



Views & Comments

Space Computing Power Networks: Fundamentals and Techniques

Linling Kuang^a, Yuanming Shi^{b,*}, Kai Liu^a, Chunxiao Jiang^a^a Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China^b School of Information Science and Technology, ShanghaiTech University, Shanghai 201210, China

1. Introduction

In the past few decades, satellite technologies have advanced considerably, enabling seamless connectivity and wide coverage across the world. As a result, remote sensing now permits the collection of massive space information, which can be leveraged to support time-sensitive services and countermeasures to serious emergencies such as natural disasters [1,2]. For example, China's Fengyun series of meteorological satellites designed for weather forecasting monitored the whole life cycle of typhoon Doksuri in 2023, precisely predicting the extreme heavy rain and mitigating its effects. However, traditional approaches that directly download raw data from remote sensing for further processing present several challenges. First, although ground station networks are already widely deployed, the issue of a short visible window (about 10 min) remains in communication between satellites and ground stations, due to the fast movement of satellites. Each satellite must wait for a period of time before accessing a ground station again, which poses a further timeliness challenge to the downloading of raw data. Second, with the adoption of high-frequency bands, the satellite-to-ground link rate has reached the gigabits-per-second level. However, with the improvement of remote sensing image resolution and the use of inter-satellite high-speed laser links (> 10 gigabits per second (Gbps)), the satellite-to-ground link is still a transmission bottleneck. Third, directly transmitting the high resolution of remote sensing raw data often presents severe privacy concerns. Therefore, it is essential to develop effective solutions to empower fast and trustworthy onboard information processing, with the space computing power network (Space-CPN) emerging as a promising architecture.

Instead of focusing solely on terrestrial networks, Space-CPN expands the concept of a computing power network [3] by incorporating innovations in satellite networks, as illustrated in Fig. 1. With advanced edge computing techniques, mature satellite communications, and novel onboard computing payloads, Space-CPN integrates the communication and computation capabilities of various types of satellites, including low-Earth-orbit (LEO), medium-Earth-orbit (MEO), and geostationary-Earth-orbit (GEO) satellites. In Space-CPN, GEO/MEO/LEO satellites typically serve as space computing centers

targeting emergency scenarios, while the ground stations act as terrestrial computing centers for scenarios with high computing power demands. In this way, computing power can be scheduled flexibly to achieve secure, fast, and accurate onboard intelligent data processing. Acting as a multilayer satellite-based distributed computing architecture, Space-CPN can support various onboard intelligent computation tasks. For example, European Space Agency (ESA) Φ-lab collaborated with Oxford University and Trillium Technologies regarding onboard artificial intelligence (AI) model training and presented their results for AI aboard Earth-observation satellites in 2023. By employing the initial model RaVAEn to compress raw images, the researchers successfully trained a machine learning model for cloud detection during the D-Orbit mission. Regarding onboard AI model inference, cloud cover assessment and removal for noisy remote sensing data are an important part of many real-time applications, including weather forecasting and disaster monitoring. The KATESU project evaluated two different corresponding deep neural networks on two onboard satellite computing platforms: Teledyne e2v's LS1046 and Xilinx ZedBoard. The outstanding performance of this project provides strong evidence for the possibility of improving onboard data quality and the efficiency of downstream processing. Overall, these preliminary achievements have demonstrated the great potential of Space-CPN for enabling future onboard intelligent services. Nevertheless, the further development of Space-CPN on a larger scale still presents several challenges.

The first challenge lies in the design of communication principles for Space-CPN. By integrating space and terrestrial networks, Space-CPN enables seamless global connectivity and provides a huge system capacity to support conventional data-oriented services. However, as the requirements of space services become complex and diversified, the need for onboard data processing also increases, which calls for the design of corresponding communication principles to support these computation tasks. Traditionally, the design of wireless data transmission is mainly based on Shannon's theorem. Using bit-based metrics such as channel capacity to measure the performance, reliable bit-level data-oriented transmission can be achieved [4]. Nevertheless, this theorem is no longer suitable for the current Space-CPN, in which communication principles are intended to serve specific computation tasks. Taking remote sensing images as an example, the Sentinel-2 satellites under the ESA's Copernicus program capture approximately 1.6 terabytes of high-resolution optical image data in each orbit during each Earth observation cycle. Directly transmitting such a huge amount of raw data to ground stations for further downstream processing would inevitably

* Corresponding author.

E-mail address: shiyym@shanghaitech.edu.cn (Y. Shi).

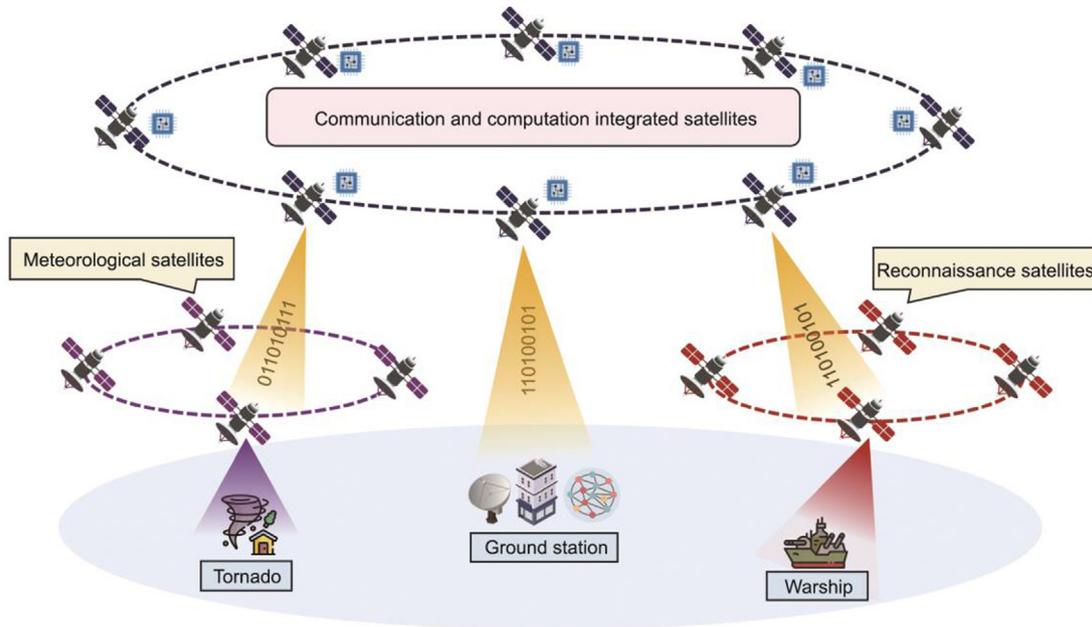


Fig. 1. The Space-CPN architecture.

lead to a severe communication bottleneck. In contrast, if the performance metrics of computation tasks are used to guide the design of information compression and transmission, only task-relevant information will be involved, allowing the communication overheads to be significantly reduced [5]. Therefore, there will be a paradigm shift from data-oriented to task-oriented communication in Space-CPN to support information compression and transmission for computation tasks.

