t_{p-1} + w(e_{p-1}, t_{p-1})$  ( $p \in \mathbb{Z}^+ + 1$ ),  $\psi(\cdot)$  maps the path and its start time to the path length. Then the DSSSP problem is defined as finding the shortest path  $\pi_{u,v,t} = \phi(u, v, t)$  ( $\phi(\cdot)$  maps the source node, the destination node and the start time to the shortest path) and its length  $L_t^{\pi_{u,v}}$  from a specific source node  $u$  to each  $v \in \mathbf{V}$  at time  $t$ .

It worth noting that  $\pi_{u,v,t} = \emptyset$  and  $L_t^{\pi_{u,v}} = \infty$  if  $v$  is not accessible from  $u$ . Problem  $\mathcal{P}0$  is a non-convex optimization problem. It has been proved to be NP-hard [69]. In other words, it is computationally prohibitive to find an optimal solution directly for the optimization problem  $\mathcal{P}0$ .

### B. Computing-Aware Routing Problem

In this paper, subtasks are the smallest unit of transmission and computation; thus, subtasks are the unit of routing. Since subtasks cannot be further partitioned, there is only one computing node on each path in the proposed computing-aware routing scheme.

#### 1) Computing-Aware Routing Problem in $G_{SF\text{D}NM}(t)$ :

As stated above, the multipath-single-computing-node routing strategy is adopted in this paper. That is, multiple independent subtasks that make up a task could be routed (i.e., transmitted and computed) on different paths simultaneously. Each subtask is the smallest unit of transmission and computation and cannot be further partitioned.

The ultimate goal of this paper is to find the optimal path for each subtask in the dynamic network  $G_{SF\text{D}NM}(t)$  established in Section V that could minimize the overall delay of each subtask, which could be formulated as follows.

**Problem  $\mathcal{P}1$ :** Let  $G_{SF\text{D}NM}(t) = (\mathbf{V}, \mathbf{E}(t), \mathbf{W}_E(t), \mathbf{W}_V(t))$  be a directed graph, where  $\mathbf{V} = \{v_1, v_2, \dots, v_n\}$  ( $n = |\mathbf{V}|$ ) is the node set,  $\mathbf{E}(t) = \{e_1, e_2, \dots, e_{m(t)}\}$  ( $m(t) = |\mathbf{E}(t)|$ ) is the edge set. Assuming that  $\tau_{u,k}$  is the  $k^{\text{th}}$  ( $k \in \mathbb{Z}^+$ ) computational task arrived at node  $u$  and its destination node is  $v$ . For subtask  $\tau_{u,k}^l \in \mathcal{T}_{u,k}$  ( $l \in \mathbb{Z}^+$ ),  $\mathbf{W}_E(t) = \{\omega_e^{u,k,l}(t) \mid$

$e \in \mathbf{E}(t)$ ,  $u \in \mathbf{V}$ ,  $t \in [0, T]\}$  is the set of edge weights and  $\mathbf{W}_V(t) = \{\omega_i^{u,k,l}(t) \mid i, u \in \mathbf{V}$ ,  $t \in [0, T]\}$  is the set of node weights. Assuming that subtask  $\tau_{u,k}^l$  is transmitted on path  $P_{u,v} = \{(u_1, v_1), (u_2, v_2), \dots, (u_q, v_q), \dots, (u_p, v_p)\}$  ( $u_1 = u$ ,  $v_p = v$ ,  $u_p = v_{p-1}$ ,  $p, q \in \mathbb{Z}^+$ ,  $p \geq q$ ) and processed by  $u_q$  which is the selected computing node on path  $P_{u,v}$ . If the start time of  $P_{u,v}$  is  $t_1$ , the length of path  $P_{u,v}$  is defined as  $L_{t_1}^{P_{u,v}, u_q} = \Psi(P_{u,v}, u_q, t_1) = \omega_{e_1}^{u,k,l}(t_1) + \omega_{e_2}^{u,k,l}(t_2) + \dots + \omega_{e_{q-1}}^{u,k,l}(t_{q-1}) + \omega_{u_q}^{u,k,l}(t_q) + \omega_{e_q}^{u,k,l}(t'_q) + \omega_{e_{q+1}}^{u,k,l}(t_{q+1}) + \dots + \omega_{e_p}^{u,k,l}(t_p)$ , where  $t_q$  is the instant that  $\tau_{u,k}^l$  arrives at node  $u_q$  (i.e., the head of edge  $e_q = (u_q, v_q)$ ),  $\Psi(\cdot)$  maps the path, the computing node and the start time to the path length. The relations of  $\{t_1, t_2, \dots, t_{q-1}, t_q, t'_q, t_{q+1}, \dots, t_p\}$  are stated as follows,

$$t_2 = t_1 + \omega_{e_1}^{u,k,l}(t_1), \quad (3a)$$

$$t_3 = t_2 + \omega_{e_2}^{u,k,l}(t_2), \quad (3b)$$

...

$$t_{q-1} = t_{q-2} + \omega_{e_{q-2}}^{u,k,l}(t_{q-2}), \quad (3c)$$

$$t_q = t_{q-1} + \omega_{e_{q-1}}^{u,k,l}(t_{q-1}), \quad (3d)$$

$$t'_q = t_q + \omega_{u_q}^{u,k,l}(t_q), \quad (3e)$$

$$t_{q+1} = t'_q + \omega_{e_q}^{u,k,l}(t'_q), \quad (3f)$$

...

$$t_p = t_{p-1} + \omega_{e_{p-1}}^{u,k,l}(t_{p-1}). \quad (3g)$$

Then the computing-aware routing problem in  $G_{SF\text{D}NM}(t)$  (i.e.,  $\mathcal{P}1$ ) is defined as finding a path  $\pi_{u,v,t}$  from the source node  $u$  to the destination node  $v$  at time  $t$  and a computing node  $u_q$  on the path that can obtain the minimum path length  $L_t^\Pi$  ( $\Pi = \{\pi_{u,v,t}, u_q\} = \Phi(u, v, t)$ ,  $\Phi(\cdot)$  maps the source node, the destination node and the start time to the shortest path). Mathematically, the problem  $\mathcal{P}1$  is formulated as

$$(\mathcal{P}1) : \min_{\Pi} (\Psi(\Pi, t)), \quad \Pi = \{\pi_{u,v,t}, u_q\}, \quad (4a)$$

$$\text{s.t. } u, v, u_q \in \mathbf{V}, \quad (4b)$$

$$t \geq 0. \quad (4c)$$

It could be concluded from the definitions stated above that there are some differences between  $\mathcal{P}0$  and  $\mathcal{P}1$ . First, both the edges and nodes in  $\mathcal{P}1$  are weighted. Second, besides the edges for transmission, in problem  $\mathcal{P}1$ , there is a computing node on the path to execute the computation of the corresponding subtask.

The shortest path of  $\mathcal{P}1$  contains multiple transmission edges and a computing node.  $\mathcal{P}1$  can be converted to  $\mathcal{P}0$  when the computation is executed at the source node  $u$  or the destination node  $v$ . Since  $\mathcal{P}0$  is NP-hard, problem  $\mathcal{P}1$  is NP-hard as well.

