SCIENCE CHINA
Information Sciences

• REVIEW •

September 2025, Vol. 68, Iss. 9, 190302:1–190302:31
https://doi.org/10.1007/s11432-024-4450-6

Special Topic: Cohesive Clustered Satellites System for 5GA and 6G Networks

A survey on communication and computing resources allocation and management for cohesive clustered satellites systems

Chen ZHANG[1,2], Xin WAN[1,2], Wanjing LI[1,2], Haojun LIU[1,2],
Dongming BIAN[1*] & Gengxin ZHANG[1,2]

[1] *College of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China*
[2] *National Engineering Research Center for Communication and Network Technology, Nanjing University of Posts and Telecommunications, Nanjing 210003, China*

Received 29 November 2024/Revised 14 March 2025/Accepted 18 May 2025/Published online 15 August 2025

Abstract With the rapid development of large-scale low Earth orbit (LEO) satellite constellations, cohesive clustered satellite (CCS) systems are considered critical spatial information infrastructures for augmenting space resource efficiency by aggregating the service capabilities of multiple satellite platforms to form an ultracohesive virtual satellite. This paper aims to fill the gaps identified in the literature by providing an extensive survey on communication and computing resources management, integrating insights from academia and the industry of CCS systems. It discusses the evolution and components of CCS systems and then introduces a native artificial intelligence architecture along with a hierarchical structure for resource management. Furthermore, it provides the emerging technologies and applications brought by CCS. On this basis, the modeling and metrics of resource management are established. Then, an exhaustive review of resource management methods encompassing both traditional and artificial intelligence algorithms is presented. Among these methods, knowledge-driven resource scheduling strategies are highlighted. Finally, some future directions of CCS systems are outlined.

Keywords cohesive clustered satellites, LEO, communication and computing resource, resource management, artificial intelligence

Citation Zhang C, Wan X, Li W J, et al. A survey on communication and computing resources allocation and management for cohesive clustered satellites systems. Sci China Inf Sci, 2025, 68(9): 190302, https://doi.org/10.1007/s11432-024-4450-6

1 Introduction

1.1 Background and motivation

To more effectively address the demands of 5G global ubiquitous interconnection [1,2], the progression of nonterrestrial networks (NTNs) has evolved into an inexorable trend [3]. Through the establishment of satellite networks collaborating with terrestrial networks, the restricted coverage of terrestrial networks in remote areas, deserts, and oceans can be surmounted, thereby achieving global seamless services [4].

The traditional transparent forwarding mode [5] used in satellites offers the advantages of straightforward deployment and low complexity. However, it does not conduct any error detection or correction on data packets, thereby restricting its application scenarios and lacking flexibility. With advancements in hardware and payload capabilities, satellites have progressively adopted the onboard regenerative forwarding mode, which encompasses functions such as signal regeneration, error detection and correction, and data packet rerouting. Given the advantages of onboard processing, 5G NTN proposes the use of complete or partial base station functions on satellites to realize "base station onboard" [6]. Consequently, the 3rd Generation Partnership Project puts forward two architectures for onboard processing: store and forwarding (S&F) and user equipment-satellite-user equipment (UE-SAT-UE) [7].

Meanwhile, the design concept of traditional geostationary Earth orbit (GEO) large satellites dictates that a single payload possesses strong capabilities and multiple functions. Nevertheless, such satellites are

[*] Corresponding author (email: bian_dm@163.com)

typically costly and lack flexibility as they are unable to adapt to changes in service and task requirements [4, 8]. In recent years, because of the advantages of miniaturization, integration, and functionalization of low Earth orbit (LEO) satellites, as well as their low manufacturing and launch costs [9], they have become a prominent focus within the industry [10].

The diversified application requirements and intelligent service paradigms under 5G NTN and the anticipated 6G ecosystem are rapidly proliferating. Nevertheless, persistent institutional fragmentation and the scarcity of heterogeneous network resources impose critical constraints on existing LEO satellite systems, hindering their capacity to effectively support these intensifying demands. For instance, although a single low-orbit satellite is part of a network, there is a notable absence of direct collaboration among satellites within the LEO constellation. Consequently, conducting intricate, intelligent, and diversified tasks such as joint data transmission, cooperative computing and observation is extremely challenging. Moreover, each satellite's resources and payload functions operate in relative isolation. Thus, they cannot provide information services that are globally open, integrated, and unified.

Therefore, with the evolution of the service scenarios of 5G NTN/6G and the development of low-orbit satellite payload technology, cohesive clustered satellite (CCS) systems integrating communication, computing, sensing, and other functions become both feasible and promising. CCS systems are poised to aggregate the service capabilities of conventional payload-integrated single-satellite platforms to form an ultracohesive virtual satellite. Generally speaking, CCS systems refer to space information systems constituted by a plurality of satellites that are distributed within one or more orbits in a specific collaborative manner. Their principal feature and design concept center around "multisatellite coordination". CCS systems can be regarded as a novel form of the LEO system and represent an inevitable future development trend. Through the aggregation, reconstruction, and orchestration of heterogeneous satellite resources, including communication, computation, and storage [11], several satellites are virtualized into a single large one. This enables the joint accomplishment of tasks such as cooperative communication, earth observation, remote sensing, and scientific experiments, thereby circumventing the limitations associated with the restricted payload capacity of an individual satellite [12].

The system composition model of a CCS is depicted in Figure 1. It clusters satellites to integrate the service capabilities of traditional single-satellite platforms with payloads to form an ultracohesive virtual satellite. This virtual satellite emerges as a crucial spatial information infrastructure that enhances the efficiency of space resources [11]. At present, satellite networks are evolving from signal coverage to capacity coverage and further to service coverage [13]. Therefore, CCS systems represent a significant advancement in the satellite field, necessitating a comprehensive review and summarization of their research issues and challenges to fully realize their potential.

1.2 Overview of related surveys and our contributions

For readability, Table 1 [13–19] summarizes the relevant reviews of satellite networks in recent years. For instance, Wang et al. [13] comprehensively reviewed the integration of satellite and terrestrial networks, mainly aiming at the standardization work, application scenarios, and performance evaluations of satellite-terrestrial networks. Thereafter, Li et al. [14] introduced edge computing technology, proposed a satellite-terrestrial integrated edge computing network architecture, and analyzed various computing offload mechanisms. For multilayer NTN, Chen et al. [15] focused on the network architecture and resource allocation strategy for communication and computing. However, review studies concerning CCS systems are scarce. Marrero et al. [16] primarily studied network topology and synchronization technology within CCS. However, their study did not involve resource management in CCS systems. A comprehensive and detailed comparison of these reviews is presented in Table 1.

It can be inferred that the current reviews associated with satellite networks predominantly concentrate on NTN (LEO) or space-air-ground integrated networks (SAGINs), with relatively scant attention paid to CCS systems. Meanwhile, resource management primarily centers around communication, and there is a dearth of reviews regarding multidimensional resource scheduling for converged communication and computing. Nevertheless, in contrast to the generalized LEO system, the principal distinctions of the factors affecting resource management between these systems and CCS are as follows.