The second challenge lies in the computing architecture of Space-CPN. Thanks to significant progress in satellite power systems, the successful onboard deployment of intelligent computation tasks with a specific power budget can be achieved. However, the onboard energy is still limited, making it difficult to centrally deploy large-scale high-performance computing units and large-scale computing clusters. For example, the power consumption of Nvidia's A100 graphic processing unit (GPU) can reach 400 W. Therefore, a low-power computing architecture is urgently needed for Space-CPN. Satellite onboard computing is conventionally implemented on processors following the von Neumann architecture, in which memory and processing units are separately deployed. This incurs the von Neumann bottleneck, however, in which information must be repeatedly shuttled between these parts, resulting in a huge energy cost. In neuromorphic computing, on the other hand, both memory and processing are governed by neurons and synapses, as this approach mimics the human brain, making it possible to avoid the overheads brought by the frequent transmission of data and instructions to improve energy efficiency. Moreover, the application of neuromorphic computing on satellites has been comprehensively evaluated [6]. The Neuro SatCom project from ESA provides one example: For a specific interference detection task, AI models on the non-neuromorphic platform Xilinx Versal consumed 136.9 J per problem, whereas those on the neuromorphic platform SpiNNaker consumed only 6.3 J per problem. These findings show the potential of the paradigm shift from the von Neumann computing architecture, with separate storage and computation, to the neuromorphic computing architecture, with integrated storage and computation.

The last challenge presented by Space-CPN involves the orchestration of various onboard computation tasks and the limited com-

munication, computation, and storage resources of dynamic space networks. First, regarding task scheduling, complex computation services require appropriate task decomposition in the heterogeneous environments of Space-CPN to realize efficient onboard execution. Typically, there are four types of computation parallelism tasks [7]: data parallelism, which distributes input data; tensor parallelism, which distributes the model weights; pipeline parallelism, which distributes the model layers; and mixture of experts, which activates partial model weights. Based on these parallelism strategies, it is crucial to first decompose the computation task and schedule the obtained sub-tasks across Space-CPN. Second, regarding resource allocation, the resources in Space-CPN can be generally divided into three categories: communication resources (e.g., time, space, spectrum, code, and orbit), computation resources (e.g., central processing units (CPUs), GPUs, and field-programmable gate arrays (FPGAs)), and storage resources (i.e., internal and external memory). Due to the ever-increasing demand for various space services, these resources are typically scarce, necessitating effective allocation approaches. However, Space-CPN exhibits large-scale, complex, and time-varying network topologies due to the diversified characteristics of thousands of satellites, including inclination, altitude, and operating direction [8], resulting in a highly dynamic environment that makes it difficult to manage communication, computation, and storage resources in a flexible way. Moreover, the uncertainties existing in Space-CPN further increase the difficulty of task scheduling and resource allocation. Due to the timeliness requirement for space information services and the high mobility of satellites, there will be fluctuations in requests for emergent tasks and in the volumes of remote sensing data, calling for robust solutions to model and tackle these uncertainties. Overall, it is necessary to identify the quantitative relationship between the computation task requirements and the scheduling of limited resources under the highly dynamic environments of Space-CPN in a robust manner.

2. Task-oriented communications

In this section, we present the task-oriented communication principles for seamlessly integrating communication and

computation in Space-CPN. This is achieved through developing a robust information bottleneck (RIB) for information compression by extracting the task-relevant information for remoting sensing data. Coded over-the-air computation (AirComp) is further developed to support ultra-low latency transmission by exploiting the waveform superposition of the wireless channels from satellites to ground stations.

2.1. RIB for information compression

In recent years, remote sensing satellites have garnered increasing attention due to their capability to swiftly acquire comprehensive coverage of targeted regions, enabling a wide range of important applications such as land-use surveys, urban studies, and hazard management. However, the use of high-resolution sensors in remote sensing satellites generates vast volumes of data that pose a significant challenge, as traditional data-oriented transmission to ground stations for further processing results in substantial communication overheads. To address this problem, numerous forthcoming satellite missions aim to implement computation tasks through collaboration between satellites and ground stations [9]. With advanced computing capabilities on satellites, onboard data processing is envisioned as a promising solution for compressing transmitted data and eliminating information redundancy, thereby improving communication efficiency. More specifically, instead of merely focusing on error-free transmission, the target of specific computation tasks serves as the metric to guide satellites in performing feature extraction, which is then sent to ground stations for fast post-processing. Unfortunately, satellite-to-ground links experience greater interference from factors such as cloud attenuation, tropospheric scintillation, and interference from other satellites, which significantly disrupt the transmission of representations, leading to substantial errors and inaccuracies. Moreover, cloud cover, which may block key features of the Earth's surface or distort signals through sun reflection, poses a major challenge for downstream remote sensing tasks that rely on distinguishing subtle inter-class and intra-class differences, leading to false classifications and decreased accuracy.

To achieve this goal, we turn to the information bottleneck (IB) principle [10], which maximizes the mutual information between the result and the label of the data sample for accuracy, while minimizing the mutual information between the feature and the input sample for compression. In Space-CPN, noise in the wireless channel is a significant factor that induces data distortion and affects inference accuracy. With the aim of further minimizing information distortion over noisy channels, we introduce a new principle called the RIB [11]. The RIB formulates the informativeness-robustness tradeoff in the encoded feature and aims to maximize the coded redundancy in order to improve robustness while retaining sufficient information for downstream inference tasks. Therefore, it improves communication robustness to channel variations without extra communication overhead. To address the computation intractability of mutual information, a tractable variational upper bound of the RIB objective is derived by means of a variational distribution parameterized by learning-based inference models.