The computing-aware routing process in  $G_{SF\text{D}NM}(t)$  can be divided into three stages: (1) finding the shortest path from the source node  $u$  to the computing node  $u_q$  to transmit the raw data of subtask  $\tau_{u,k}^l$ , (2) processing the computation of  $\tau_{u,k}^l$  on a specific computing node  $u_q$ , (3) finding the shortest path from  $u_q$  to the destination node  $v$  to transmit the computation result of subtask  $\tau_{u,k}^l$ . It is worth noting

<sup>4</sup> $t$  is the instant when data is transmitted to the head of  $e$  (i.e.,  $e_h$ ). It is also called "the start time of  $w(e, t)$ ", or "the time of  $w(e, t)$ " for short.



that stage (1) and stage (3) are both typical DSSSP problem (i.e.,  $\mathcal{P}0$ ) whose optimal results are  $\pi_{u,u_q,t} = \phi(u, u_q, t)$  and  $\pi_{u_q,v,t'_q} = \phi(u_q, v, t'_q)$ , respectively. Assuming that the path length of  $\pi_{u,u_q,t}$  and  $\pi_{u_q,v,t'_q}$  are  $L_t^{\pi_{u,u_q}} = \psi(\pi_{u,u_q,t}, t)$  and  $L_{t'_q}^{\pi_{u_q,v}} = \psi(\pi_{u_q,v,t'_q}, t'_q)$ . Then the problem  $\mathcal{P}1$  can be rewritten as the following problem  $\mathcal{P}2$ .

**Problem  $\mathcal{P}2$ :** Let  $G_{SFDDNM}(t) = (\mathbf{V}, \mathbf{E}(t), \mathbf{W}_E(t), \mathbf{W}_V(t))$  be a directed graph. (The definition of  $G_{SFDDNM}(t)$  here is the same as that in  $\mathcal{P}1$ .) For any subtask  $\tau_{u,k}^l \in \mathcal{T}_{u,k}$  ( $l \in Z^+$ ), find a computing node  $u_q$  ( $u_q \in \mathbf{V}$ ), whose node weight is  $\omega(u_q, t_q) = \omega_{u_q}^{u,k,l}(t_q)$ , which could minimize the path length  $L_t^\Pi$  ( $\Pi = \{\pi_{u,v,t}, u_q\} = \{\pi_{u,u_q,t} \cup \pi_{u_q,v,t'_q}, u_q\} = \Phi(u, v, t)$ ). Mathematically, the problem  $\mathcal{P}2$  is formulated as

$$(\mathcal{P}2):$$

$$\min_{u_q} (\psi(\phi(u, u_q, t), t) + \omega(u_q, t_q) + \psi(\phi(u_q, v, t'_q), t'_q)), \quad (5a)$$

s.t. (4b), (4c) and

$$t \leq t_q < t'_q. \quad (5b)$$

For a specific subtask,  $u, v, t$  are known. The mapping of  $\psi(\cdot)$  is given in the definition of  $\mathcal{P}0$ . The mapping of  $\phi(\cdot)$  can be obtained by solving the DSSSP problem defined in  $\mathcal{P}0$ . In addition,  $t_q$  and  $t + q'$  can be calculated by (3d) and (3e).

Given the above discussion, a computing-aware routing process can be separated into the transmission process (i.e., stage (1) and stage (3)) and the computing process (i.e., stage (2)). In this way, the problem  $\mathcal{P}1$  can be converted to problem  $\mathcal{P}2$  which divides the optimization procedure of  $\mathcal{P}1$  into finding the shortest transmission path with a specific computing node and finding the optimal computing node with the corresponding shortest transmission path.

## VII. COMPUTING-AWARE ROUTING BASED ON SNAPSHOT-FREE DYNAMIC NETWORK MODEL

Section V provides an accurate model of the LEO satellite network studied in this paper. In this section, we discuss how to solve the computing-aware routing problem (presented in Section VI-B1) in the established dynamic network model.

Since the computing-aware routing problem in dynamic networks is NP-hard; thus, we propose the following GA-based algorithm to solve this problem as presented in Algorithm 1. In summary, it outputs the set of the optimal routing path and the computing node  $\Pi = \{\pi_t(u, v), u_q\}$  and the corresponding path length  $L^*$ .

As Algorithm 1 shows, before running the algorithm, the sets of VNs and edges should be generated based on the dynamic network model construct method introduced in Section II first. In addition, the impact factor set of edges  $\Lambda_E$ , the impact factor set of VNs  $\Lambda_V$  and subtask  $\tau_{u,k}^l$ 's parameter set  $\Lambda_{\tau_{u,k}^l}$  are the input data. Algorithm 1 outputs the set of the optimal routing path and the computing node  $\Pi = \{\pi_t(u, v), u_q\}$ , the corresponding minimum overall delay  $L^*$ , the updated impact factor set of edges  $\Lambda_E$  and the updated impact factor set of VNs  $\Lambda_V$ .

### Algorithm 1: GA-based computing-aware routing

---

**Data:** The VN set and edge set of the dynamic network model  $\mathbf{V}, \mathbf{E}(t)$

**Data:** Impact factor sets of edges and VNs  $\Lambda_E = \{\mathbb{D}, \mathbb{R}\}$ ,  $\Lambda_V = \{\mathbb{C}, \mathbb{S}, \mathbb{E}\}$

**Data:** Subtask  $\tau_{u,k}^l$ 's parameter set  $\Lambda_{\tau_{u,k}^l} = (\widetilde{C_{u,k}^l}, \widetilde{N_{u,k}^l}, \widetilde{S_{u,k}^l}, \vartheta_{u,k}^l, \alpha_{u,k}^l, \beta_{u,k}^l, t_{u,k})$