(1) **Dynamic system configuration.** In CCS systems, the high-speed motion of LEO makes the Doppler frequency offset and time delay between satellite-ground links large, making synchronization implementation and system performance difficult. More importantly, the topology, scale, and interconnection relationships of satellites in CCS undergo highly dynamic changes. Therefore, this highly

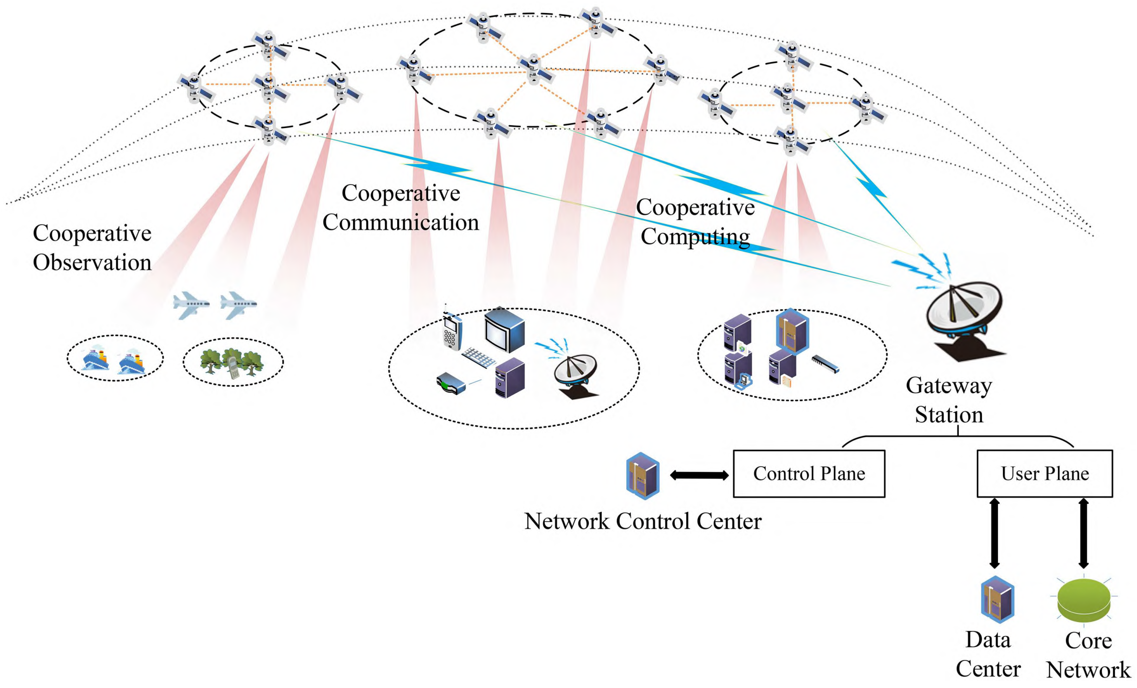

Figure 1 (Color online) Architecture of a CCS system.

Table 1 Existing surveys related to NTNs. DSS: distributed satellite system; SAGSIN: space-air-ground-sea integrated network; JCC-SAGIN: joint communication and computing-embedded SAGIN.

Reference	Year	Network	Management structure	Comm.& Computing	Emerging tech	Traditional algorithm	Data-driven algorithm	Knowledge-driven algorithm	Contribution
[13]	2020	Satellite-terrestrial network	×	×	√	×	×	×	Focusing on the architecture of satellite–terrestrial network, providing standards and key applications
[16]	2022	DSS	×	×	×	×	×	×	Summarizing synchronization techniques for DSS
[17]	2023	Satellite network	×	√	√	×	×	×	Presenting vision and challenges for satellite computing
[14]	2023	Satellite-terrestrial network	√	√	×	√	×	×	Analyzing computing offloading mechanisms and application scenarios
[18]	2023	SAGSIN	×	×	×	√	√	×	Focusing on SAGSIN from a resource optimization perspective in 6G networks.
[19]	2024	SAGIN	×	×	×	√	√	×	Focusing on resource allocation methods within SAGIN
[15]	2024	SAGIN	×	√	√	√	√	×	Focusing on enabling technologies, applications, and resource management modeling and optimization methods in JCC-SAGIN
Ours	2025	CCS	√	√	√	√	√	√	Providing a comprehensive analysis and survey for managing heterogeneous resources in CCS, including evolution, architecture, traditional, and data- and knowledge-driven AI algorithms

dynamic network topology poses higher challenges to the complexity and performance requirements of CCS resource management and scheduling methods.

(2) **Multisatellite coordination.** An essential characteristic of CCS systems is their accomplishment of diverse tasks via cooperative control and regulation among individual satellites. A crucial consideration in resource management and scheduling is how to combine the service capabilities of traditional single-satellite platforms with integrated payloads to create an ultracohesive virtual satellite. This virtual satellite can then fulfill the requirements of various application scenarios, such as cooperative communication, cooperative computing, and cooperative beam hopping.

(3) **Heterogeneous resources and diverse tasks.** Within a CCS system, heterogeneous communication and computation resources are both diverse and widely distributed. Consequently, the system must manage the substantial volume of data that is generated, exchanged, and processed among and within heterogeneous network layers. Additionally, the multiplicity of communication services and computing tasks, the nonuniformity of spatiotemporal distribution, and the variability of quality of service (QoS)

requirements have become increasingly prominent. Thus, the resource management architecture of CCS must be capable of enabling resource sharing and dynamic scheduling across and within heterogeneous layers, thereby fulfilling the requirements of different communication and computing services and tasks.

Considering the aforementioned issues, our objective is to fill the gaps identified in the literature by providing an extensive survey on heterogeneous resource management within CCS, contributing to both theoretical research and practical system design advancements. This study provides an innovative native artificial intelligence (AI) CCS system architecture and hierarchical heterogeneous resource scheduling structure and deeply reviews resource management models and methods in CCS. To the best of our knowledge, this paper represents the first exhaustive review to detail knowledge-driven intelligent resource scheduling strategies in CCS. The main contributions of this paper are outlined as follows.

(1) This study sets itself apart from existing surveys by highlighting several critical distinctions. Unlike other surveys on NTN or SAGIN, this research focuses on communication and computing resource allocation and management in CCS. It emphasizes the evolution of CCS by considering its communication payload and onboard computing hardware and discusses resulting emerging technologies, such as cooperative computing, cooperative transmission, and multisatellite cooperative beam-hopping. It also addresses new resource management challenges inherent in CCS to support these emerging technologies. Then, it is primarily devoted to an in-depth review and analysis of key capabilities in cohesive heterogeneous resource management, including resource scheduling orchestration, heterogeneous resource characterization and modeling, and resource scheduling algorithms.

(2) Through the analysis of the enabling and emerging technologies within CCS systems, an innovative native AI CCS system architecture is constructed, accounting for both the approach of AI for CCS and CCS for AI. This transforms resource management in CCS from interactive control signaling to information exchange and knowledge update among network elements. Correspondingly, a hierarchical heterogeneous resource layout and scheduling structure for CCS is proposed. Considering the constraints of computing capabilities and management ranges of different layers, resource scheduling AI models of different scales are deployed layer by layer to intelligently manage and schedule multidimensional resources.

(3) A deep literature review on the resource management of CCS is conducted, emphasizing modeling and optimization strategies. In addition to traditional optimization techniques, intelligent decision-making methods are analyzed. On the basis of a survey of data-driven intelligent resource scheduling methods and per the current cutting-edge development of artificial intelligence technology, this paper comprehensively explores and reviews knowledge-driven resource scheduling strategies for CCS. To the best of our knowledge, it is the first to provide the principles and advantages of knowledge-driven approaches and expound on the feasibility of their application in CCS systems.

Figure 2 illustrates the structural organization of this survey. The remainder of this article is structured as follows. Section 2 summarizes the architecture and evolution of CCS, with a particular emphasis on an innovative native artificial intelligence CCS system architecture and a hierarchical heterogeneous resource scheduling framework. Section 3 explores the emerging technologies facilitated by CCS, which also constitute the application scenarios for resource management. Sections 4 and 5 deliberate on resource management modeling and optimization approaches, encompassing both traditional and AI-based algorithms. Section 6 concludes our work and contemplates some prospective research directions. Finally, all abbreviations used in this article have been compiled in Appendix A.