Furthermore, to address the inherent and unavoidable cloud noise in satellite imagery of the Earth's surface, we adopt a Fisher-RIB (F-RIB) approach to increase the robustness to cloud noise interference by introducing Fisher information to the RIB, as shown in Fig. 2. Fisher information, which quantifies uncertainty in parameter estimation, serves as a compelling metric for assessing the sensitivity of extracted features to input data with latent noise [12]. The F-RIB refines the informativeness-robustness tradeoff by maximizing mutual information between the inference results and corresponding labels, while minimizing the sensitivity

of features to noise-induced input. Notably, Fisher information evaluates the parameter uncertainty caused by cloud noise, such as color distortion and brightness variation. By minimizing the trace of the Fisher information matrix of features, the F-RIB suppresses sensitivity to cloud interference and preserves feature consistency between cloudy and clear inputs, thereby increasing the network's robustness to cloud-induced perturbations.

2.2. Coded over-the-air computation for information transmission

Space-CPN is able to support various onboard satellite AI services. For AI model training, the cloud server, ground stations, and satellites can formulate a client-server architecture for onboard federated learning based on the remote sensing data collected by the satellites [13]. In each training iteration, the ground stations aggregate model updates from different satellites, and the new global model is calculated at the cloud server. For another type of computation task, when the satellites and ground station collaboratively finish an inference task, a vertically split neural network architecture can be employed [14], where the ground station must calculate an intermediate feature map based on the weighted sum of the local feature maps from satellites. Overall, it can be observed that local model aggregation is a key procedure for these satellite AI-based computation tasks. Nevertheless, as the number of satellites—together with the dimensions of the model parameters—grows, this procedure would inevitably lead to a severe latency issue. To achieve a more efficient aggregation procedure, AirComp acts as a promising solution [15] that aims to compute a class of nomographic functions over wireless channels. Based on the waveform superposition of the wireless channels from the satellites to the ground stations, the signals at each satellite are set to transmit simultaneously and directly realize a summation operation over the air. In this way, communication and computation can be integrated, further reducing the aggregation latency [16].

Nevertheless, affected by various factors such as strong interference and unstable weather conditions, the communication link between a ground station and LEO satellites typically exhibits a low signal-to-noise ratio (SNR). In addition, existing wireless communication systems for satellite communications mainly deploy digital modulation chips, so an arbitrary linear scaling modulation scheme is not supported. Therefore, it is impossible to directly employ uncoded AirComp in Space-CPN. Digital AirComp is thus proposed as an alternative to perform quantization on transmitted data. Among the existing quantization schemes, lattice-quantization-enabled coded AirComp [17] stands out as a superior solution because it does not rely on perfect channel state information on the satellite side, which is difficult to estimate due to the large propagation delay and high mobility. To be specific, a lattice can be characterized as a discrete additive subgroup, where each lattice point is formulated by the linear combination of corresponding basis vectors. Then, the lattice quantizer is defined as a map that quantizes a point to its nearest lattice point based on the Euclidean distance, which further leads to the concept of a Voronoi region denoting the set of all points quantized to the same lattice point. Based on the closure of the lattice, which ensures that the summation of any two lattice points is still a lattice point, each element that comes from the summation operation on any other two elements in the lattice still belongs to this lattice, as revealed in the formulation of a lattice point. This essentially provides a fundamental guarantee that the points in the lattice are naturally suited for aggregation in AirComp. The lattice quantization on the data makes it compatible with the digital modulation chips on the communication system of the satellites. Moreover, the Voronoi region over the decoded result makes it possible to resist the effect caused by noise and interference. Therefore, lattice-based

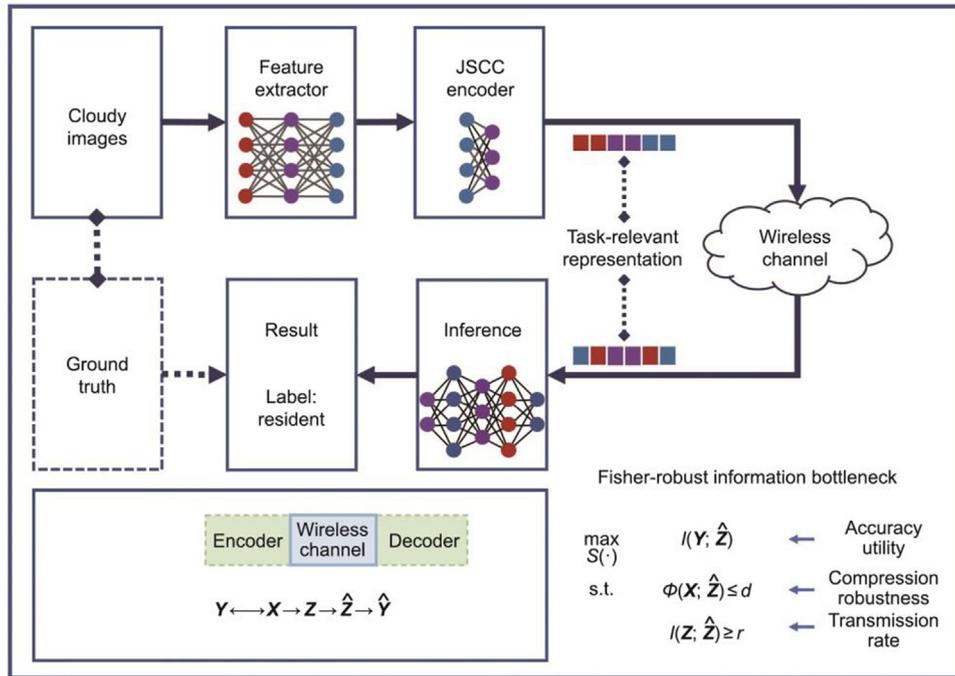


Fig. 2. The F-RIB employed in satellite inference tasks. JSCC: joint source and channel coding; $I(\cdot)$: mutual information; $\Phi(\cdot)$: Fisher information; Y : task objective; X : original data; Z : feature representation; \hat{Z} : noisy feature representation; \hat{Y} : result; d : dimension constraint; r : transmission rate constraint; S : feature extractor.

quantization provides an effective way to enable coded AirComp, and specific procedures can be described as follows. First, the local data on satellites passes through the lattice quantization, where a dither vector (which is generated independently based on a random uniform distribution) and a precoding factor are employed to ensure a uniform quantization error and control transmission power, respectively. Next, simultaneous transmission from all satellites is performed to realize the summation operation over the air with the channel noise injected. Finally, with lattice quantization and post-processing at the ground station, the decoded result can be obtained.