**Result:** The optimal path and computing node set  $\Pi = \{P^*, u_q^*\}$  and the minimum overall delay  $L^*$

**Result:** The updated  $\Lambda_E$  and  $\Lambda_V$

- 1 Initialize  $P^* \leftarrow \emptyset$ ,  $u_q^* \leftarrow \emptyset$ ,  $L^* \leftarrow \infty$ , the instant set  $\mathbf{T}^* \leftarrow \emptyset$ ;
- 2 Calculate  $\mathbf{W}_E(t)$ ,  $\mathbf{W}_V(t)$  based on  $\Lambda_E$ ,  $\Lambda_V$  and  $\Lambda_{\tau_{u,k}^l}$  with Equ. (1) and (2);
- 3 Set  $G_{SFDDNM}(t) \leftarrow (\mathbf{V}, \mathbf{E}(t), \mathbf{W}_E(t), \mathbf{W}_V(t))$ ;
- 4 Set  $\Lambda'_{\tau_{u,k}^l} = (\widetilde{C_{u,k}^l}, 0, \widetilde{S_{u,k}^l}, \vartheta_{u,k}^l, \alpha_{u,k}^l, \beta_{u,k}^l, t_{u,k})$ ;
- 5 Calculate  $\mathbf{W}'_E(t)$  based on  $\Lambda_E$  and  $\Lambda'_{\tau_{u,k}^l}$  with Equ. (1);
- 6 Set  $G'_{SFDDNM}(t) \leftarrow (\mathbf{V}, \mathbf{E}(t), \mathbf{W}'_E(t), \mathbf{W}_V(t))$ ;
- 7 **foreach**  $u_q \in \mathbf{V}$  **do**
- 8      $(\pi_1, \mathbf{T}_1) \leftarrow GA(G_{SFDDNM}(t), u, u_q, t_{u,k})$ ;
- 9      $d_1 \leftarrow L(\pi_1, t_{u,k})$ ;
- 10     $d_2 \leftarrow w_{u_q}^{u,k,l}(t_{u,k} + d_1)$ ;
- 11     $(\pi_2, \mathbf{T}_2) \leftarrow GA(G'_{SFDDNM}(t), u_q, v, t_{u,k} + d_1 + d_2)$ ;
- 12     $d_3 \leftarrow L(\pi_2, t_{u,k} + d_1 + d_2)$ ;
- 13     $L_{temp} \leftarrow d_1 + d_2 + d_3$ ;
- 14    **if**  $L_{temp} < L^*$  **then**
- 15        $\Pi = \{P^*, u_q^*\} \leftarrow \{\pi_1 \cup \pi_2, u_q\}$ ;
- 16        $L^* \leftarrow L_{temp}$ ;
- 17        $\pi_1^* \leftarrow \pi_1$ ;
- 18        $\mathbf{T}^* \leftarrow \mathbf{T}_1$ ;
- 19        $T_{proces}^* \leftarrow d_2$ ;
- 20    **end**
- 21 **end**
- 22 **for**  $i \leftarrow 1, 2, \dots, |\pi_1^*| - 1$  **do**
- 23     $R_{i,j}^r([\mathbf{T}^*(i), \mathbf{T}^*(i+1)]) \leftarrow 0$ ;
- 24 **end**
- 25  $C_{u_q^*}^r([\mathbf{T}^*(|\pi_1^*|), \mathbf{T}^*(|\pi_1^*|) + T_{proces}^*]) \leftarrow 0$ ;

---

Except the initialization (line 1), Algorithm 1 contains three parts. The first part is the graph generation (line 2 to line 6). The algorithm prepares the key data structure  $G_{SFDDNM}$  and assigns weight to it according to physical constraints. Inside the algorithm, the sets of edge weights and node weights of the dynamic network model  $G_{SFDDNM}(t)$  are calculated based on  $\Lambda_E$ ,  $\Lambda_V$  and  $\Lambda_{\tau_{u,k}^l}$  with Equ. (1) and (2) (line 2 and line 3). Because the data volume of  $\tau_{u,k}^l$  can be reduced to a few bits after being computed, this paper only considers the propagation delay and ignores the transmission delay in the computing result transmission process. Therefore, a new edge weight set is calculated and a new dynamic network model  $G'_{SFDDNM}(t)$  is generated (line 4 to line 6).

The second part is the optimal path and computing node finding. The main loop (line 7) iterates through all nodes. In each iteration,  $u_q$  is selected<sup>5</sup> for computing the subtask, and

<sup>5</sup>Note that the complexity of the proposed computing-aware routing scheme can be significantly reduced if the computing nodes are selected in an appropriate range and order. This paper mainly focuses on evaluating how much overall delay can be reduced by the proposed computing-aware routing scheme compared with the ground-offloading approach. The computing node selection strategy will be investigated in-depth in the following research.

the overall delay under this scenario is evaluated as  $L_{temp}$  which has three components (i.e.,  $d_1$ ,  $d_2$  and  $d_3$ ).  $d_1$  is caused by the transmission of raw data from the source node  $u$  to the computing node  $u_q$ . The time at which the subtask is created is  $t_{u,k}$ . The shortest path  $\pi_1$  and the instance set  $\mathbf{T}_1$  (recording the instances when  $t_{u,k}$  first arrived at each node on  $\pi_1$ ) are first solved with GA (line 8), and the delay  $d_1$  is calculated thereafter (line 9).  $d_2$  is caused by the processing of the subtask. Since the subtask is arrived at  $t_{u,k} + d_1$  which affects the computation delay, the delay is computed as  $w_{u_q}^{u,k,l}(t_{u,k} + d_1)$  (see line 10).  $d_3$  is caused by the computing result transmission from the computing VN  $u_q$  back to the destination VN  $v$ . Similar to the calculation of the first part of the delay, the shortest path is first solved with GA (line 11) and the delay is calculated next (line 12). In this step, the transmission delay of each edge is ignored because the computation results are usually only a few bits; therefore,  $G'_{SFDNM}(t)$  rather than  $G_{SFDNM}(t)$  is adopted as the input of GA. After calculating the delay under the assumption that node  $u_q$  is selected for computing, the global state is updated to find the best node. Here the optimal path  $P^*$ , the optimal computing node  $u_q^*$  and the minimum overall delay  $L^*$  are updated in line 15 and 16, respectively. Furthermore, for updating  $\Lambda_E$  and  $\Lambda_V$  in the next part, the optimal path from  $u$  to  $u_q^*$ , the instance set  $\mathbf{T}$  and the processing delay of the optimal computing node are recorded in line 17–19.

The third part is impact factor sets update. After the second part, the route for current subtask is resolved as  $\Pi \leftarrow \{P^*, u_q^*\}$ . However, the offloading of the subtasks occupies computation and transmission resources of the network. To reflect these changes, line 23 to line 25 subtracts the resources occupied by  $\Pi$  from the impact factor values of the time when  $\tau_{u,k}^l$  arrives at  $u_q$  and each edge on  $\pi$ .

It is worth noting that the proposed computing-aware scheme and the ground-offloading scheme are not mutually exclusive, but complementary. If computing resources on satellites are not sufficient, tasks can be offloaded to ground servers for computing. By adding ground servers as nodes of the dynamic network model, these ground servers will be selected for computing if the delay of on-board computing is larger than the delay of ground offloading.

## VIII. SIMULATION RESULTS AND ANALYSES

In this section, we evaluate the GA-based computing-aware routing scheme proposed in Section VII, and analyze the simulation results. We aim to answer the following research questions:

- **RQ1:** How much computing capability is required for LEO satellites to enable the proposed scheme to effectively reduce delay?
- **RQ2:** How does the transmission capability affect the performance of the proposed scheme?
- **RQ3:** What kind of tasks can be accelerated by the proposed scheme?

Furthermore, we would like to emphasize that the following assumptions and settings are made in the simulations.

- **Sufficient energy and storage resources.** As shown in Table II, existing onboard computing systems can provide thousands of GFLOPS of computing capability. Therefore, we assume that modern LEO satellites can afford the energy consumption when the computation capability is less or equal than 400 GFLOPS. The detailed influence of energy consumption will be discussed and evaluated in future work.
- **Zero computation delay for ground-offloading routing scheme.** Ground servers (such as supercomputers) usually have large amounts of computational resources. We round the computation delay towards zero for a fair comparison.
- **Platform-agnostic evaluation.** A satellite platform contains multiple components. For example, the onboard computing system determines the computing capability, whereas the signal transmitters and receivers determine the transmission capability. Since the components can be composed at will, a variety of configurations can be found for LEO satellites. Rather than evaluating specific LEO satellite configurations exhaustively, we reveal how the factors affect the performance in general.
- **Scheme-level comparison.** Since the pathfinding algorithm adopted in the ground-offloading scheme can also be applied to the transmission process of the proposed computing-aware scheme, for fairness, the pathfinding algorithm for both schemes is set as the GA algorithm. In other words, we focus on scheme-level rather than algorithm-level comparison in this paper. Furthermore, since task partitioning strategy also affects the routing performance, for a fair comparison, subtask is assumed to be the smallest unit of transmission and computation in both schemes.

