Intelligent Sensing, Communication, Computation and Caching for Satellite-Ground Integrated Networks

Yongkang Gong, Member, IEEE, Haipeng Yao, Senior Member, IEEE, and Arumugam Nallanathan, Fellow, IEEE,

Abstract—Satellite-ground integrated networks (SGINs) are regarded as promising architectures for sensing heterogenous measurements, reducing network congestion and for providing pervasive intelligence services in support of terrestrial users. In SGINs, both low, medium and geostationary orbit based satellites are deployed for achieving global coverage and for supporting communication services for terrestrial users. However, the integration of task sensing, computation, communication and caching functionalities is quite challenging, which leads to low real-time task processing capabilities in dynamically fluctuating complex network environments. Hence, we propose an edge-intelligence-driven collaborative SGIN architecture and construct a framework relying on multiple planes, including the sensing plane, forwarding plane, control plane, intelligence plane and application plane. Furthermore, we construct an integrated sensing, communication, computation, caching and intelligence service paradigm for SGINs. Moreover, a centralized and distributed integrated learning framework is established for computation offloading and network function virtualization. Our simulation results show that the proposed learning framework is superior to the existing baseline methods in terms of its average transmission rate. Finally, we list a suite of potential research directions and solutions.

I. INTRODUCTION

Traditional terrestrial networks fail to provide seamless, uniform coverage in regions of low user-density and cannot support the explosive proliferation of mobile users, services and applications [1]. Fortunately, satellite-ground integrated networks (SGINs) succeed in filling the coverage-holes in support of seamless global coverage, even in remote areas [2].

Hence, there is an increasing number of research contributions on the integration of satellite communications, space networks as well as of terrestrial networks. In order to achieve efficient resource sharings, it is necessary to establish holographic SGINs to provide sensing, communication, computation and caching services for the users. Specifically, Liu et al. [3] systematically classified the related integrated sensing and communication (ISAC) methods and defined the salient performance metrics as well as the limits of ISAC. Mai et al. [4] established a Network AI paradigm and utilized the in-band telemetry technology for collecting network information from the forwarding plane, which was capable of incorporating the network status into the data packet header to guarantee a certain minimum level of network monitoring granularity. However, the elements of SGINs are extremely heterogeneous and of high-dimension. This fact results in grave open challenges.

Artificial intelligence (AI) techniques have been widely introduced into communication, computation and caching systems, which tend to operate in the face of uncertainty. Specifically, Gu et al [5] proposed a learning-based intelligent network function virtualization and traffic flow scheduling mechanism for terrestrial cloud computing scenarios. As a further advance, Wang et al. [6] introduced a distributed deep reinforcement learning (DRL) framework for reducing the communication overhead in edge AI scenarios. Furthermore, Dai et al. [7] conceived secure content caching policies via integrating DRL with blockchain in vehicular networking scenarios. As a recent advance development, Qiu et al. [8] harnessed a blockchain-based collaborative Q-learning method for enhancing both the learning efficacy and privacy protection. However, the aforementioned centralized or distributed DRL algorithms rely on high-overhead interaction and computation. Furthermore, it is quite hard to find globally optimal solutions, while only relying on local observations.

As another limitation, traditional network planes, including the application plane, data plane and control plane, represent closed interfaces, which are not programmable, thus result in high construction as well as maintenance cost, and poor scalability. Hence, it is necessary to develop innovative network plane structures, which can achieve efficient resource scheduling and accurate task sensing [9]. Based on the aforementioned traditional structure, new sensing and intelligence planes can be introduced to facilitate fine-grained multi-dimensional network resource sensing as well as heterogeneous task sensing. These solutions are also expected to carry out load balancing and congestion control as well as to construct flow tables by harnessing powerful AI-based frameworks. Finally, the integrated sensing-forwarding-control-intelligence-application plane structure of SGINs has to be designed for facilitating autonomous learning, adaptation and optimization.

To resolve the aforementioned challenges, we propose a collaborative centralized-distributed framework for supporting...
sensing, communication, computation and caching with the assistance of AI in SGIN scenarios. Our main contributions are as follows.

- **We conceive a sensing, communication, computation, caching and intelligence-aided SGIN architecture, and establish a matching sensing-forwarding-control-intelligence-application plane structure for resource and task sensing, computation offloading and distributed cooperative caching.**

- Next, we propose a collaborative centralized-distributed collaborative learning framework for multi-dimensional network resource sensing and task sensing, communication resource pre-allocation, computation offloading and distributed cooperative caching. This solution is capable of further accelerating the intrinsic integration of communication, networking, computation, caching, routing and sensing resources.