However, the implementation of this approach is accompanied by a high performance cost due to distortion of the aggregate data. First, the quantization performed on the local data introduces a quantization error between the transmitted data and the actual local data. Second, channel noise naturally adds a distractor during transmission in wireless channels. Together, these two factors result in a distortion that can be characterized by the mean squared error (MSE) between the aggregated model and the desired model in a typical collaborative training scenario. Inspired by Ref. [18], the convergence behavior of the learning process with the proposed transmission scheme can be analyzed. Under a non-convex objective and a decaying learning rate, after a certain number of global update steps, the upper bound of the weighted average norm of the global gradients will be affected by this MSE, which eventually leads to an error bound. To improve the learning performance, the denoising factors employed in post-processing can be jointly designed to mitigate the distortion of the aggregation.

3. Brain-inspired computing

In this section, we leverage neuromorphic computing models and hardware to present an energy-efficient onboard computing architecture for Space-CPN. To further train the brain-inspired spiking neural networks (SNNs) in Space-CPN, we propose satellite

federated and decentralized neuromorphic learning network architectures. We then suggest highly efficient inter-satellite link scheduling and satellite-ground coordinated transmission strategies for distributed model training.

3.1. Satellite neuromorphic computing principles

To implement energy-efficient neuromorphic computing, a corresponding hardware architecture is necessary. From a bottom-up view, the design of neuromorphic hardware can be interpreted in two levels: a unit architecture for neurons and a system architecture formulated by connected units [19]. Existing unit architectures can be generally divided into three categories. The first category is the digital unit, which leverages discrete values and Boolean logic-based gates to realize computation. The reconfigurability and flexibility of this scheme have facilitated many implementations of neuromorphic computing on FPGAs and application-specific integrated circuit chips. The second category is the analog unit, which harnesses the similarity between biological neurons and analog circuits. Based on the efficient realization of neuron behavior through the natural dynamics of circuits, this scheme is also widely adopted for neuromorphic implementation, including field-programmable analog arrays. The last category is the mixed implementation of the previous two schemes; this approach is usually employed to tackle the limitations of analog schemes.

In addition, there are two major types of system architectures. The first type comprises ideal architectures that strictly mimic the organization of biological neurons. Each neuron maintains its own state, and inter-neuron communication is carried out in parallel over massive synapses. The integrated memory and computation enables a high number of synaptic operations per second. However, as the number of neurons increases, the scalability of this architecture significantly decreases because of an increasingly large circuit overhead. Therefore, this architecture is not suitable for large-scale neuromorphic computing. The second type of

practical architecture provides a more reasonable strategy in which a number of neurons are grouped as a core, where the neurons keep their own data but share computation paths. This organization not only largely reduces the effective circuit area per neuron but also improves the memory efficiency through the sharing of common parameters. It is widely employed in neuromorphic processors, including International Business Machines Corporation (IBM)'s TrueNorth and Intel's Loihi.

The introduced hardware architectures provide a promising solution to realize onboard neuromorphic computing and improve the energy efficiency. Nevertheless, to fully unleash the potential of the integrated storage and computation, neuromorphic software that can well adapt to such a hardware architecture is required. To this end, the SNN, which takes another level of inspiration from biological neurons and incorporates the time dimension into its design, is widely adopted to serve as the architecture for neuromorphic intelligent models [20]. An SNN generally uses a leaky integrate-and-fire model to characterize neurons, in which a gradually decreased membrane potential is adopted to record the temporal dynamics. The weighted sum of the previous layer's output contributes to the membrane potential during forward propagation, and each neuron emits a spike as output if its membrane potential reaches a certain threshold. Based on this special neuron structure, the computational flow can be described as the transition of sparse, discrete, and single-bit spikes. Employing this spike-based communication and event-driven mechanism in an SNN leads to extreme sparsity in computation, which significantly improves the energy efficiency in comparison with conventional models. This approach has been employed in the design of various complex AI models, such as transformer and diffusion models.

Despite the strengths described above, the SNN's unique structure of neuromorphic neurons presents new challenges in its training process [20]. The discontinuity relation between output spikes and membrane potentials leads to a non-differentiable problem during backpropagation. Although a conventional approach of directly converting a well-trained artificial neural network (ANN) to an SNN can be used to avoid this problem, this is accompanied by a significant performance loss due to the conversion of high-precision activations. To solve this problem, the idea of a surrogate gradient can be adopted, in which smooth functions are leveraged to approximate the original discontinuous relation. In this way, the gradient can be conveniently estimated and backpropagation can be performed effectively.

3.2. Satellite neuromorphic computing and learning network architectures

Benefiting from the energy efficiency brought by neuromorphic computing, the SNN serves as a potential software architecture for onboard intelligent data processing. The question that remains is how to obtain a well-trained SNN based on the data collected by different satellites in constellations. Traditionally, centralized training of an SNN can be enabled by downloading raw remote sensing data to a ground station. However, the communication bottleneck caused by the massive raw data, scarce radio resources, and long propagation delay, together with the privacy concerns due to directly downloading sensitive information, make such an approach no longer suitable.

3.2.1. Satellite federated neuromorphic learning architecture

To this end, we propose a novel satellite federated neuromorphic learning framework for an onboard SNN, as illustrated in Fig. 3. Each training iteration can be characterized as five steps [13,21]. First, each satellite performs local updates based on its own dataset. Next, each orbit calculates an intra-orbit model based on the results of local training. Here, by taking the stable ring

topology formed by satellites in the same orbit and the high-speed inter-satellite links into consideration, it is possible to leverage the Ring-AllReduce algorithm to achieve parallel model aggregation in each orbit and further improve communication efficiency in this step. Then, satellite-ground links are used to realize intra-orbit model downloading at ground stations, which follows the global model aggregation at the cloud server. Finally, global model broadcasting is performed to ensure that each satellite maintains the global model for the next iteration. Nevertheless, the low data rate and short link duration of satellite-ground links impose severe transmission overheads during the model downloading procedure. To address this issue, we formulate the space-ground cooperative model downloading problem as a network flow problem due to the fact that, once the link-establishment conditions are satisfied, a single satellite can communicate with multiple ground stations simultaneously, while a ground station can also serve multiple satellites during a certain time. This further leads to a real-time satellite-ground model collaborative routing scheme, which significantly reduces transmission delays.