2 Evolution and architecture

This section describes the evolution of the CCS system, along with the development of communication payloads and computing hardware onboard. Subsequently, the system components of CCS are elucidated. After introducing software-defined networking (SDN) and network function virtualization (NFV) technologies, which are the key technologies enabling CCS, it focuses on discussing system architecture and resource management design.

2.1 Evolution of CCS

Evidently, the satellite payload technology and system architecture of CCS are in a continuous state of evolution. The payload has advanced from transparent forwarding to regenerative forwarding, and its function changed from single task to integration of communication, navigation, remote sensing and space

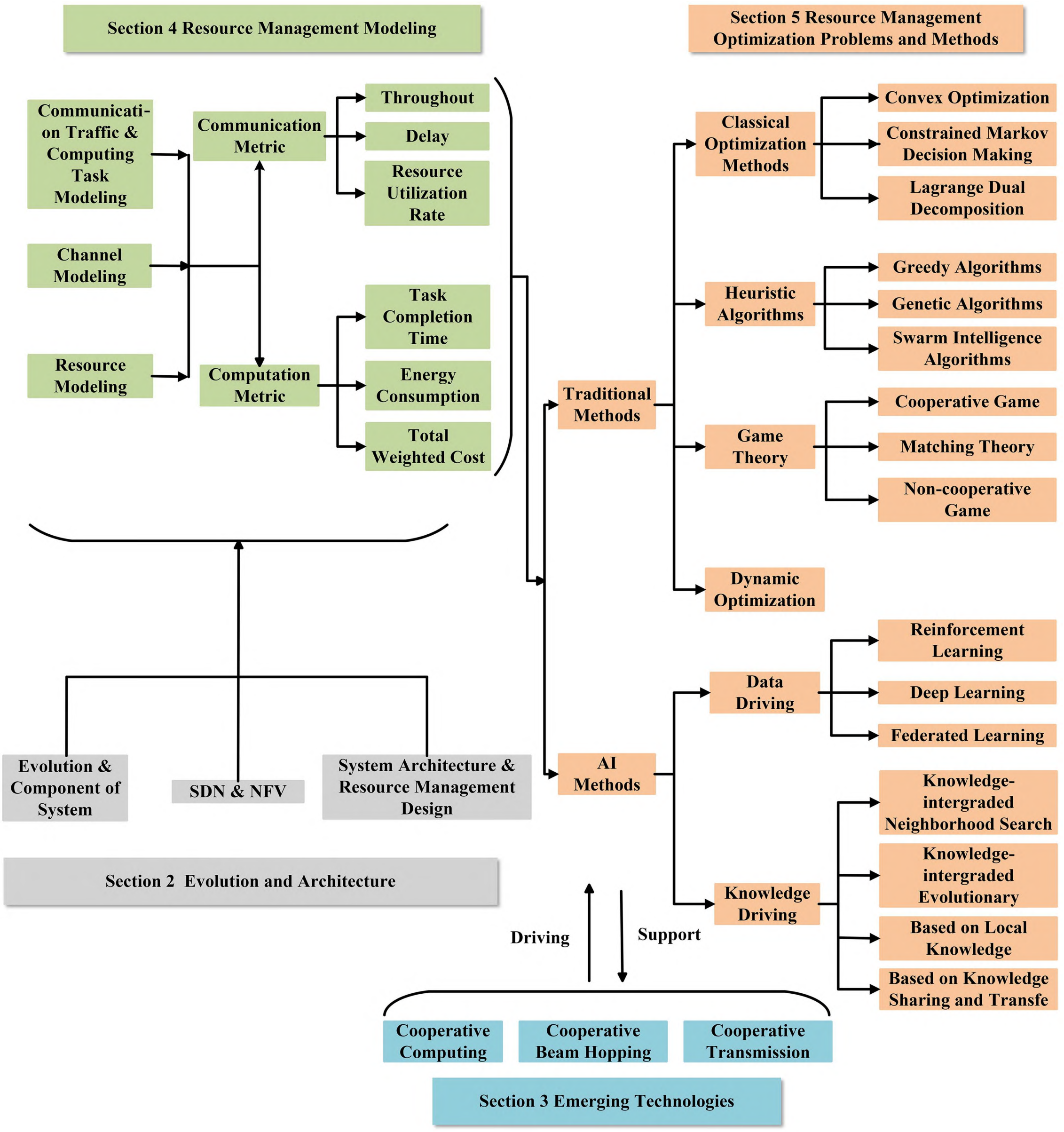
Section 4 Resource Management Modeling
Section 5 Resource Management Optimization Problems and Methods
Communication Traffic & Computing Task Modeling
Channel Modeling
Resource Modeling
Communication Metric
Computation Metric
Throughput
Delay
Resource Utilization Rate
Task Completion Time
Energy Consumption
Total Weighted Cost
Traditional Methods
Classical Optimization Methods
Convex Optimization
Constrained Markov Decision Making
Lagrange Dual Decomposition
Heuristic Algorithms
Greedy Algorithms
Genetic Algorithms
Swarm Intelligence Algorithms
Game Theory
Cooperative Game
Matching Theory
Non-cooperative Game
Dynamic Optimization
AI Methods
Data Driving
Reinforcement Learning
Deep Learning
Federated Learning
Knowledge Driving
Knowledge-intergraded Neighborhood Search
Knowledge-intergraded Evolutionary
Based on Local Knowledge
Based on Knowledge Sharing and Transfe
Evolution & Component of System
SDN & NFV
System Architecture & Resource Management Design
Section 2 Evolution and Architecture
Driving
Support
Cooperative Computing
Cooperative Beam Hopping
Cooperative Transmission
Section 3 Emerging Technologies

Figure 2 (Color online) Structure of this paper.

computing [20]. With the decreasing manufacturing and launch costs of LEO, along with the maturity of inter-satellite link and space networking technologies, the CCS system has progressively transformed from satellite formations and clusters to satellite constellations [16]. Moreover, it has been integrated with terrestrial networks to supply seamless, all-weather information services [21].

More importantly, the cooperation manner between satellites in the CCS system is also evolving with a diverse trend. In terms of cooperative communication, multiple satellites are utilized for multi-satellite and multi-beam transmission, or established distributed MIMO to enhance the communication diversity gain [22]; by employing a phase array antenna, multi-satellite cooperative beamforming can also be performed to suppress co-frequency interference, thereby improving the system capacity. In terms of cooperative observation, CCS cooperatively performs observation tasks and transmits its data to the ground station via scheduled beams to achieve stereo observation [11]. In terms of collaborative computing, by employing promising technologies such as edge computing, model pruning and knowledge distillation, the computing tasks can be offloaded between multiple satellites, aiming to improve service quality, as well as circumvent the issues of limited local computing, storage, and power consumption resources [15].

Consequently, CCS systems have emerged as a focal point within both academia and the industry in

Table 2 CCS systems launched and planned in recent years.

Country	Satellite system	Satellite function and orbit	Number of satellites
China	Ling que	Remote sensing SSO+LEO	132 in planning
	Hongtu-1 (01 Unit)	Remote sensing LEO	4 satellites formation
	Tian suan	Communication, computation LEO	6 in planning
	Tian yi	Remote sensing LEO	120 in planning
	Sail Space (G60)	Communication LEO	648 of 1st generation
USA	Dove	Remote sensing SSO+LEO	About 100+
	Kuiper	Communication LEO	3000+ in planning
	Starlink/StarSheild	Communication/communication, remote sensing and computing integration LEO	4408 of 1st gen, 42000+ total planned
UK	OneWeb	Communication LEO	648 total planned

recent years. Table 2 presents a summary of the innovative CCS systems that have either been launched or are in the planning stage. From Table 2, a significant evolution in satellite functionality is observed, transitioning from transparent forwarding to onboard processing capabilities. And the payload type displays a wide variety, encompassing communication, remote sensing, computing, as well as integrated payloads.