- Finally, we develop an SGIN prototype, which consist of controllers having multiple layers, levels and domains. Moreover, our simulation results show that the proposed learning framework is superior both to the traditional baselines, such as random scheduling (RS) as well as to a greedy policy relying on full offloading (GPFO) and to actor-critic (AC) solutions.

**II. COLLABORATIVE SGIN ARCHITECTURE**

As shown in Fig. 1, the collaborative SGIN architecture consists of three segments, including the space-aerial-ground segment. Specifically, the space segment consists of low earth orbit (LEO), medium earth orbit (MEO) and geostationary earth orbit (GEO) satellites [10], provided for the seamless coverage of deserts, oceans and mountains. The aerial segment includes multiple high-altitude platforms (HAPs), civil aircraft and unmanned aerial vehicles (UAVs) at altitudes spanning from 0.1 KM to 30 KM. The ground segment refers to the terrestrial cellular networks supporting high rate services. The resultant SGIN architecture amalgamates the advantages of different segments for providing an extended 3D coverage range, flexible deployment, robustness and high system capacity for a massive number of users in order to achieve fair and equitable resource sharing. The specific functions of each SGIN segment are presented next.

**A. Space Segment**

The GEO satellites are at an altitude of 35786 KM over the equator, and their operational cycle is synchronized with the planet’s rotation cycle. The GEO constellation has a wide coverage range. Explicitly, each GEO satellite covers 42% of the earth and hence 3 or 4 satellites can achieve global coverage, except for the two poles. Additionally, each terrestrial station can benefit from uninterrupted communication via GEO constellations, which have a simple networking architecture. They can be readily used for communication, meteorology, broadcasting and data forwarding.

By contrast, the MEO satellites have an altitude of 2000-36000 KM, and a single satellite covers 12%-38% of the planet. Hence, dozens of MEO satellites are required for covering the globe. However, the resultant communication period is limited to about 100 minutes between each user and MEO satellite because of their relative motion.

Finally, the orbit altitude of LEO constellations is less than 2000 KM and the coverage area is smaller than that of MEO and GEO satellites. The LEO satellites move at a high speed relative to ground users, and their constellation has the following characteristics. Firstly, we need thousands of satellites to form large-scale constellations for global coverage. Furthermore, the time-varying inter-satellite links (ISLs) lead to a complex constellation. Moreover, the low orbit altitude has a low propagation delay, but a single satellite’s coverage duration may be less than 5 minutes. Hence, it needs a seamless routing mechanism among multiple satellites to guarantee the quality of service.

**B. Aerial Segment**

The aerial segment relies on passenger aircraft for information collection, transmission and processing as a complement of terrestrial networks. These mobile nodes may be divided into HAPs and UAVs. Specifically, solar-charged HAPs can be deployed at 10-50 KM altitudes [11], which facilitates a coverage radius of 50 KM for ground users. Additionally, they can also be equipped with aerial base stations to provide classic telecommunication services to cover remote suburbs, the oceans, mountains and deserts. As an anecdotal example, the softbank company cooperated with AeroVironment to set up the HAPSMobile project having a 19 KM flight altitude. The stratospheric platforms utilized the ‘Grob520’ aircraft to test the communication performance in Germany.

Furthermore, the flight altitude of UAVs is limited by aviation authorities typically below 300 m, which can help assist the terrestrial networks in case of rallies and sporting events to enlarge the coverage range or to cope with the traffic peaks. Simultaneously, several enterprises and operators embarked on launching UAV projects, such as the ‘European Perfume’, the ‘Nokia F-Cell’ and the ‘Huawei Digital Air’.

**C. Ground Segment**

In the ground segment, the heterogenous terrestrial networks include cellular and mobile ad-hoc networks, WLANs and device to device (D2D) communication. Moreover, the operational wireless networks can have ultra-high peak rate, high connection density, low end-to-end transmission delay and also support high mobility. Furthermore, they operate in multiple frequency bands, utilize massive multiple-input multiple-output schemes and multiple-carrier techniques to achieve high spectral efficiency and high system capacity. They also support multi-access edge computing, software defined networking and network slice management to achieve efficient resource management and scheduling. However, they still face challenges, such as insufficient spectral resources, high network construction costs and limited base station coverage range.