Another problem is that the heterogeneity among satellites hinders the effectiveness of the training process. This heterogeneity can be generally categorized into two types: statistical heterogeneity and hardware heterogeneity. Statistical heterogeneity arises from the diverse characteristics of the local datasets captured by satellites in different orbits and spectral environments, which result in variations in data distribution. More specifically, it refers to differences in the architecture and performance of the computing units on satellites. For example, satellites in different orbits may be equipped with different combinations of computing units (e.g., CPU, GPU, and FPGA), resulting in performance disparities between nodes. Hardware heterogeneity refers to differences in the architecture and performance of onboard computing units among satellites, which eventually lead to variations in computation capabilities—namely, differences in the number of local updates that can be performed in each training iteration.

To address these issues, we propose a federated heterogeneous hybrid averaging algorithm to improve the Ring-AllReduce averaging in the original framework. This is achieved by innovating weighted averaging to correct the convergence bias caused by statistical and hardware heterogeneity during the learning process. The data heterogeneity parameter is calculated based on the size ratios of satellite local datasets, and the hardware heterogeneity parameter is determined by the maximum number of local training rounds on each device. Integrating these two parameters, normalized averaging is then employed for intra-orbit model aggregation, thereby effectively mitigating the issues caused by satellite heterogeneity.

3.2.2. Decentralized satellite neuromorphic computing architecture

Although the satellite federated neuromorphic learning approach presented above provides a general framework for onboard SNN training, the frequent usage of satellite-ground links still poses a significant challenge. Even though the development of ground station networks is becoming mature, the short visible window (e.g., about 10 min for a LEO satellite) still results in a transmission bottleneck between satellites and ground stations. Moreover, although the presented network flow strategy effectively increases the utility of satellite-ground links, the relatively low data rate caused by the long distances and severe interference still poses a natural limitation on the maximum volume of flow in the formulated network. Therefore, it is necessary to achieve fully onboard decentralized learning for SNNs and mitigate the interaction between ground stations, which means that the aggregation of intra-orbit models must be realized via inter-plane inter-satellite links. However, in comparison with the stable ring topology formulated by satellites in the same orbit, inter-plane communication

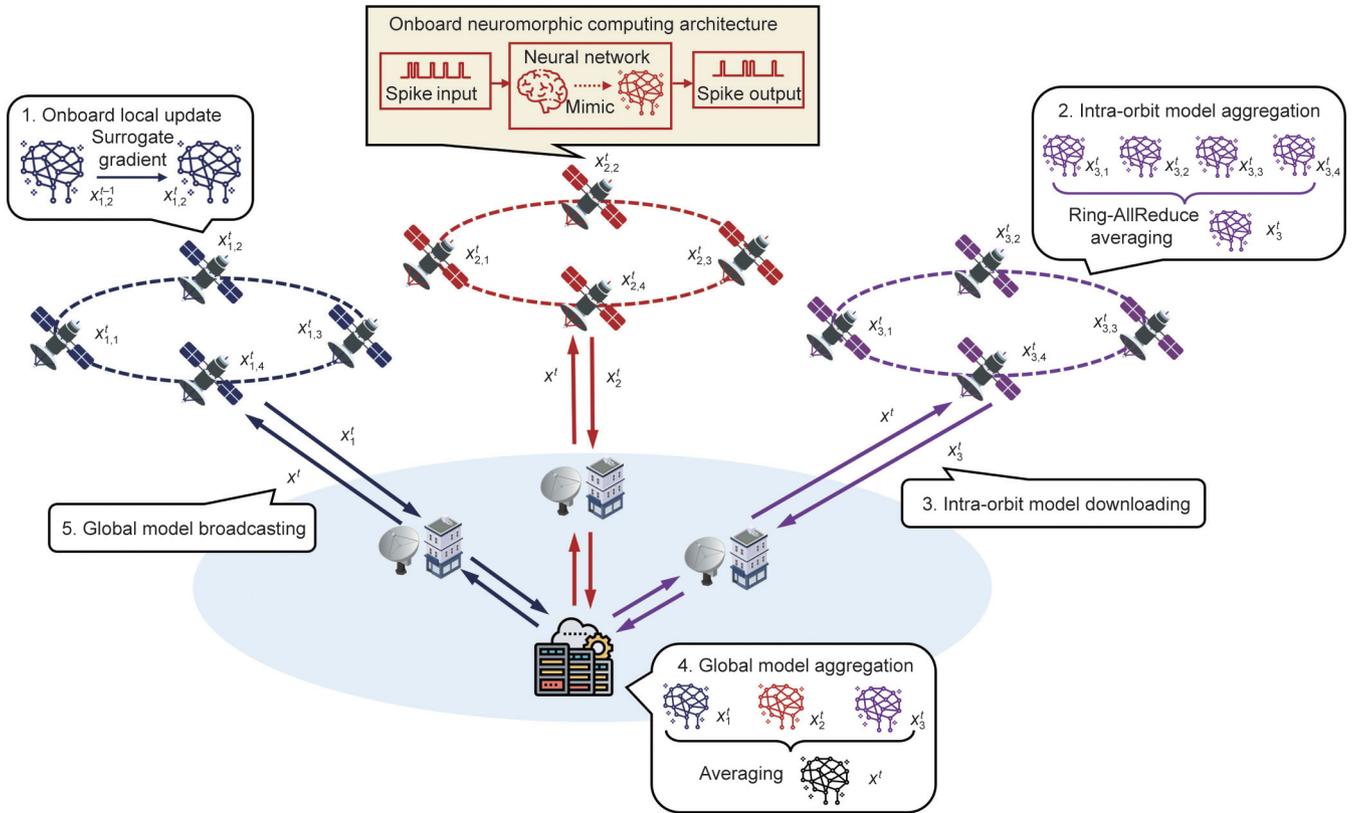


Fig. 3. Satellite neuromorphic computing and learning network architecture in Space-CPN. $\{x_{i,j}^t\}$: local model for satellite j at orbit i in the t th iteration; $\{x_i^t\}$: intra-plane model for orbit i in the t th iteration; x^t : global model in the t th iteration.

topology is relatively unstable because of the diversity of the movement characteristics of different orbit planes (e.g., altitude, inclination, and direction).