The simulation parameters are presented in Table III.

TABLE III  
SIMULATION PARAMETERS

6	
Parameter	Value
Radius of the earth $R_e$	6,371,393 m
Mass of the earth $M_e$	$5.965 \times 10^{24}$ kg
Earth rotation angular velocity $\omega_e$	$7.29211510 \times 10^{-5}$
Gravitational $G$	$6.67428 \times 10^{-11}$
Kepler constant $K$	$3.9860 \times 10^{14}$
Velocity of light $c$	299,792,458 m/s
Number of orbits	10
Satellites per orbit	[10,12,10,12,10,12,10,12,10,12]
Orbit altitude $h$	[200,300,400,500,600,200,300,400,500,600] km
Orbit inclination $i_0$	$90^\circ$
Max data rate of ISL	5 Gbps
Channel number per ISL	1
Max data rate of SGL	0.2 Gbps
Max computing capability of satellites $C$	100 GFLOPS
The number of tasks arrived at each VN per second $\lambda$	1/60
Distribution parameters of subtask number per task $(\mu_n, \sigma_n)$	(3,1)
Distribution parameters of subtask's computation requirement $(\mu_C, \sigma_C)$	(50,2) GFLO
Distribution parameters of sub-task's data volume $(\mu_N, \sigma_N)$	(0.1,0.02) Gbps

### A. Reduced Overall Delay (RQ1)

Fig. 4 and Fig. 5 shows how the computing capability of LEO satellites affects the performance of computing-aware routing schemes. The simulation covers computing capability from 100 GFLOPS to 400 GFLOPS, and the remaining parameters follow Table III. For each computing capability, the delay of the computing-aware routing scheme is calculated, and the result is normalized with the corresponding delay of the benchmark routing scheme.

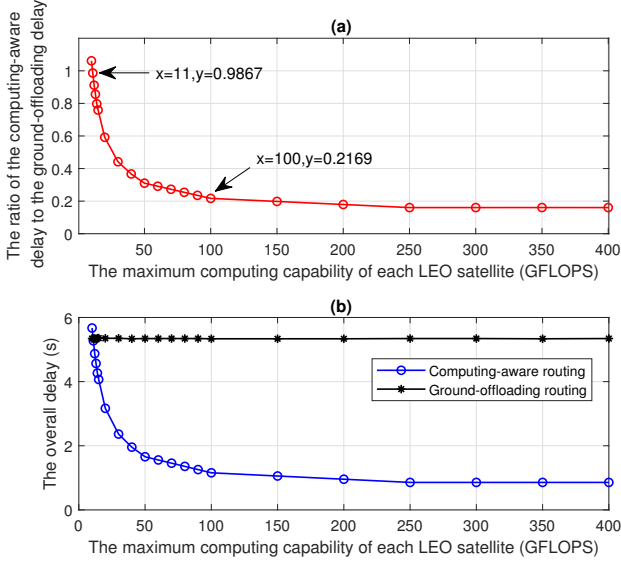


Fig. 4. (a) The ratio of the computing-aware delay to the ground-offloading delay versus the maximum computing capability of each LEO satellite  $C$ . A ratio of 1.0 means that two methods have the same performance. The lower the ratio, the better the proposed method. (b) The overall delays of the computing-aware routing scheme and the ground-offloading routing scheme

Fig. 4 (a) demonstrates shortened delays from the proposed computing-aware routing scheme. It can be concluded that the proposed computing-aware routing scheme can reduce the overall delay at most computing capabilities ( $C > 11$  GFLOPS). Especially when the maximum computing capability is set to 100 GFLOPS, the overall delay is reduced to 21.69% compared to the benchmark scheme (i.e.,  $3.61\times$  speedup). Since existing onboard computing systems can already provide these computing capabilities, the proposed scheme is not only effective but also feasible.

Furthermore, we notice that when  $C < 11$  GFLOPS, the performance of the ground-offloading scheme is better, which indicates that the time saved by reducing the amount of data transmitted is less than the increased computation delay due to insufficient computing capability. In this condition, tasks can be offloaded to ground servers for computing. In other words, the proposed computing-aware scheme and the ground-offloading scheme are not mutually exclusive, but complementary. By adding ground servers as nodes of the dynamic network model, these ground servers will be selected for computing if the delay of on-board computing is larger than the delay of ground-offloading.

In addition, the curve in Fig. 4 (a) shows that the stronger the computing capability, the greater the reduction. It flattens

out as more computing capability is added to satellites. This is because the computation delay of the proposed scheme converges to 0 after the computing capability is increased to a certain value. Meanwhile, the other part (i.e., transmission delay) of the overall delay remains constant. In this way, the overall delay of the proposed scheme converges to a constant value and the curve flattens out. The diminishing return indicates the cost-effectiveness of the computing-aware routing scheme. The most cost-saving hardware configuration is allocated to each satellite with 100 GFLOPS of computing capability. It is less than 10% of the most powerful existing onboard computing systems listed in Table II. In other words, not only can computing-aware routing achieve significant performance boost with advanced hardware, the advantage also persists with budget hardware with minimum costs.

The reasons for the change in ratio in Fig. 4 (a) can be explained by Figure 4 (b): as the computing capability of LEO satellites grows, the proposed approach can leverage more onboard resources to complete the tasks without transmitting them back to the ground. Therefore, the delay of the computing-aware routing gradually decreases, while the delay of ground-offloading routing stays still.

To further explain why computing-aware routing decreases the delay, Fig. 5 presents a delay decomposition of both the computing-aware routing scheme and the ground-offloading routing scheme.

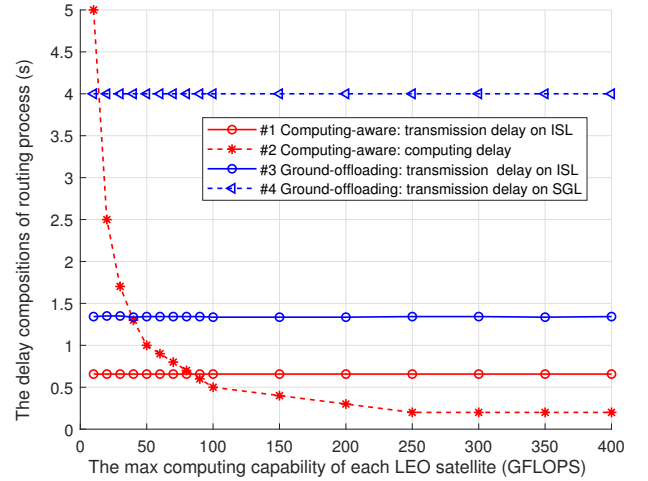


Fig. 5. Delay decomposition of computing-aware routing scheme v.s. ground-offloading routing scheme. The overall delay of computing-aware routing mainly consists of the transmission delay on ISL (#1), and the processing delay (#2). Similarly, the overall delay of ground-offloading routing mainly consists of the transmission delay on ISL (#3), and the transmission delay on SGL (#4).

Fig. 5 shows that the processing delay of the computing-aware routing scheme decreases when the computing capability of LEO satellites increases, whereas other delays remain unchanged. Furthermore, it can be concluded that the transmission delay of the computing-aware routing scheme is much lower than that of the benchmark scheme. It demonstrates that the proposed computing-aware routing scheme decreases delay successfully by transmitting computation results instead

of raw data. When the computing capability is greater than 11 GFLOPS, the delay reduced by transmitting computing results is larger than the increased computation delay; the ratio in Fig. 4 decreases when the computing capability of LEO satellites increases.