**D. Decentralized Collaborative Satellite-Ground Scenario**

To solve the above challenges, we propose a collaborative decentralized SGIN framework for intelligent sensing, communication, computation and caching relying on a sophisticated task-oriented routing mechanism. Subsequently, we propose a base to satellite (B2S) resource pre-allocation scheme for overcoming any intermittent channel blocking.
Furthermore, an ISL resource management technique will be proposed for parallel transmissions as well as for collaborative services in order to improve the SGIN capacity. Then, we will propose a distributed deep reinforcement learning aided computation offloading framework for beneficially scheduling the satellite CPU cycle frequency, in order to minimize the total energy consumption and delay for stationary users. Next, we present a decentralized model-based energy harvesting solution for maximizing the sum-rate of mobility-aware HAP networks. Finally, based on the content distribution requests, the network topology, traffic distribution and file popularity inferred from multiple servers, we harness distributed learning for optimizing the caching contents while reducing the teletraffic and minimizing the delay. Simultaneously, we propose novel cooperative D2D caching policies for improving the content placement location with the aid of feedback information gleaned from the environment, which can be obtained by exploiting the social relationships among users as well as their physical distances and file popularity.

III. THE FIVE-LAYER SGIN STRUCTURE

The five-layer SGIN architecture is constituted by the sensing plane, forwarding plane, control plane, intelligence plane and application plane, whose specific functions will be presented in the following subsections.

A. Sensing Plane

As shown in Fig. 2, the sensing plane analyses the application tasks and network state information in support of the fine-grained multi-dimensional network resource sensing and intelligent task sensing via the red arrows from the forwarding plane and application plane. Specifically, the network nodes are not programmable, hence the corresponding sensing rules are embedded into the nodes. Their state and data flow can only be sensed via passive monitoring. Accordingly, spontaneous measurement rules or monitoring instructions can be utilized to detect the state of the network nodes, their traffic characteristics and performance. However, passive sensing [12] often fails to accurately infer the network state and task types. By contrast, when the network nodes have intelligent programmable capabilities, the sensing planes become capable of acquiring fine-grained multi-dimensional resource awareness, concerning any service level information, packet detection success information and anomalous data flow detection. Moreover, the sensing plane can introduce intelligent network flow classification techniques for achieving accurate task sensing based on the packet length and data flow duration.

B. Forwarding Plane

The forwarding plane consists of all physical entities in SGINs, which can support the data forwarding, processing and
Fig. 2: The integrated sensing, forwarding, control, intelligence and application plane structure of SGINs.

monitoring. Next, a software defined network [13] controller is introduced into this SGIN and each network element only undertake data forwarding instead of embedding the upper network control functionality, because the corresponding control rules are released via the OpenFlow southbound interface. Hence, each network element can forward the received data packet according to the released control rules, and these programmable network elements can infer the specific network states for achieving multi-dimensional resource and heterogeneous task sensing in the sensing plane.

C. Control Plane

The control plane is utilized for connecting the intelligence plane with the forwarding plane via the red arrow in Fig. 2, whose function is to issue the control policy from the intelligence plane to the forwarding plane and maintain the global control of SGINs. Next, we propose a decentralized network structure for further decomposing the time-space distribution and network element function, which consists of multiple ground controllers, satellite controllers and main controllers. Furthermore, the associated distributed intra-domain control system includes satellite controllers, ground controllers and multiple network nodes. By contrast, the inter-domain control system has multiple main controllers, which can be harnessed for managing the intra-domain control system and for cross-domain cooperation in SGINs.

D. Intelligence Plane

The intelligence plane has to generate the optimal control policy based on the SGINs’ state maps and on the corresponding task types. Specifically, the intelligence plane gathers, stores, manages and extracts the network state information and data flow information for constructing the global network state maps, which can be utilized for further optimizing and orchestrating the network resources for improving the quality of service (QoS) for ground users.

E. Application Plane

The application plane constitutes the top layer of SGINs and it directly interacts with the network tasks associated with specific requirements, such as supporting ultra dense connections, high-mobility and high-throughput, as well as near-real-time and high-reliability operation. Moreover, the application plane can provide network interfaces for the associated tasks, while the sensing plane may classify the different task types. Next, the task requirements can be forwarded to the intelligence plane for supporting resource orchestration.
IV. INTELLIGENT CENTRALIZED-DISTRIBUTED LEARNING FRAMEWORK

As shown in Fig. 3, we propose a collaborative centralized-distributed networking framework, which can be utilized for sensing, communication, computation and caching functions in SGINs. They are presented as follows.