There are typically two inter-plane aggregation schemes for intra-orbit models. The first of these schemes relies on the principle of gossip averaging: Each orbit plane only communicates with its neighbors for the exchange of model information. This ensures a low communication overhead for model transmission. However, the slow diffusion of local updates inevitably decreases the convergence rate of training. The second scheme aims at global aggregation. By establishing an aggregation path across orbit planes, a global model can be derived so that the convergence behavior is not degraded in comparison with the centralized learning scheme. Nevertheless, the establishment of this path may be directly hindered by unreliable cross-seam inter-satellite links. More importantly, the frequent transmission of model information on such a path can lead to a severe communication overhead. To mitigate the drawbacks of these two schemes, we propose employing the idea of a relay sum to realize inter-plane model aggregation. More specifically, based on the criterion of a minimum diameter to accelerate the spread of model updates, we first construct a spanning tree on the current inter-plane communication topology, considering only inter-satellite links that satisfy the minimum communication rate requirement. Then, in each iteration of model aggregation, through the usage of additional memory, the summation of the received model updates is stored and exchanged between neighbors, which effectively mitigates the information attenuation caused by iterative averaging. In this way, a faster convergence rate can be achieved while avoiding cross-seam links with low communication rates and consuming only the same communication overhead as the gossip aggregation scheme, as illustrated in Fig. 4.

4. Robust resource allocation

In this section, to quantitatively characterize the relationship between the limited network resources and computation service requirements in Space-CPN, we propose robust resource-allocation modeling and algorithmic approaches to address the challenges of dynamic network topologies and uncertain task demands.

4.1. Robust reinforcement learning for satellite microservice deployment

Space-CPN enables onboard inference capabilities for downstream remote sensing applications, such as land-use classification, disaster and environmental monitoring, and weather nowcasting. These computation tasks involve large data volumes, leading to a substantial computation burden that a single satellite cannot handle effectively. Moreover, different downstream inference tasks often involve redundant module deployment and computation, resulting in inefficiencies. Furthermore, Space-CPN encounters challenges due to the rapidly changing satellite topology and the highly heterogeneous computing environment, making onboard inference even more challenging. To address these coupled challenges, we propose a microservice-empowered onboard inference architecture for downstream tasks that divides the entire inference process into several low-coupled modules. These modules are assigned to corresponding satellite nodes for the sequential execution of specific inference tasks. This microservice-empowered onboard inference architecture offers advantages in terms of portability, scalability, and resilience in software development and maintenance. As illustrated in Fig. 5, the microservice-empowered onboard inference architecture includes several sequential steps, including microservice deployment, task requesting, task

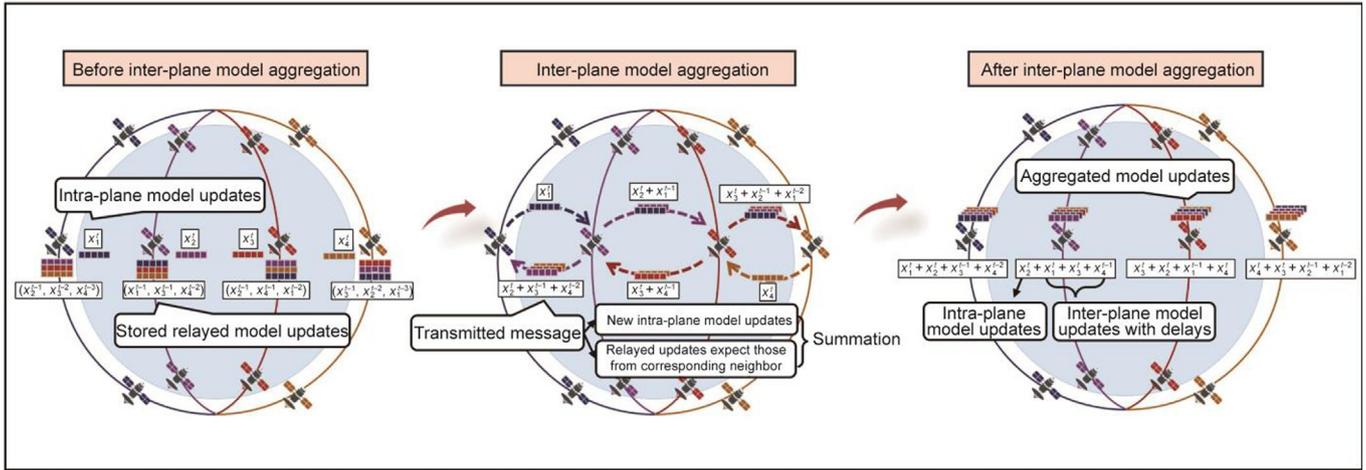


Fig. 4. Inter-plane aggregation method for decentralized satellite neuromorphic learning.

scheduling, data transmission, and result transmission. Thus, a statistically predetermined microservice deployment strategy is crucial in order for the entire process to proceed smoothly.

The essence of microservice deployment based on the statistical information of task requests is to decompose and distribute inference computation tasks across Space-CPN. The main challenge lies in the fact that, in addition to completing the preassigned tasks established at launch, Earth-observation satellites must be ready to handle emergent observation tasks such as forest fire monitoring and downburst nowcasting, based on ground instructions. These additional tasks often occur unexpectedly, with uncertainties in both occurrence time and frequency. To address these challenges, we propose the application of a robust optimization

approach to capture the uncertainty in task requests. The objective function aims to minimize the resource consumption of microservice deployment, including computation resources (e.g., CPU and GPU), memory, and power consumption. The optimization variables represent the microservice deployment strategy—that is, which microservice should be deployed on which satellite. The constraints of the problem include quality-of-service constraints such as latency, integer constraints for deployment strategies, and resource limits that cannot exceed the inherent resources of each satellite. To model the uncertainty of additional task requests, parameters such as the number of task requests from different regions are introduced into the robust optimization model and are bounded by an uncertainty set to increase robustness.