2.2 Evolution of payload and hardware on board

2.2.1 *Flexible payload of communication*

The flexible payload is set to become the standard-configuration of the new generation communication satellites [3], eventually replacing the payload of transparent transponder to fulfill the multi-mission requirements and adapt to the changes within the CCS system [23]. It mainly includes the following key technologies.

Regenerative processing. This methodology operates on the digital baseband data obtained after waveform digitization, demodulation and decoding [24], which has been discussed for many decades. However, with the demand for onboard base stations and lightweight core-net onboard represented by Starlink and O3b, regeneration processing has been given new connotations and challenges. In addition to signal processing, high-level protocol analysis, routing and switching are also carried out. Moreover, reprogrammable functionality is also supported to avoid suffering from obsolescence of technology [25].

Digital transparent processor (DTP). It amalgamates the advantages of transparent forwarding and regeneration processing. By leveraging digital channelization technology, it actualizes the functions of channel demultiplexing, switching, and multiplexing within the digital domain. It is capable of accomplishing the routing and switching procedures of any sub-channel signal from any input beam to any output beam, and can even achieve cross-band switching through resource pooling [26], it attains the flexible allocation of beam, frequency, and power resources in accordance with application requirements [27]. More importantly, DTP based processing results in payload designs agnostic to air-interface evolutions [24]. DTP has been employed in INMARSAT-4, SES 12, SES 17 satellites.

Beam hopping. In conventional multi-beam high throughput satellite, system can only allocate available resources in the limit of a single beam. As a sequence, many satellite manufacturers and operators such as Spaceway3 series satellites, European quantum satellite, have begun to test and verify flexible beam configuration in recent years [28, 29]. The basic idea of beam hopping is that at a certain moment, not all beams work, but only a part of the beams are activated on demand. That is, where and when there is demand, there will be beam coverage and illuminating to service [30]. Different from earlier beam switching [31], the current beam-hopping technology usually employs phased array antennas to accomplish rapid beam forming and direction control. The principal advantage is that the direction, size, shape, and EIRP of each beam can be adjusted as requirements [32, 33].

2.2.2 *Onboard computational capacity*

Space-based computing involves the processing of data and signals on platforms situated in space, whereas onboard computational capacity denotes the velocity and efficacy of data processing during the execution of computational tasks by space-based computing systems. Recently, the next-generation onboard computer, which is based on embedded cluster computing, has been designed with an architecture that

Table 3 Performance metrics of high-performance aerospace computing processors.

CPU model	Country/region	Clock frequency	Number of IP cores	Computational ability	Framework
TSC695F	Europe	25 MHz	1	20 DMIPS	SPARC
RAD750	USA	200 MHz	1	400 DMIPS	PowerPC
GR712RC	Europe	100 MHz	2	200 DMIPS	SPARC
DAHLIA	Europe	1.6 GHz	4	4000 DMIPS	ARM
HPSC	USA	800 MHz	8	7360 DMIPS	ARM
Loongson 1F	China	266 MHz	2	200 MIPS	MIPS
BM3883	China	1 GHz	8	16 GIPS	SPARC
Yulong 810	China	1 GHz	8	12 TOPS	ARM

consists of a master processing node, a monitoring and detection node, and a slave processing node. This architecture promotes distributed parallel computing, enhancing the overall computing efficiency and performance of the system. Applications of space-based computing encompass satellite remote sensing anomaly detection and image recognition [34].

Concurrently, to support space-based computing, the necessity for high-performance space computing (HPSC) processors is paramount. In the aerospace domain, the reliability of chips outweighs their performance. Only radiation-relistened chips can ensure the normal operation. With the advancement of semiconductor technology, the mainstream onboard computers primarily use processors based on architectures such as PowerPC, SPARC, ARM, and MIPS [35]. Table 3 showcases the recent developments in onboard computing products [36].

On the other hand, in line with the development of the diversification and integration of the missions within the CCS system, the electronic system solution for space-based computing demands the utilization of a processing architecture that combines FPGA + DSP + CPU. This combination enables the capacity for functional reconstruction. Eventually, considering the requirement for complex parallel computing of AI on board in the near future, the research and implementation of GPU on board become even more crucial on the space computing platform.

2.3 System components of CCS

Space segment. The space segment of CCS primarily consists of satellites, which can be categorized into three types: constellations, clusters and formations [20]. These satellites are usually distributed on one or more orbits to jointly realize designated space missions, including earth observation, remote sensing imaging, early warning, collaborative computing, and communication [11]. In addition, there are two emerging approaches, fractionated and federated [12], which have not yet been implemented in practical. Moreover, considering the communication links between the nodes of CCS, the space segment can be classified as Ring, Star, Mesh and Tree topologies.

Ground segment. As depicted in Figure 1, the ground segment is comprised predominantly of user equipment (UE), gateway station (GW), operation and maintenance subsystems, tracking, telemetry and command (TT&C) subsystem. The types of UE are highly diverse, with requirements for both communication and computing services. The main ones include: handheld terminal devices (such as satellite phones and satellite-direct-to-cell mobile phones), Internet of Things (IoT) devices (encompassing communication services like data collection, as well as computing services such as edge processing), ground stations (large fixed satellite communication stations), and airborne and shipborne satellite communication equipment. GWs serve as the intermediaries bridging the terrestrial and satellite networks. They receive remote sensing data, communication data, and computational data through satellite feed links. The remote sensing and computational data can be transmitted to the system data center for additional processing, during which the raw data is converted into remote sensing and computational results as the user's requests. Meanwhile, the communication data is relayed to the core terrestrial network to enable data exchange. Furthermore, the network control center (NCC) within operation and maintenance subsystems also constitutes a vital part of the CCS system, responsible for management and evaluation of the entire network, users and services, as well as resource scheduling and management.

2.4 Resource virtualization via SDN and NFV

Traditional network management and control approaches prove inadequate when it comes to satellite networks, which feature huge, distributed and dynamic nodes. By introducing SDN into satellite networks, the data plane and control plane are separated, allowing the centralized controller to perform complex operations, including routing decision, resource management [37]. Furthermore, the application of NFV technology allows the decoupling of physical hardware, which facilitates the pooling of virtualized resources. This provides the SDN controller with a global view of network resources, thereby enhancing resource flexibility and utilization [38]. Within the CCS system's heterogeneous networks, barriers exist in terms of physical hardware and network protocols. Fortunately, SDN and NFV technologies are capable of fulfilling the interaction needs among these heterogeneous networks [39], thereby establishing a basis for the unified management and efficient scheduling of resources.

SDN controllers can be deployed in a hybrid manner across ground gateway stations and satellites, with each being assigned the responsibility of supervising network operations at different levels. In the context of the CCS system, the strategy placement of SDN controllers across multiple satellites facilitates adaptable distributed control strategies and enhances network resilience [40]. Typically, a hybrid deployment configuration consists of two models [40, 41]: a ground-based primary controller with satellite support and a satellite-based primary controller with ground support. The former arrangement reduces control latency while maintaining global oversight, whereas the latter shifts control to the satellite-based controller in certain scenarios, such as remote locations or emergency situations, with the ground-based controller acting as a secondary node, offering essential resource information and support.