A. Sensing Functions

- Fine-Grained Multi-Dimensional Network Resource Sensing

The multi-dimensional network resources of SGINs exhibit heterogeneity and dynamic fluctuations, which makes it hard to sense the associated multi-domain network resources by employing traditional resource sensing. Hence, a cooperative convolutional neural network (CNN) based multi-domain resource sensing model is established to generate a multi-dimensional network resource storage matrix, which may also be compressed depending on the specific network environment, user behaviors and network states, such as any service level information and anomaly data flow.

- Intelligent Sensing of Heterogeneous Tasks

In SGINs, the sensing granularity may affect the efficacy of network control. The recognition of traditional task characteristics based on tags depends on fixed matching rules, which makes it hard to achieve accurate recognition of heterogeneous multi-dimensional tasks. Hence, we can establish differentiated representation models based on accurate task flow characteristics and introduce deep reinforcement learning-based intelligent network flow classification, which can be exploited for learning the data packet length and data flow duration. Moreover, this can distill the heterogeneous features into high-dimensional characteristic vectors, maximize the separation among the classes of heterogeneous tasks and achieve accurate task sensing.

B. Communication Functions

- G2S Resource Pre-Allocation

When the ground users transmit tasks to the satellites (G2S), each channel may encounter blockage. Based on the beamforming vector sets, we can predict any future channel blockage probability via a recurrent neural network model. Subsequently, we can sort the channel blockage probabilities into an ordered sequence and assign an available channel to each user, while bearing in mind the blockage probability, the corresponding transmission delay, improving the transmission rate and reducing the total system-level energy consumption metrics. Hence, the proposed G2S resource pre-allocation scheme may assign more transmission power and bandwidth to those specific channels, which are more likely to be blocked. This can reduce both the data transmission delay and the
energy consumption. The ISL resource management schemes are presented next.

- ISL Resource Management Schemes

The transmission links among multiple satellites are termed as inter-satellite links (ISL), which have to efficiently utilize the backhaul links for improving the satellite nodes’ collaboration gains. First, we analyse the task priority in each satellite’s coverage area and establish the service level agreement-based network model. Then, we can harness the proposed collaborative central-distributed framework for modeling each satellite by an agent, and optimize the computational resources via interacting with the associated network environment, such as the maximum transmission rate, channel gains and random tasks. Subsequently, the resultant state, action and reward space can be invoked for training the distributed agents to improve the cooperative gains among multiple agents.

C. Computation Functions

We classify the associated computation scenarios into two types, i.e., distributed static users and mobility-aware task scheduling mechanisms. The specific functions are presented in the following.

- Resource Scheduling for Distributed Static Users

For a single edge node scenario associated with static users, a model-based computation offloading framework can be utilized for optimizing the offloading decisions, CPU cycle frequency and transmission power in the face of time-varying channel gains, which can further minimize the delay and energy consumption. Next, since there is high interference among the edge nodes, a multi-agent deep deterministic policy gradient algorithm can be conceived for optimizing the communication as well as computational resources and assign the most appropriate channel to each ground user, which is beneficial for reducing the mutual interference among multiple edge networks.

- Mobility-Aware Task Scheduling for Rechargeable Users

When the ground users offload their tasks to remote servers, such as satellites or HAPs, the stochastic task arrivals, the potential HAP locations and the limited battery storage of ground users affect the QoS. Hence, a centralized-distributed integrated multi-agent proximal policy optimization technique is proposed for task scheduling, for adapting to the dynamic HAP locations and arranging for the payments of ground users, which can further maximize the sum rate. The resultant framework can process stochastic tasks originating from the sensing plane such as latency-sensitive as well as computation-intensive tasks, and partially offload them to remote satellites or HAPs, which can further boost the integration of the sensing plane, application plane and intelligence plane.

D. Caching Functions

In SGINs, the limited computation and caching resources cannot fulfill the actual user requirements, such as requesting popular caching contents and reducing the service delay. Due to time-variant content updates and user demand transformations, it is difficult to collect massive amounts of useful data for predicting the content popularity. Hence, we propose non-stationary popularity prediction-based ISL caching in order to reduce the ratio of caching repetition and to improve the caching hit ratio.

- Non-Stationary Popularity Prediction-Based ISL Caching

In practical scenarios, the content popularity is constantly changing and the total number of content requests from past to present cannot accurately reflect the future content updates. Hence, we can model the content popularity based on traffic bursts, file life cycles and ISL location in support of proficient caching replenishment. For improving the caching utilization efficacy, a DRL-based federated learning mechanism can be utilized to enhance the node cooperation capability and to protect the data privacy.