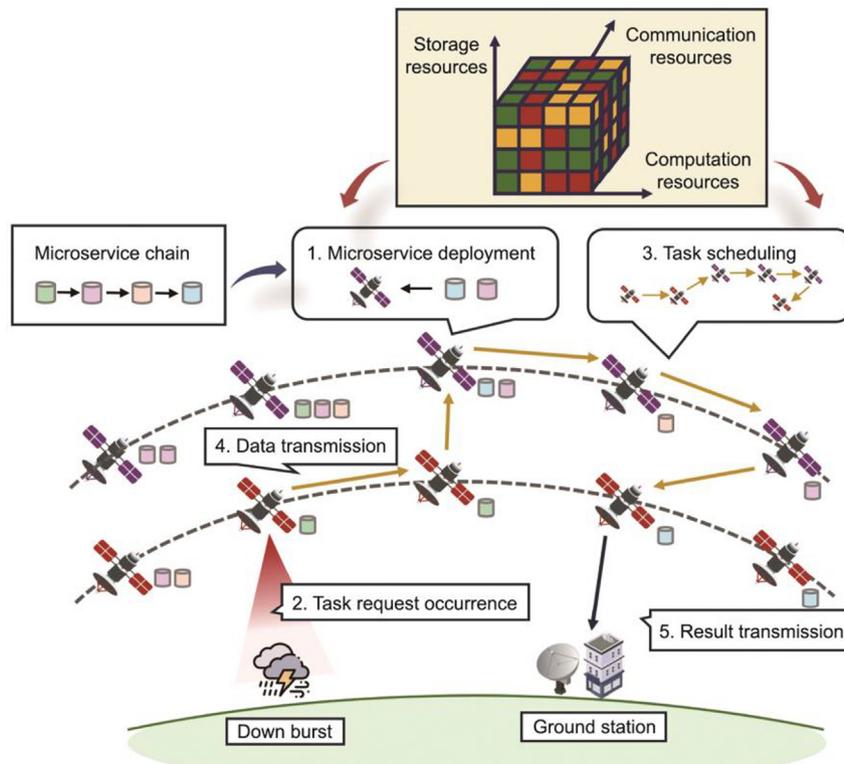


Fig. 5. The microservice-empowered onboard processing architecture in Space-CPN.

However, the uncertainty in quality-of-service constraints, which arises from the variability of the number of input data for the inference tasks, makes it difficult to solve this problem using traditional methods such as commercial solvers (e.g., Gurobi) alone. To address this issue, we reformulate the original problem as a partially observable Markov decision process (MDP) given its MDP characteristics, which makes it possible to obtain sub-optimal solutions using reinforcement learning.

We first formulate the box uncertainty set of the original problem as a latent state of the environment, representing unknown additional tasks that exist but cannot be directly observed by the agent. The reward received by the agent is affected by both the observable state and this latent component; to account for this, we introduce an additional reward impact term before the reward calculation. Based on this formulation, the problem can be naturally transformed into a partially observable MDP because the rest follows the standard MDP structure. To ensure the robustness of the solution, a perturbation agent is introduced into the traditional reinforcement learning framework to simulate environmental perturbations [22]. In each training round, the perturbation agent first applies a perturbation based on the current state; then, the robust agent makes decisions without direct access to the perturbation information, receiving rewards from both the environment and the effects of the perturbation. Through this adversarial training process, the robust agent ultimately learns a microservice deployment strategy that performs effectively under extreme scenarios.

4.2. Distributionally robust optimization for satellite task scheduling

Thanks to advanced remote sensing technologies, a large amount of high-resolution images are being captured by satellites, with great potential for supporting diverse downstream tasks. However, the conventional image-processing approach relies on downloading raw data to ground stations. Because of the long distance and poor channel conditions of satellite-to-ground communication links, this procedure inevitably incurs severe communication overheads, which are unacceptable for time-sensitive applications such as disaster prediction and intelligence reconnaissance. To address this issue, task scheduling in Space-CPN serves as an effective solution to enable fast and reliable data processing by coordinating mature on-orbit computing capabilities. The goal of satellite task scheduling is to seek an efficient routing strategy for the processing requests of images generated on remote sensing satellites. In each timestamp, a set of satellites first capture a series of remote sensing images of target areas. Owing to the limitations of a single satellite's computation resources, the original image-processing tasks at these satellites are decomposed, and corresponding image slices are then sent to other idle satellites through inter-satellite communication links for further processing. Eventually, the ground station collects the processing results from the satellites responsible for computation. This task scheduling scheme unleashes the computing power in the constellation. Nevertheless, due to the complex resource distribution and dynamic environments in Space-CPN, the routing scheme employed will have a significant impact on the performance of computation tasks, essentially resulting in a many-to-many knapsack matching problem for system optimization.

More specifically, since response time plays an important role in many space applications, the average image-processing delay naturally serves as the performance metric of this system and can then be decoupled into several parts. The first part is the image-transmission delay, which depends on the data volume and the bandwidth allocated from the source to destination node. The second part is the computational processing delay, which can be calculated based on the computational power allocated to the corresponding data. The third part is the satellite handover delay, which refers to the initialization delay caused by nodes matched for the first time. The last part is the results-feedback delay, which

corresponds to the time consumed by satellites sending results back to the ground station.

The fluctuations existing in the data volume of the captured remote sensing images are an important constraint. However, due to the satellites' rapid movement, the scenes the satellites cover change rapidly, resulting in unpredictable data volume. This eventually leads to uncertainty in satellite task scheduling, which refers to an uncertain amount of remote sensing data to be processed in the same task due to scene differences, requirements, weather, and time. Through stochastic optimization, this uncertainty can be modeled by a probability distribution, although such a distribution is difficult to estimate precisely in the dynamic environment of Space-CPN. Moreover, without the help of prior distributional knowledge, robust optimization directly minimizes the worst case cost, which may lead to a very conservative result when the worst case is very rare. To mitigate the drawbacks of these two schemes, we propose leveraging the idea of distributionally robust optimization. Based on a substantial amount of historical data for approximation, partial prior knowledge can be employed and an ambiguity set that encompasses the true distribution can be constructed to model the constraint of uncertain image volume. In this way, the randomness of stochastic optimization can be inherited, while the strong conservativeness of robust optimization is simultaneously mitigated. Thereafter, taking the bandwidth capacity, path, computational capacity, and handover constraints into consideration, a distributional robust many-to-many knapsack matching problem can be formulated [23]. With kernel density estimation employed for distribution fitting and the Wasserstein distance adopted for ambiguity set construction, the original problem can be transformed into a programmable explicit expression version and then decomposed into multiple subproblems, which can be efficiently solved using an alternating optimization approach.