The proposed scheme can accelerate the offloading with a computing capability as low as 11 GFLOPS under the provided settings. When the computing capability is increased to 100 GFLOPS, a trivial computing capability supported by most LEO satellites, a  $3.61\times$  speedup can be observed. In general, the reduction becomes more significant as computing capability grows.

### B. Impact from Transmission Capability (RQ2)

Fig. 6 and Fig. 7 show how the transmission rates of SGL and ISL affect the performance of the computing-aware routing scheme. The simulation covers transmission rate of SGL from 0.2 Gbps to 10 Gbps and transmission rate of ISL from 0.25 Gbps to 20 Gbps. The remaining parameters follow Table III.

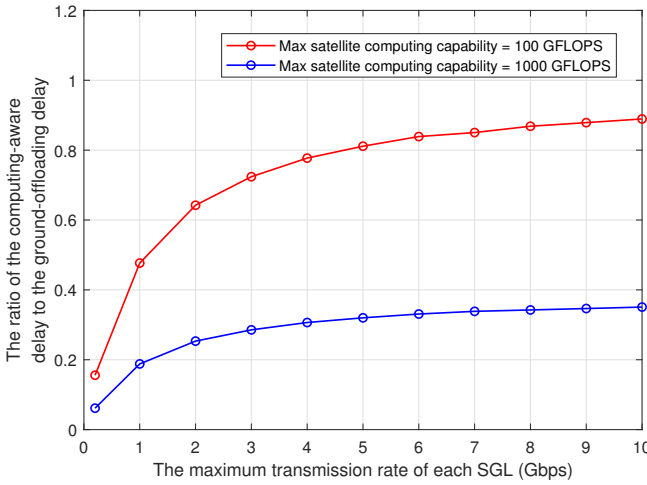


Fig. 6. Delay of the computing-aware routing scheme under different SGL transmission rates, normalized to the ground-offloading scheme.

In Fig. 6, the leftmost points presents the situation of the most common cases at the moment, where the transmission rate of SGL is assumed to be 0.2 Gbps. Similarly, the rightmost points present the ideal scenarios where the transmission rate of SGL is set to 10 Gbps. It could be concluded that, with the most practical SGL transmission rate setting at present, i.e., 0.2 Gbps, satellites with 100 GFLOPS and 1000 GFLOPS of computing capability can save 84.43% and 93.86% of the overall delay, respectively. Moreover, in Fig. 6, the ratio increases with the increase of SGL transmission rate. This is because the ground-offloading scheme transmits a much larger amount of data over the SGL than the proposed computing-aware scheme; therefore, as the SGL transmission capacity increases, the reduction in transmission delay over the SGL is much greater for the ground-offloading scheme than for the proposed computing-aware scheme. Although the advantage

of computing-aware routing scheme diminishes with a higher quality SGL, it is still significant: when the bandwidth is boosted to 10 Gbps, the satellite with 100 GFLOPS achieves a speedup of 11.07% nevertheless.

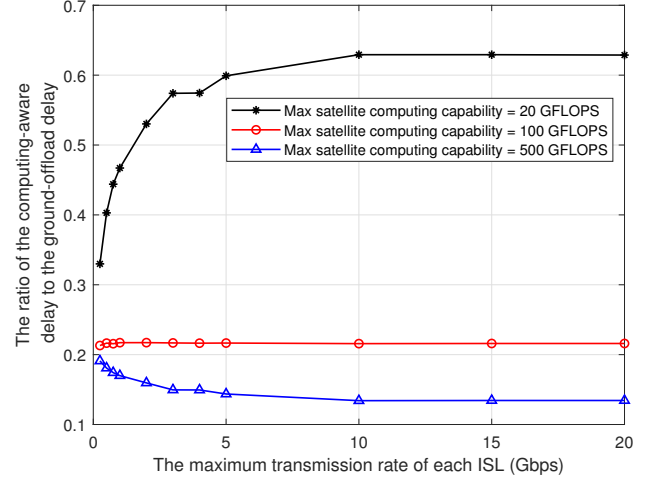


Fig. 7. Delay of computing-aware routing under different ISL transmission rates, normalized to the ground-offloading delay.

Fig. 7 shows the performance at different ISL rates. The ratio can be approximated with  $(T_{comp} + x \times \Delta_{ISL}) / (T_{trans,SGL} + y \times \Delta_{ISL})$ , where  $T_{comp}$  is the computation delay of the proposed scheme,  $T_{trans,SGL}$  is the transmission delay on SGL of the ground-offloading scheme.  $\Delta_{ISL}$  is the average delay on each ISL (i.e., one hop).  $x$  and  $y$  are the average hop number of transmitting raw data on ISL for the proposed scheme and the benchmark scheme, respectively. Since the proposed scheme performs onboard computing,  $x < y$ . After deducing the above equations, the following conclusion can be obtained: when  $T_{comp}/T_{trans,SGL} = x/y$ , the ratio constantly equals  $x/y$ ; when  $T_{comp}/T_{trans,SGL} > x/y$ , the ratio is larger than  $x/y$  and vice versa. In addition, when increasing the transmission capability of ISL,  $\Delta_{ISL}$  will decrease and converge to 0; therefore, the three curves all converge to  $T_{comp}/T_{trans,SGL}$ . With this conclusion, we can determine whether to perform the proposed scheme (onboard computing) or perform the ground-offloading scheme instead by estimating  $T_{comp}/T_{trans,SGL}$ .

The proposed scheme outperforms the ground-offloading scheme for all the realistic ISL/SGL configurations used in simulations. For the worst case of SGL ( $C = 20$  GFLOPS,  $R_{ISL} = 20$  Gbps), the proposed scheme still reduces 37.12% of the baseline delay. For the worst case of ISL ( $C = 100$  GFLOPS,  $R_{SGL} = 10$  Gbps), the proposed scheme still reduces 11.07% of the baseline delay.

### C. Impact from Task Properties (RQ3)

Fig. 8, Fig. 9, and Fig. 10 show how the data volume and computing requirement of subtasks affects the performance of

the computing-aware routing scheme, respectively. The simulation covers data volumes from 1 MB to 1 GB, and computing requirements from 50 Giga floating-point operations (GFLO) to 800 GFLO. The remaining parameters follow Table III. In Fig. 8, Fig. 9, and Fig. 10, the delay of the computing-aware routing is calculated, and the result is normalized with the corresponding delay of ground-offloading.

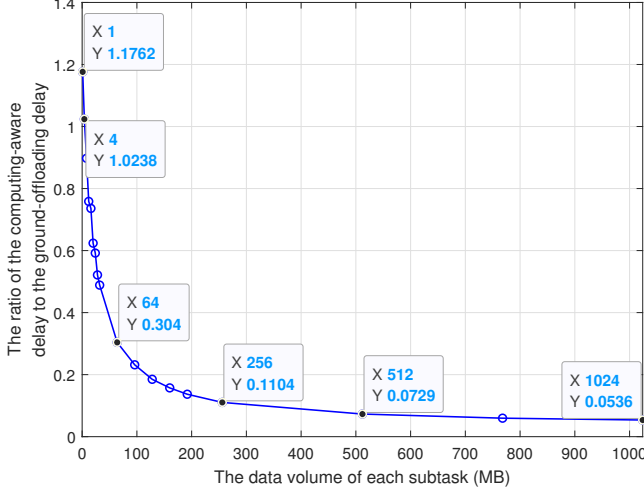


Fig. 8. Performance impact from subtasks' data volumes. The delay of the computing-aware routing scheme is normalized with the delay of the ground-offloading scheme. In other words, a ratio lower than 1.0 means that the proposed method is better.