NFV utilizes virtualization technology to separate network functions from hardware equipment. Combining with network slicing, NFV transforms physical network resources into logical entities, enabling the formation of separate virtual networks customized for different services. This promotes flexible and efficient network management and resource allocation [1, 10, 38, 39]. The application of NFV to mobile edge computing (MEC) allows for the virtualization and dynamic management of edge computing resources [42, 43], which is employed for satellite-ground collaborative network [40], enhancing resource utilization efficiency and the quality of network services.

2.5 System architecture and resource management design

By aggregating, reconstructing, and orchestrating heterogeneous satellite resources, including computation and communication, the CCS system enables resource sharing and provides information-centric and intelligent on-demand services [11]. Due to the characteristics of complex system, dynamic topology, heterogeneous resources, diversified and spatiotemporal variation service requirements within the CCS system, its resource scheduling problem presents challenges when it comes to being modeled and resolved using traditional approaches. Therefore, an artificial intelligence method must be employed for multi-dimensional resource scheduling.

This paper proposes an innovative native AI CCS system architecture. This shifts the resource management of CCS from interactive control signaling to information exchange and knowledge update among network elements. Based on this, a hierarchical heterogeneous resource layout and scheduling structure for CCS is constructed. In accordance with the constraints of computing capabilities and management scopes of different layers, AI models for resource scheduling of different scales are deployed layer by layer to manage and schedule multi-dimensional resources in an intelligent manner. Therefore, by considering both the approach of AI for CCS and CCS for AI, the network and AI support each other and enhance mutual benefit, as follows.

2.5.1 *Native AI-based CCS system architecture*

As mentioned above, AI can effectively assist the system in decision-making and scheduling for complex heterogeneous multi-dimensional resources. However, traditional "superimposed and externa" AI not only brings massive transmission and computing overhead problems but also fails to fully leverage the advantages of resource sharing and distributed collaboration in the CCS system.

Therefore, the CCS system architecture with native intelligence is proposed. Based on the computing and storage capabilities of different equipment of the network control center, regional autonomous control sub-centers and edge satellite payloads, the resource scheduling intelligent models are designed with different scales. As shown in Figure 3, due to the powerful computing and processing capabilities of the network control center, a global resource scheduling large model can be deployed. On this basis, by utilizing deep separable convolution, model pruning, model quantization and knowledge distillation technologies, lightweight neural network models are designed to maintain model performance while reducing model parameters and increasing computational speed, which can be deployed in regional autonomous

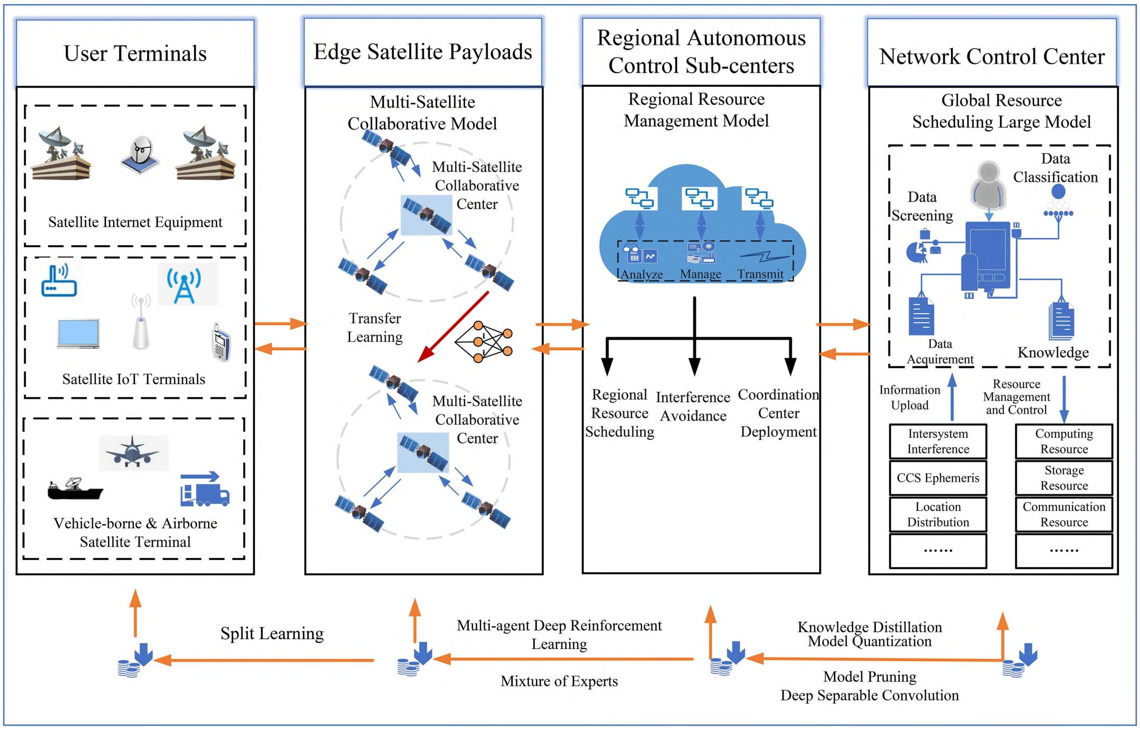
User Terminals
Satellite Internet Equipment
Satellite IoT Terminals
Vehicle-borne & Airborne Satellite Terminal
Edge Satellite Payloads
Multi-Satellite Collaborative Model
Multi-Satellite Collaborative Center
Transfer Learning
Multi-Satellite Collaborative Center
Regional Autonomous Control Sub-centers
Regional Resource Management Model
Analyze
Manage
Transmit
Regional Resource Scheduling
Interference Avoidance
Coordination Center Deployment
Network Control Center
Global Resource Scheduling Large Model
Data Classification
Data Screening
Data Acquirement
Knowledge
Information Upload
Resource Management and Control
Intersystem Interference
CCS Ephemeris
Location Distribution
......
Computing Resource
Storage Resource
Communication Resource
......
Split Learning
Multi-agent Deep Reinforcement Learning
Mixture of Experts
Knowledge Distillation
Model Quantization
Model Pruning
Deep Separable Convolution

Figure 3 (Color online) Native artificial intelligence-based CCS system architecture.

control sub-centers. Further, in response to the complex scheduling tasks and diverse heterogeneous resources on board, different sub-models are designed by using multi-agent deep reinforcement learning and the mixture of experts (MoE) approach, forming an intelligent agent for satellite resource scheduling to meet users' diverse QoS requirements. MoE consists of multiple independent sub-models (experts). According to different tasks, only one or part of the experts are activated during the inference. The gating network calculates the weight of each expert based on the characteristics of the input data, then assigns the input data to the corresponding expert. Finally, the outputs of all activated experts are weighted and summed to generate the result. Compared with dense models, its pre-training speed is faster. Compared with models with the same number of parameters, its inference speed is faster. Thus, the utilization rate of satellite resources is optimized and complex resource scheduling problems are handled. By leveraging the short-term strong correlation of user services and resource requirements, as well as the long-term similarity of channels and electromagnetic environments, transfer learning and knowledge sharing are carried out among multiple satellites to address the issues of limited onboard processing capacity and high inter-satellite collaboration costs. This approach aims to reduce computational and collaborative overhead, enhance edge service performance and solve the problem of real-time resource allocation and scheduling on board. Specifically, driven by the different QoS requirements of computing and communication tasks, the collaboration mechanism of the CCS system is utilized to facilitate data exchange and intelligent model parameter sharing between satellites through knowledge sharing. Satellites use the intelligent model parameters obtained from knowledge sharing as the data foundation of transfer learning and fine-adjust the pre-trained model according to the QoS requirements of the current user services, realizing the real-time allocation of heterogeneous resources on board. Regarding heavy computing tasks, by employing split learning, cooperative training and model segmentation are implemented between the terminals and edge satellites with the aim of reducing the interaction overhead. On the other hand, the distributed collaboration advantages of the CCS system are fully utilized, combining with the cloud-edge-terminal network to support channel parameters aggregation, collaborative data distribution, and model parameters passing, adapting to the high dynamics in the CCS system.