5. Conclusions

Space-CPN is a promising architecture to enable seamless global connectivity and support various downstream computing tasks. Nevertheless, in order to establish Space-CPN, it is necessary to address fundamental challenges across communication strategies, computing and networking architectures, resource-allocation methods, and application scenarios in space networks. To tackle these issues, we first proposed the use of task-oriented communication solutions for information compression and transmission by extracting task-relevant information and folding the communication goals for computation. Second, we proposed the development of neuromorphic computing and networking architectures to support ultra-high-energy-efficient onboard data processing. Finally, we proposed robust optimization methods to quantitatively characterize the relationship between computation task requirements and network resources in dynamic and uncertain space environments. We hope this article will serve as a guideline for exploring opportunities to enable various promising onboard computing services in Space-CPN.

CRedit authorship contribution statement

Linling Kuang: Resources, Data curation, Conceptualization. **Yuanming Shi:** Writing – original draft, Methodology, Investigation. **Kai Liu:** Writing – review & editing, Validation, Formal analysis. **Chunxiao Jiang:** Visualization, Supervision, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

- [1] Abdelsadek MY, Chaudhry AU, Darwish T, Erdogan E, Karabulut-Kurt G, Madoery PG, et al. Future space networks: toward the next giant leap for humankind. *IEEE Trans Commun* 2023;71(2):949–1007.
- [2] Zhou D, Sheng M, Li J, Han Z. Aerospace integrated networks innovation for empowering 6G: a survey and future challenges. *IEEE Commun Surv Tutor* 2023;25(2):975–1019.
- [3] Tang X, Cao C, Wang Y, Zhang S, Liu Y, Li M, et al. Computing power network: the architecture of convergence of computing and networking towards 6G requirement. *China Commun* 2021;18(2):175–85.
- [4] Qin Z, Liang L, Wang Z, Jin S, Tao X, Tong W, et al. AI empowered wireless communications: from bits to semantics. *Proc IEEE* 2024;112(7):621–52.
- [5] Shi Y, Zhou Y, Wen D, Wu Y, Jiang C, Letaief KB. Task-oriented communications for 6G: vision, principles, and technologies. *IEEE Wirel Commun* 2023;30(3):78–85.
- [6] Lagunas E, Ortiz F, Eappen G, Daoud S, Martins WA, Querol J, et al. Performance evaluation of neuromorphic hardware for onboard satellite communication applications. *IEEE Commun Mag* 2024;31(6):78–84.
- [7] Deng S, Zhao H, Huang B, Zhang C, Chen F, Deng Y, et al. Cloud-native computing: a survey from the perspective of services. *Proc IEEE* 2024;112(1):12–46.
- [8] Wu SP, Liao W. Scalable contact graph modeling for low Earth orbit satellite mega constellations. *IEEE Netw* 2025;39(3):257–62.
- [9] Bui VP, Dinh TQ, Leyva-Mayorga I, Pandey SR, Lagunas E, Popovski P, et al. Semantic image encoding and communication for Earth observation with LEO satellites. *IEEE Trans Cogn Commun Netw* 2025;11(2):1210–24.
- [10] Goldfeld Z, Polyanskiy Y. The information bottleneck problem and its applications in machine learning. *IEEE J Sel Areas Inf Theory* 2020;1(1):19–38.
- [11] Xie S, Ma S, Ding M, Shi Y, Tang M, Wu Y. Robust information bottleneck for task-oriented communication with digital modulation. *IEEE J Sel Areas Commun* 2023;41(8):2577–91.
- [12] Pensia A, Jog V, Loh PL. Extracting robust and accurate features via a robust information bottleneck. *IEEE J Sel Areas Inf Theory* 2020;1(1):131–44.
- [13] Shi Y, Zeng L, Zhu J, Zhou Y, Jiang C, Letaief KB. Satellite federated edge learning: architecture design and convergence analysis. *IEEE Trans Wirel Commun* 2024;23(10):15212–29.
- [14] Yang P, Wen D, Zeng Q, Zhou Y, Wang T, Cai H, et al. Over-the-air computation empowered vertically split inference. *IEEE Trans Wirel Commun* 2024;23(12):19634–48.
- [15] Wang Z, Zhao Y, Zhou Y, Shi Y, Jiang C, Letaief KB. Over-the-air computation for 6G: foundations, technologies, and applications. *IEEE Internet Things J* 2024;11(14):24634–58.
- [16] Letaief KB, Shi Y, Lu J, Lu J. Edge artificial intelligence for 6G: vision, enabling technologies, and applications. *IEEE J Sel Areas Commun* 2022;40(1):5–36.
- [17] Nazer B, Gastpar M. Compute-and-forward: harnessing interference through structured codes. *IEEE Trans Inf Theory* 2011;57(10):6463–86.
- [18] Zhu J, Shi Y, Zhou Y, Jiang C, Chen W, Letaief KB. Over-the-air federated learning and optimization. *IEEE Internet Things J* 2024;11(10):16996–7020.
- [19] Shrestha A, Fang H, Mei Z, Rider DP, Wu Q, Qiu Q. A survey on neuromorphic computing: models and hardware. *IEEE Circuits Syst Mag* 2022;22(2):6–35.
- [20] Eshraghian JK, Ward M, Neftci EO, Wang X, Lenz G, Dwivedi G, et al. Training spiking neural networks using lessons from deep learning. *Proc IEEE* 2023;111(9):1016–54.
- [21] Tao M, Zhou Y, Shi Y, Lu J, Cui S, Lu J, et al. Federated edge learning for 6G: foundations, methodologies, and applications. *Proc IEEE*. In press.
- [22] Pinto L, Davidson J, Sukthankar R, Gupta A. Robust adversarial reinforcement learning. In: *Proceedings of the 34th International Conference on Machine Learning*; 2017 Aug 6–11; Sydney, NSW, Australia; 2017.
- [23] Sun J, Chen X, Jiang C, Guo S. Distributionally robust optimization of on-orbit resource scheduling for remote sensing in space-air-ground integrated 6G networks. *IEEE J Sel Areas Commun* 2025;43(1):382–95.