From Figure 8 we can conclude that the proposed computing-aware routing scheme could reduce the overall delay over a wide range of data volume ( $> 4$  MB) with current network settings. The superiority of the computing-aware routing becomes more and more significant when the data volume of each subtask increases. Since the computing-aware routing replaces transmissions of large-scale raw data with final results, the amount of data to transfer is greatly reduced. The reduced demand also opens up opportunities in future data-intensive applications: for example, for applications transferring 1GB of data, the task execution efficiency can be boosted by 17.66x with computing-aware routing.

Figure 9 shows how the computing requirement of subtasks affects the performance of the proposed computing-aware routing. The simulation covers computing requirements from 50 GFLO to 800 GFLO, and the remaining parameters follow Table III.

From Figure 9 we can conclude that the computing-aware routing scheme has linear scalability concerning the computing requirement of each subtask. Because the onboard computing resources are limited in nature, computing-aware routing is not intuitively suitable for tasks with extreme computing demands. Despite this, for most tasks (400 GFLO or less), the proposed computing-aware routing scheme still handles them well; for corner cases where computing-aware routing is not the best approach, its performance downgrades gracefully in a linear way as the computing requirements grow.

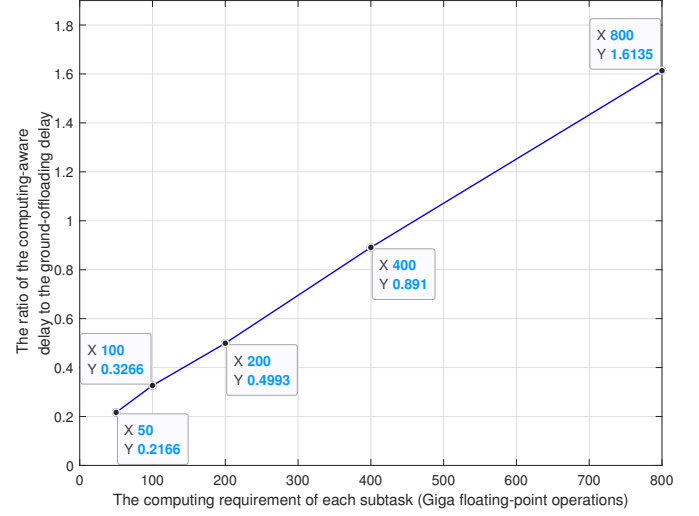


Fig. 9. Performance impact from subtasks' computing requirements. The delay of the computing-aware routing is normalized with the ground-offloading delay. In other words, a ratio lower than 1.0 means that the proposed method is better.

Figure 10 presents a comprehensive view of the delay of the computing-aware routing scheme, encompassing both the variables of subtasks' data volumes and computing requirements. It provides a method for compensating the limitations of computing-aware routing on tasks with high computation costs. Specifically, the strategy to determine whether to use the computing-aware routing scheme or fallback to the conventional ground-offloading routing for a specific task can be pre-computed by the ground station or GEO satellites.

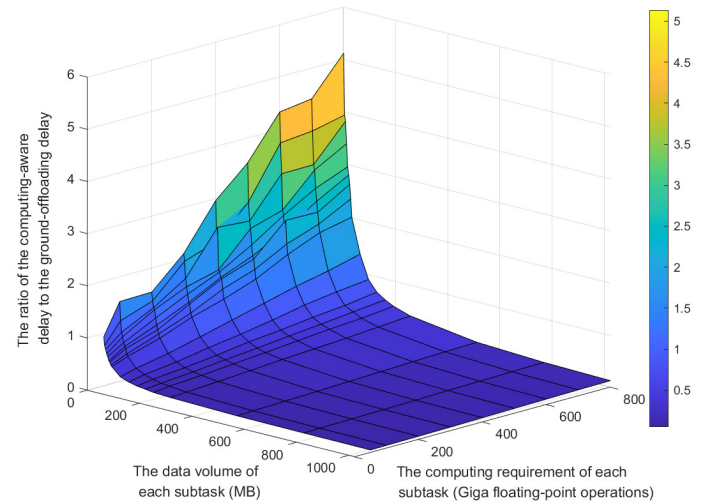


Fig. 10. The impact of subtasks' data volume and computing requirements on the delay of the computing-aware routing scheme (normalized to the ground-offloading delay). A ratio of 1.0 means two methods have equally good performance. The lower the ratio, the better the proposed method is.

When the deployment of a satellite network has completed, the parameters of onboard resources (e.g. computing capability) and the data rate of ISLs and SGLs are known in



advance. Therefore, the ground station can predict the overall performance of computing-aware routing for a specific task by considering its computing requirement and data volume. The evaluation results (as Figure 10 shows) can be uploaded to the satellite network in advance, and all newly generated tasks follow the predetermined threshold accordingly.

The proposed scheme is applicable to most kinds of tasks. Unless the task has minimal data volume (less than 4MB) or extremely high computation requirements (more than 400 GFLO), the proposed scheme can be used for improving offloading performance.

## IX. CONCLUSION

This paper investigates the LEO satellite network routing to fulfill new requirements of space missions. It first analyzes the challenges in LEO satellite networks, including the highly dynamic network topology, limited onboard resources, and intensive computational demands.

Aiming at tackling the challenges, this paper proposes a computing-aware routing scheme for LEO satellite networks. The paper first models the dynamic set of satellites as a snapshot-free network with time-varying weights. Then the computing-aware routing problem in the dynamic network is formulated as a combination of multiple DSSSP problem. In addition, a GA-based method is proposed to approximate the results in reasonable time. Simulation results demonstrate the applicability of the proposed approach, where the overall delay can be reduced in a wide range of network settings.

In the future, we will study how to further optimize the proposed computing-aware routing scheme to reduce its complexity. Furthermore, we will work on the task splitting mechanism in computing-aware routing.