In summary, from both AI for CCS and CCS for AI perspectives, a native intelligent CCS system architecture is established, thereby revolutionizing CCS resource management and control from signalling interaction to information interaction and knowledge updates among network elements. When significant changes occur in key system parameters such as constellation configuration, satellite tasks, and spectrum

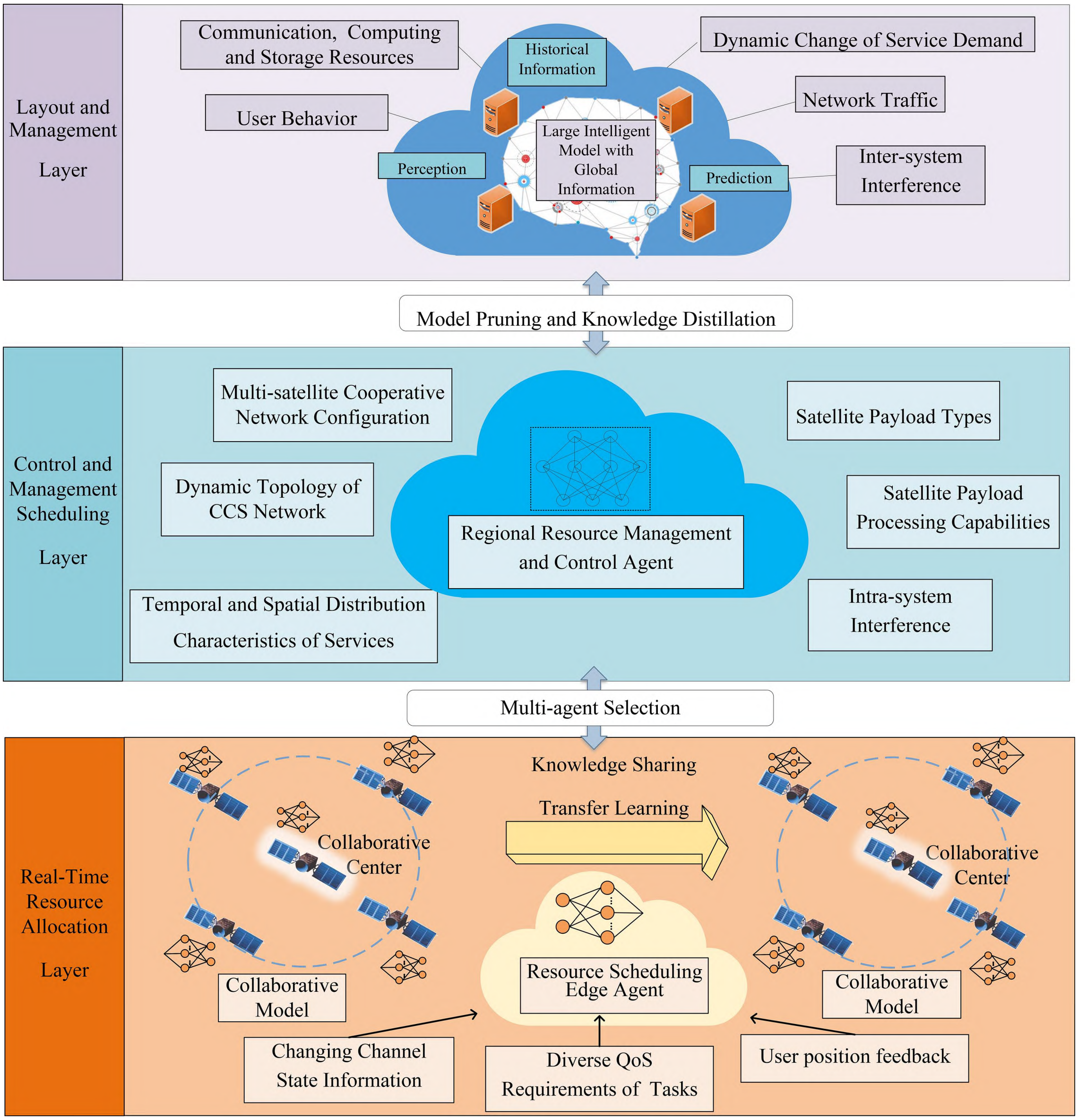
Layout and Management Layer
Communication, Computing and Storage Resources
Historical Information
Dynamic Change of Service Demand
Network Traffic
User Behavior
Large Intelligent Model with Global Information
Perception
Prediction
Inter-system Interference
Model Pruning and Knowledge Distillation
Control and Management Scheduling Layer
Multi-satellite Cooperative Network Configuration
Satellite Payload Types
Dynamic Topology of CCS Network
Satellite Payload Processing Capabilities
Regional Resource Management and Control Agent
Temporal and Spatial Distribution Characteristics of Services
Intra-system Interference
Multi-agent Selection
Real-Time Resource Allocation Layer
Knowledge Sharing
Transfer Learning
Collaborative Center
Collaborative Center
Resource Scheduling Edge Agent
Collaborative Model
Collaborative Model
Changing Channel State Information
Diverse QoS Requirements of Tasks
User position feedback

Figure 4 (Color online) CCS hierarchical layout and scheduling structure for heterogeneous resources.

resources, the regional autonomous control sub-centers will detect that the current resource scheduling policy does not meet the users' QoS requirements, resulting in degraded system performance. In this case, they will retrain or fine-tune the intelligent model according to the updated system configurations, then upload the newly intelligent model parameters through TT&C. In addition, the minimal incremental method is usually adopted. Only the changed parameters are uploaded, and there is no need to update all the parameters of the model, which further reduces the upload overhead.

2.5.2 *Heterogeneous resource hierarchical scheduling structure*

In order to achieve efficient management and scheduling of heterogeneous resources in the CCS system, this paper proposes a hierarchical layout and scheduling structure, as shown in Figure 4. From the uppermost to the lowermost level, the layout and management layer for the entire network, the control and scheduling layer for regions, and the real-time allocation layer for satellites are established. Considering the constraints of computational capability and the control scope of network elements at different hierarchical levels. AI-assisted resource scheduling agents of different scales are deployed at each layer. The goal is to achieve intelligent control and scheduling of multi-dimensional resources of communication, computation, and storage in the entire network, regions, and satellites.

Entire-network-oriented layout and management layer. By comprehensively considering the historical, perceived and predicted information, including but not limited to user behavior and distribution, dynamic change of service demand, real-time perception and prediction of network traffic, and the situation of inter-system interference outside the system, this layer utilizes the large intelligent model with global information to carry out system resources top-level planning, management and operation.

Region-oriented control and scheduling layer. It considers resource planning pool assigned by the upper layer, dynamic topology of regional network, temporal and spatial distribution of services, interference within the system, satellite payload types and processing capabilities. In the regional autonomous control sub-centers of this layer, the lightweight resource management and control agent, after model pruning and knowledge distillation, is used to execute multi-satellite cooperative scheduling on a large spatial and temporal scale to meet the task requirements of communication and computing in the regional subnetwork.

Edge-oriented real-time resource allocation layer. It considers changing channel state information, high-speed motion of LEO and dynamic time-varying cooperative relationship between satellites. Onboard resource schedulers are driven by diverse QoS requirements of computing and communication tasks. Through knowledge sharing and transfer learning among satellites, the computing costs of onboard resource management edge model can be reduced to solve the problem of limited onboard processing resources and achieve real-time allocation of heterogeneous resources of satellites.