- Distributed Cooperative Caching for ISL and D2D

In the proposed SGIN scenario, we can determine the number of caching user sets to pre-cache file contents. Specifically, we propose a distributed learning-based D2D caching policy to determine the user sets in terms of file popularity and D2D distance. Next, this policy can determine the optimal reward function via interacting with the network environment. Finally, the framework can help efficiently exploit the D2D caching space and assist the ISL caching, which is beneficial for inferring the optimal caching replacement policy of SGINs.

V. SIMULATION RESULTS AND PERFORMANCE ANALYSIS

In this section, we report on our simulation experiments for validating the efficiency of the proposed learning framework compared to the AC, GPFO and RS benchmarks.

A. The SGINs Considered

As shown in Fig. 4, we designed an SGIN prototype to verify the aforementioned sensing, communication, computation and caching functions in terms of multiple scenarios. The case study consists of a data plane simulation platform, as well as a control plane and demonstration verification plane. Specifically, the data plane includes virtual SGINs, which can be interacted with via SGIN sensing, communication, computation and caching systems. Next, the network state extraction can be achieved via task and resource sensing actuators. Moreover, the control plane consists of centralized ground controllers and distributed satellite controllers, which can be arranged by relying on multi-layer and multi-domain deployment. Subsequently, the control plane extracts the underlying communication, network, computation, caching, routing and sensing resources in order to achieve beneficial matching between the tasks and resources.

B. Performance Analysis

In the simulation scenario, there are a total of 200 ground users, 5 HAPs and multiple LEO satellites. The computational capability of each HAP and LEO is uniformly distributed at $[3 \cdot 10^8 \text{ cycle/s}, 6 \cdot 10^8 \text{ cycle/s}]$ and $[1 \cdot 10^8 \text{ cycle/s}, 2 \cdot 10^8 \text{ cycle/s}]$ [14]. Additionally, the local processing capability of each ground user is uniformly distributed in the interval of $[1000 \text{ cycle/s}, 3000 \text{ cycle/s}]$. Subsequently, to characterize the SGIN sensing, forwarding, control, intelligence and application planes structure, as shown in Fig. 5, we explore the impact of task generation rate on the average transmission rate. When more tasks are generated, the proposed learning framework has a higher average transmission rate than the three baseline
methods, since it utilizes a centralized structure for global network management, and makes offloading decisions by relying on prompt distributed execution, which helps each ground user to adapt to time-varying network environments and task types.

VI. FUTURE RESEARCH DIRECTIONS

Although the proposed SGIN architecture can support sensing, communication, computation and caching functions, some future challenges still have to be tackled, specific details are shown as follows.

A. Complex Constellation Configuration and Dynamic Topology

The constellation configuration among ISLs and the dynamic network topology make it hard to determine the optimal computation offloading policy. Moreover, the evolution from a single-point connection to SGINs relying on complete satellite constellations having prompt routing convergence is a significant research challenge, especially in support of millions of ground users and thousands of links. Hence, an efficient multi-layer, multi-domain collaborative routing scheme based on reinforcement learning has to be deployed to achieve prompt routing convergence.

B. Data Privacy Protection

The lack of mutual trust potentially hinders resource sharing and content downloaded, because potential attacks may cause private data leakage and poor QoS during offloading. Block-chain aided federated learning protection [15] may come to rescue for preventing data tampering, for setting up distributed ledgers and for protecting data privacy. Specifically, while performing computation offloading, we can design a distributed task transaction protocol to protect the data privacy, and then only transmit the pre-trained model parameters instead of local data to the remote server.

C. Fault Localization and Routing Repair

The large-scale spatial dimension and limited satellite processing capability may result in poor network resilience. Hence, improving the fault localization accuracy and network robustness is another one important challenge in the face of limited network resources. To this effect, a real-time global network state sensing model must be constructed for accurate fault localization and intelligent routing repair. Specifically, we can utilize the principal component analysis technique to analyze the fault types and construct the backup routing tables to achieve the optimal fault path switch.

VII. CONCLUSIONS

Intelligent SGIN sensing, communication, computation and caching architectures were proposed based on a sensing-forwarding-control-intelligence-application plane structure. A centralized-distributed collaborative learning framework was...
also conceived for fine-grained multi-dimensional network resource sensing as well as heterogeneous tasks sensing, computation offloading and distributed cooperative caching. Moreover, an SGIN use case was developed for validating the related network functions. Subsequently, simulation results were presented to show that our proposed learning framework achieves superior average transmission rate compared to the AC, RS and GPFO benchmarks. Finally, the paper listed a range of novel research directions for future SGINs.

REFERENCES