## REFERENCES

- [1] A. J. Plaza and C.-I. Chang, *High performance computing in remote sensing*. CRC Press, 2007.
- [2] L. Zhang, L. Zhang, and B. Du, "Deep learning for remote sensing data: A technical tutorial on the state of the art," *IEEE Geoscience and Remote Sensing Magazine*, vol. 4, no. 2, pp. 22–40, 2016.
- [3] A. Plaza, Q. Du, Y.-L. Chang, and R. L. King, "High performance computing for hyperspectral remote sensing," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 4, no. 3, pp. 528–544, 2011.
- [4] C.-I. Chang, H. Ren, and S.-S. Chiang, "Real-time processing algorithms for target detection and classification in hyperspectral imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 39, no. 4, pp. 760–768, 2001.
- [5] P. Ghamisi, N. Yokoya, J. Li, W. Liao, S. Liu, J. Plaza, B. Rasti, and A. Plaza, "Advances in hyperspectral image and signal processing: A comprehensive overview of the state of the art," *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 4, pp. 37–78, 2017.
- [6] A. J. Plaza, "Special issue on architectures and techniques for real-time processing of remotely sensed images," *Journal of Real-Time Image Processing*, vol. 4, no. 3, pp. 191–193, 2009.
- [7] A. Alonso, M. Reyes, and Z. Sodnik, "Performance of satellite-to-ground communications link between artemis and the optical ground station," in *Optics in Atmospheric Propagation and Adaptive Systems VII*, vol. 5572, 2004, pp. 372–383.
- [8] B. Yost, S. Weston, G. Benavides, F. Krage, J. Hines, S. Mauro, S. Etchey, K. O'Neill, and B. Braun, "State-of-the-art small spacecraft technology," 2021.
- [9] T. M. Lovelley and A. D. George, "Comparative analysis of present and future space-grade processors with device metrics," *Journal of Aerospace Information Systems*, vol. 14, no. 3, pp. 1–14, 2017.
- [10] X. Pan, R. Liu, and X. Lv, "Low-complexity compression method for hyperspectral images based on distributed source coding," *IEEE Geoscience and Remote Sensing Letters*, vol. 9, no. 2, pp. 224–227, 2012.
- [11] P. Liu, H. Chen, S. Wei, L. Li, and Z. Zhu, "Hybrid-traffic-detour based load balancing for onboard routing in LEO satellite networks," *China Communications*, vol. 15, no. 6, pp. 28–41, 2018.
- [12] M. Madni, S. Iranmanesh, and R. Raad, "DTN and Non-DTN routing protocols for inter-cubesat communications: A comprehensive survey," *Electronics*, vol. 9, no. 3, p. 482, 2020.
- [13] S. El Alaoui, "Routing optimization in interplanetary networks," 2015.
- [14] X. Zhang, Y. Yang, M. Xu, and J. Luo, "ASER: Scalable distributed routing protocol for LEO satellite networks," in *Proc. IEEE LCN*, 2021, pp. 65–72.
- [15] Y. Lu, F. Sun, and Y. Zhao, "Virtual topology for LEO satellite networks based on earth-fixed footprint mode," *IEEE Communications Letters*, vol. 17, no. 2, pp. 357–360, 2013.
- [16] S. C. Burleigh, "Contact graph routing," Tech. Rep., 2011.
- [17] T. Zhang, J. Li, H. Li, S. Zhang, P. Wang, and H. Shen, "Application of time-varying graph theory over the space information networks," *IEEE Network*, vol. 34, no. 2, pp. 179–185, 2020.
- [18] C. Jiang and X. Zhu, "Reinforcement learning based capacity management in multi-layer satellite networks," *IEEE Transactions on Wireless Communications*, vol. 19, no. 7, pp. 4685–4699, 2020.
- [19] E. Köhler, K. Langkau, and M. Skutella, "Time-expanded graphs for flow-dependent transit times," in *European symposium on algorithms*. Springer, 2002, pp. 599–611.
- [20] T. Zhang, H. Li, S. Zhang, J. Li, and H. Shen, "Stag-based QoS support routing strategy for multiple missions over the satellite networks," *IEEE Transactions on Communications*, vol. 67, no. 10, pp. 6912–6924, 2019.
- [21] J. Whitbeck, M. D. de Amorim, V. Conan, and J. Guillaume, "Temporal reachability graphs," in *Proc. ACM Mobicom*, August 2012, pp. 377–388.
- [22] G. Araniti, N. Bezirgiannidis, E. Birrane, I. Bisio, S. Burleigh, C. Caini, M. Feldmann, M. Marchese, J. Segui, and K. Suzuki, "Contact graph routing in DTN space networks: overview, enhancements and performance," *IEEE Communications Magazine*, vol. 53, no. 3, pp. 38–46, 2015.
- [23] F. Tang, H. Zhang, and L. T. Yang, "Multipath cooperative routing with efficient acknowledgement for LEO satellite networks," *IEEE Transactions on Mobile Computing*, vol. 18, no. 1, pp. 179–192, 2019.
- [24] B. Soret and D. Smith, "Autonomous routing for LEO satellite constellations with minimum use of inter-plane links," in *Proc. IEEE ICC*. IEEE, 2019, pp. 1–6.
- [25] H. Tan and L. Zhu, "A novel routing algorithm based on virtual topology snapshot in LEO satellite networks," in *Proc. IEEE CSE*, 2014, pp. 357–361.
- [26] J. Shen, C. Wang, A. Wang, X. Sun, S. Moh, and P. C. K. Hung, "Organized topology based routing protocol in incompletely predictable Ad-hoc networks," *Computer Communications*, vol. 99, no. C, pp. 107–118, 2017.
- [27] M. Werner, "A dynamic routing concept for ATM-based satellite personal communication networks," *IEEE Journal on Selected Areas in Communications*, vol. 15, no. 8, pp. 1636–1648, 1997.
- [28] D.-N. Yang and W. Liao, "On multicast routing using rectilinear steiner trees for LEO satellite networks," *IEEE Transactions on Vehicular Technology*, vol. 57, no. 4, pp. 2560–2569, 2008.
- [29] E. Ekici, I. F. Akyildiz, and M. D. Bender, "A distributed routing algorithm for datagram traffic in LEO satellite networks," *IEEE/ACM Transactions on Networking*, 2001.
- [30] J. A. Ruiz de Azúa, A. Calveras, and A. Camps, "Internet of satellites (IoSat): Analysis of network models and routing protocol requirements," *IEEE Access*, vol. 6, pp. 20 390–20 411, 2018.
- [31] H. Li, T. Zhang, Y. Zhang, K. Wang, and J. Li, "A maximum flow algorithm based on storage time aggregated graph for delay-tolerant networks," *Ad Hoc Networks*, vol. 59, pp. 63–70, 2017.
- [32] D. Zhou, M. Sheng, B. Li, J. Li, and Z. Han, "Distributionally robust planning for data delivery in distributed satellite cluster network," *IEEE Transactions on Wireless Communications*, vol. 18, no. 7, pp. 3642–3657, 2019.
- [33] P. Yuan, Z. Yang, Y. Li, and Q. Zhang, "An event-driven graph-based min-cost delivery algorithm in earth observation DTN networks," in *Proc. WCSP*, 2015, pp. 1–6.
- [34] F. He, Q. Liu, T. Lv, C. Liu, H. Huang, and X. Jia, "Delay-bounded and minimal transmission broadcast in leo satellite networks," in *Proc. IEEE ICC*, 2016, pp. 1–7.