In summary, based on the proposed heterogeneous resource layout and scheduling hierarchical structure of CCS, constrained by the computational capability and control scope of network elements at different hierarchical levels, different scales of AI-assisted resource scheduling agents (models) are deployed at each layer, laying the foundation for efficient utilization of heterogeneous resources.

3 Emerging technologies

This section primarily presents the utilization of the synergistic multi-satellite functions within CCS. The emerging technologies thereby achieved predominantly comprise cooperative computing, cooperative beam hopping, and cooperative transmissions. A state of the art is provided regarding the application of these technologies in CCS systems, with a particular emphasis on the analysis of the influence and challenges brought about by the resource management and scheduling in communication and computing.

3.1 Cooperative computing

For the current LEO system, by employing MEC, UE can offload local computing tasks to satellites [44]. Prior research has thoroughly examined task execution delays [45] and energy consumption [46], allocating computational and communication resources to maintain system performance [47–51]. Additionally, some studies have concentrated on parallel computing within the MEC framework [42] and resilience against network failures [52]. On the other hand, for the CCS system, this survey principally focuses on the substantial space computational tasks onboard, such as processing massive remote sensing image data [53]. In the CCS system, through multi-satellite distributed collaborative computing, the computing power of each collaborative satellite in the cluster is fully utilized to achieve efficient matching of computing resources and task requirements. In cluster satellites, cooperative computing between satellites requires consideration of the following issues, which refer to different resource allocation issues.

Allocation and routing of computing capability. The advent of multi-satellite collaborative computing has revolutionized the traditional single-satellite computation manner in satellite networks. However, the dynamic and time-varying network has resulted in low utilization of satellite resources and unbalanced computing load [54]. Consequently, it is imperative to investigate resource scheduling strategies for multiple satellites under intricate constraints among satellite nodes, employing a distributed satellite agent architecture [54]. The resource scheduling strategies include the allocation of inter satellite computing resources and end-to-end computing power flow planning, which enables better matching between onboard computing resources and computing demands. Furthermore, bandwidth contention on multi-hop links during computational data flow cannot be ignored. As a result, the end-to-end transmission demands for processing distributed data present significant challenges for dynamic collaborative resource management. To tackle these issues, Ref. [55] suggested a service graph-driven inter-satellite collaborative computing framework, which reimagines the collaborative computation problem as a graph mapping task from service graphs to network graphs. The work of [56] proposed a distributed collaborative computing approach rooted in multi-dimensional gradients, selecting path nodes involved in task processing to sequentially accomplish the computation and transmission of services. By capitalizing on space collaborative computing, the latency associated with task processing is considerably reduced when compared to both cloud computing and local computing alternatives.

Collaborative computation methodologies. Within the CCS system, a cooperative computing relationship has been formed between the orchestrating satellite and the satellites it coordinates, mandating the consideration of data security and privacy preservation concerns. To tackle these challenges, federated learning (FL) is employed for the collaborative processing of remote sensing data computation tasks [57], where the orchestrating satellite and its coordinated satellites ensure data confidentiality and security by exchanging only local model parameters, rather than raw data. Nevertheless, FL's parallel training of shared AI models on individual local datasets imposes substantial computational demands on participating devices. Split learning (SL) effectively addresses these issues by splitting the AI model into a device-side model and a server-side model at the cutting layer, training the AI model between devices and edge servers. Ref. [58] proposed a cluster based parallel training algorithm, and proposes a dual time scale algorithm that can jointly make cutting layer selection decisions on a large time scale, and jointly make device clustering and radio spectrum allocation decisions on a small time scale, reducing training latency. Thus, in the CCS system by employing SL, the model can be divided between collaborative satellites to improve computational efficiency through collaborative training. In these distributed computing methods, the computation of models and data of different scales on the CCS system requires the planning of collaborative computing to match the design of the collaborative model. The virtualization representation of computing resources is closely related to the model parameters, parameter transfer, and computational consumption of the distributed collaborative model.

Aggregation and semantic of data. Collaborative computing will generate a large number of onboard computing tasks and aggregation of computing data during the collaborative process. Aiming at the problems of poor aggregation ability and high aggregation delay of inter-satellite data, Ref. [54] suggested a data aggregation algorithm tailored for dynamic time-varying networks. This algorithm establishes an inter-satellite communication link model to execute data protocol and aggregation tasks. Furthermore, unlike conventional bit-level coordinated transmission protocols, semantic-aware technology facilitates efficient, information-centric coordinated transmission within the CCS system [59]. Thus, to maximize the utility of information per unit energy consumption within a limited connection window, Ref. [11] proposed a novel semantic-aware coordinated transmission scheme for CCS, by employing a multi-agent double and dueling deep Q-learning algorithm to improve the efficiency of cooperative observation tasks. Through semantic perception, inter satellite aggregation of data streams becomes more efficient, but for aggregated satellites, perceptual computing consumes more computational load on board.

In summary, key aspects of cooperative computing for CCS include computational resource management, collaborative methods and data processing. Resource management and scheduling optimize computing and storage resources through offloading and design end-to-end coordination of computing flow. Collaborative methods like federated and split learning improve the efficiency of distributed computing by sharing models and parameters while maintaining privacy and reducing latency. Semantic learning can help improve the processing efficiency of massive data aggregation. Regarding the application of collaborative computing for onboard computing tasks, experiments have also been conducted on remote sensing satellites such as the Chaohu-1, Changguang Satellite, and TaiJing-3, where the deployment of the Xingxi series space-based computing platforms has enabled multi-satellite collaborative computing. Through collaborative computing, these systems achieve real-time perception of spatial high-speed object anomalies. In the future, there will be a demand for AI as a service (AIaaS) provided by satellite, which will make the corresponding computing tasks on board more diverse, posing greater challenges to the collaborative deployment of computing tasks. At the same time, in the collaborative computing of cluster satellites, research on computing resource management, collaborative methods, and data processing has improved the efficiency of CCS systems in handling a large number of computing tasks. However, the application of related technologies also brings additional computing requirements, which require the design of resource models that take into account additional resource consumption.

3.2 Cooperative beam hopping

As previously mentioned, the traditional multibeam satellite system employs a rigid resource allocation strategy leading to resource inefficiencies [60]. Beam hopping technology lays the groundwork for flexible beam management. However, the current beam-hopping strategies for LEO system predominantly focus on single-satellite scenarios. In contrast, the research on multi-satellite beam-hopping remains in its infancy and requires more in-depth exploration for the CCS system. Additionally, current LEO beam-hopping is mainly designed for communication services. Yet, within the CCS system, there is an urgent need to conduct research on cooperative beam-hopping strategies dedicated to remote sensing applications or communication-sensing integrated functions. Thus, the application of beam hopping technology within the CCS system must also take into account the following design considerations.

Cell to satellite association. In contrast to the static of GEO systems, due to the high dynamics and multiple coverage of LEO satellites, the cooperative hopping beam of the CCS system must first address the association between the cell and the service satellites [61]. This association should consider satellite-UE distance, service time, load balancing, and a weighting of the above factors, which is determined by diverse QoS of services [62]. A satellite-terrestrial coordinated multi-satellite beam hopping scheduling framework is proposed in [63]. In this framework, the complex multi-satellite beam hopping problem is decomposed into a long-term sub-problem and a short-term sub-problem. The long-term sub-problem focuses on the cell-satellite association issue. It is resolved by a low-complexity iterative algorithm implemented in the network operation control center. The objective of this algorithm is to minimize the traffic load gap among satellites while taking interference avoidance into account. Specifically, this low-complexity iterative algorithm addresses the cell-satellite association problem in the following way. Initially, based on long-term traffic statistics, each cell is greedily allocated to a satellite to achieve the maximum possible inter-satellite load balance. Subsequently, an iterative process is designed. This process further reduces the load gap between satellites while simultaneously considering the avoidance of inter-satellite interference.