- [35] Z. Tang, Z. Feng, W. Han, W. Yu, B. Zhao, and C. Wu, "Improving the snapshot routing performance through reassigning the inter-satellite links," in *Proc. IEEE INFOCOM WKSHPS*, 2015, pp. 97–98.
- [36] B. George, S. Kim, and S. Shekhar, "Spatio-temporal network databases and routing algorithms: A summary of results," in *International Symposium on Spatial and Temporal Databases*. Springer, 2007, pp. 460–477.
- [37] T. Zhang, H. Li, S. Zhang, and J. Li, "A storage-time-aggregated graph-based QoS support routing strategy for satellite networks," in *GLOBECOM 2017-2017 IEEE Global Communications Conference*. IEEE, 2017, pp. 1–6.
- [38] B. George and S. Shekhar, "Time aggregated graphs," 2009.
- [39] Z. Song, Y. Hao, Y. Liu, and X. Sun, "Energy-efficient multiaccess edge computing for terrestrial-satellite internet of things," *IEEE Internet of Things Journal*, vol. 8, no. 18, pp. 14 202–14 218, 2021.
- [40] C. Ding, J.-B. Wang, H. Zhang, M. Lin, and G. Y. Li, "Joint optimization of transmission and computation resources for satellite and high altitude platform assisted edge computing," *IEEE Transactions on Wireless Communications*, vol. 21, no. 2, pp. 1362–1377, 2022.
- [41] Q. Tang, Z. Fei, B. Li, and Z. Han, "Computation offloading in LEO satellite networks with hybrid cloud and edge computing," *IEEE Internet of Things Journal*, vol. 8, no. 11, pp. 9164–9176, 2021.
- [42] N. Waqar, S. A. Hassan, A. Mahmood, K. Dev, D.-T. Do, and M. Gidlund, "Computation offloading and resource allocation in MEC-enabled integrated aerial-terrestrial vehicular networks: A reinforcement learning approach," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–14, 2022.
- [43] J. Ren, G. Yu, Y. Cai, and Y. He, "Latency optimization for resource allocation in mobile-edge computation offloading," *IEEE Transactions on Wireless Communications*, vol. 17, no. 8, pp. 5506–5519, 2018.
- [44] Y. Zhou, L. Tian, L. Liu, and Y. Qi, "Fog computing enabled future mobile communication networks: A convergence of communication and computing," *IEEE Communications Magazine*, vol. 57, no. 5, pp. 20–27, 2019.
- [45] T. Dang and M. Peng, "Joint radio communication, caching, and computing design for mobile virtual reality delivery in fog radio access networks," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 7, pp. 1594–1607, 2019.
- [46] Y. Liu, J. Zhou, D. Tian, Z. Sheng, X. Duan, G. Qu, and V. C. M. Leung, "Joint communication and computation resource scheduling of a UAV-assisted mobile edge computing system for platooning vehicles," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–16, 2021.
- [47] T. Zhang, Y. Xu, J. Loo, D. Yang, and L. Xiao, "Joint computation and communication design for UAV-assisted mobile edge computing in IoT," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 8, pp. 5505–5516, 2020.
- [48] S. Chen, Y. Zheng, W. Lu, V. Varadarajan, and K. Wang, "Energy-optimal dynamic computation offloading for industrial IoT in fog computing," *IEEE Transactions on Green Communications and Networking*, vol. 4, no. 2, pp. 566–576, 2020.
- [49] A. Geist, C. Brewer, M. Davis, N. Franconi, S. Heyward, T. Wise, G. Crum, D. Petrick, R. Ripley, C. Wilson *et al.*, "Spacecube v3. 0 NASA next-generation high-performance processor for science applications," 2019.
- [50] T. Flatley, A. Giest, D. Petrick, G. Crum, and M. Davis, "SpaceCube v3. 0 single-board computer," Patent 11,026,331, June, 2021.
- [51] "Q7S specifications - datasheet," <http://xiphos.com/wp-content/uploads/2015/06/XTI-2001-2020-e-Q7S-Spec-Sheet.pdf>, 2020.
- [52] "Q8S specifications - datasheet," <http://xiphos.com/wp-content/uploads/2020/06/XTI-2001-2025-f-Q8S-Rev-B-Spec-Sheet-1.pdf>, 2020.
- [53] "RAD5545 SpaceVPX single-board computer," <https://www.baesystems.com/en-media/uploadFile/20210404061759/1434594567983.pdf>, 2017.
- [54] "CFC-500: Compact on-board computer - datasheet," <https://www.innoflight.com/product-overview/cfcs/cfc-500/>, 2020.
- [55] "Integrated avionics unit - datasheet," [https://www.moog.com/content/dam/moog/literature/Space\\_Defense/spaceliterature/avionics/moog-integrated-avionics-unit-datasheet.pdf](https://www.moog.com/content/dam/moog/literature/Space_Defense/spaceliterature/avionics/moog-integrated-avionics-unit-datasheet.pdf), 2020.
- [56] "iX5-100 spacecloud solution," <https://unibap.com/en/our-offer/space/spacecloud-solutions/ix5100/>.
- [57] F. C. Bruhn, N. Tsog, F. Kunkel, O. Flordal, and I. Troxel, "Enabling radiation tolerant heterogeneous GPU-based onboard data processing in space," *CEAS Space Journal*, vol. 12, no. 4, pp. 551–564, 2020.
- [58] "iX10-100 spacecloud solution," <https://unibap.com/en/our-offer/space/spacecloud-solutions/ix10100/>.
- [59] "SpaceCube v2.0 hybrid data processing system," [https://spacecube.nasa.gov/SpaceCube\\_v2\\_BriefPDF](https://spacecube.nasa.gov/SpaceCube_v2_BriefPDF).
- [60] N. G. Franconi, A. D. George, A. D. Geist, and D. Albajies, "Signal and power integrity design methodology for high-performance flight computing systems," in *Proc. IEEE SCC*, 2021, pp. 27–38.
- [61] I. Del Portillo, B. G. Cameron, and E. F. Crawley, "A technical comparison of three low earth orbit satellite constellation systems to provide global broadband," *Acta Astronautica*, vol. 159, pp. 123–135, 2019.
- [62] C. Carrizo, M. Knappek, J. Horwath, D. D. Gonzalez, and P. Cornwell, "Optical inter-satellite link terminals for next generation satellite constellations," in *Proc. Free-Space Laser Communications XXXII*, vol. 11272, 2020, pp. 8 – 18.
- [63] V. Frost and B. Melamed, "Traffic modeling for telecommunications networks," *IEEE Communications Magazine*, vol. 32, no. 3, pp. 70–81, 1994.
- [64] R. Radhakrishnan, W. W. Edmonson, F. Afghah, R. M. Rodriguez-Osorio, F. Pinto, and S. C. Burleigh, "Survey of inter-satellite communication for small satellite systems: Physical layer to network layer view," *IEEE Communications Surveys Tutorials*, vol. 18, no. 4, pp. 2442–2473, 2016.
- [65] D. Frigioni, A. Marchetti-Spaccamela, and U. Nanni, "Fully dynamic algorithms for maintaining shortest paths trees," *Journal of Algorithms*, vol. 34, no. 2, pp. 251–281, 2000.
- [66] P. G. Franciosa, D. Frigioni, and R. Giaccio, "Semi-dynamic breadth-first search in digraphs," *Theoretical Computer Science*, vol. 250, no. 1-2, pp. 201–217, 2001.
- [67] C. Demetrescu, D. Frigioni, A. Marchetti-Spaccamela, and U. Nanni, "Maintaining shortest paths in digraphs with arbitrary arc weights: An experimental study," in *Proc. ACM WAE*, 2000.
- [68] Sunita and D. Garg, "Dynamizing dijkstra: A solution to dynamic shortest path problem through retroactive priority queue," *Journal of King Saud University - Computer and Information Sciences*, 2018.
- [69] L. Lin, C.-G. Yan, C.-J. Jiang, and X.-D. Zhou, "Complexity and approximate algorithm of shortest paths in dynamic networks," *Chinese Journal of Computers*, vol. 30, no. 4, p. 608, 2007.