Interference avoidance. Since the hopping beam adopts full-frequency multiplexing to improve resource utilization, the co-frequency interference between satellites and beams in the CCS system cannot be ignored. It is of crucial importance to utilize the same frequency beams with spatial isolation while they are being activated or to alternate their operation in a temporal manner [64]. However, for high-demand service areas, simultaneous operation of adjacent beams with same-frequency may occur. Due to the complexities of interference mitigation methods and their impact on resource efficiency, adopting a precoding strategy is recommended. That is, only the beams suffering serious co-frequency interference in the same beam hopping time slot need to be precoded [65]. Furthermore, the integrated of terrestrial and satellite communication systems has highlighted inter-system interference, making spectrum sensing based interference avoiding vital for enhancing the beam-hopping capabilities of CCS systems [66].

Joint beam control and beam-hopping pattern design. Unlike traditional fixed beam, by utilizing phased-array beam shaping, it dynamically changes the beam position size to achieve load balancing [67]. Subsequently, taking into account co-frequency interference, beam revisit, beam dwell time, switching overhead and other factors, the beam hopping pattern design involving multi-satellite and multi-beam cooperation is carried out. On the other hand, for the collaborative beam hopping in the communication and sensing integrated CCS system, the design of the beam-hopping pattern is pivotal for the CCS system. For the communication function, it is crucial for achieving agile coverage. Regarding the radar function, it enables wide-area detection and targeted gazing at key areas. To solve these problems, polling beam is employed to complete the wide-area communication user service application and radar target position perception; on this basis, dynamic agile beams are employed to complete communication and sensing tasks. Thus, the beam pattern arrangement problem of communication and sensing integrated CCS system is transformed into the p-center problem [68] in the plane geometry. Its goal is that the communication user and radar target are covered with the least number of beams. By this method, the optimal arrangement of the beam position and hopping pattern not only solves the unevenness of the communication traffic problem, but also considers the wide-area detection of radar targets. Thus, it reduces the system delay, and lays the foundation for the efficient utilization of system resources.

In summary, the CCS system can effectively improve the quality of services such as communication and remote sensing, and realize the flexible scheduling of beam resources by using multi-satellite cooperative beam hopping, and can combine with beamforming to play to their respective advantages. Moreover, Huawei Technologies Company [69] has developed an experimental platform that is built following 5G

NTN protocol to validate cooperative beam hopping. They are seeking opportunities for in-orbit verification.

On the other hand, as previous mentioned, the CCS system is characterized by its multi-fold coverage and dense beam configuration, which significantly enhances capacity and coverage. However, this advancement introduces new challenges in beam-hopping. Key considerations include cells to satellites association, co-frequency interference avoidance, effective joint resource scheduling and beam control are also necessary, all of which complicate the optimization problem [66]. Moreover, due to the high dynamism of low-orbit satellites, the complexity and unpredictability of network topology, the diversity and temporal and spatial inequalities of business requirements, the differences in the capabilities of satellite-to-ground devices, and the complexity of the electromagnetic environment, the multi-satellite joint beam hopping of CCS systems has become difficult to mathematically model and solve. Therefore, current research focuses on the use of artificial intelligence methods to assist in decision scheduling. However, it is worth noting that the traditional artificial intelligence method based on data training [63] relies heavily on the transmission of a large amount of training data between user end points, satellites, and gateway stations, the computing and storage resource limitations of the equipment itself, and the signaling coordination between satellites and ground, which will become a challenge for the CCS system to achieve cooperative hopping beam.

3.3 Cooperative transmissions

Collaborative transmission of the CCS system employs multiple transmission points within a satellite cluster to deliver signals to users [12] increasing system capacity. By collaborating the beam resources of multiple satellites to serve a common coverage area, the system capacity can break through the resource limitations of different satellites serving successively. This collaborative transmission relies on precise coordination among cluster satellites. Formation control ensures that time delays and Doppler frequency shifts of received signals are approximately uniform [70], reducing the effects of incomplete information sharing and unprecise time-frequency synchronization during high-dynamic processes.

Building on the precise control of the CCS system's formation [9, 71–73], distributed MIMO and cooperative beamforming techniques enhance collaborative transmission performance through diversity gain and signal gain, respectively.

Distributed MIMO. Employing MIMO technology to satellite communication systems can effectively harness diversity gain to enhance satellite communication performance and increase information transmission rates [74]. On the basis of the original single satellite MIMO, multiple satellites in the CCS system utilize distributed MIMO to generate spatial diversity. The transmission model from multiple satellites to multiple user terminals in the same region is equivalent to a virtual MIMO system. Among them, multiple satellites form a giant antenna array and exchange data with multiple nodes on the ground [75]. In MIMO satellite communications that rely solely on spatial division multiplexing, channel capacity is determined by the geometric distribution of antennas [76] and inversely related to channel correlation [77]. The work of [78] has constructed a multi-antenna system using two satellites in different orbits, optimizing the layout of ground receiving antennas to maximize satellite communication system capacity. In consideration of the impact of satellite perturbation, fast moving and Doppler frequency shift, an optimization is carried out for the antenna layout scheme. With this scheme, the satellites within the formation can collaboratively form a large antenna array, thereby generating larger diversity gain and substantially enhancing the channel capacity of the distributed satellite MIMO system [79]. Through distributed MIMO technology, satellite clusters can distribute data processing tasks across multiple nodes, reducing the need for data transmission back to the ground and thus lowering latency [80].

Cooperative beamforming. It employs phased antenna arrays to boost signal gain towards the communication direction while mitigating interference in other directions. In CCS systems, several small satellites can constitute a loosely coupled phased array system by means of distributed phased arrays. This eliminates the requirement for a single satellite to be equipped with extremely large antenna arrays and remarkably diminishes manufacturing and launch costs [81]. Precise time and phase synchronization among distributed phased arrays is of utmost importance, thereby mandating accurate satellite formation control. On the one hand, by precisely controlling the specific radiation pattern of a large-scale satellite antenna array and generating directional transmission beams, the gain of signals in specific directions and positions can be improved; On the other hand, in order to better utilize scarce spectrum resources, satellite clusters can use full frequency reuse between beams of different satellites and multiple beams of the same satellite to meet the demand for high-capacity data transmission rates. The core issue that needs to be addressed in the above design is severe interference between multiple satellite beams [80]. The method of DPS (dynamic point selection) picks the single best transmission point based on CSI (channel state information), which can lessen inter-cell interference and it is appropriate for satellite cluster networks with numerous satellites [12]. Nevertheless, acquiring ideal real-time CSI is impractical in a satellite system. The work of [82] has proposed a multi-satellite cooperative robust beamforming strategies with imperfect CSI to effectively counteract co-channel interference and improve system capacity. Thus, in the CCS system, a key aspect of cooperative transmission is to coordinate the beam resources of different satellites on the basis of balancing CSI effectiveness and jamming performance.

In summary, leveraging precise formation control, the CCS system significantly boosts data transmission performance through cooperative transmission by utilizing distributed MIMO and cooperative beamforming techniques. On the 5G NTN broadband satellite communication system, CICT (China Information and Communication Technologies Group Corporation) has completed the end-to-end technical test verification in the orbit environment, built a 6G satellite ground fusion principle verification platform, and verified the multi beam cooperative transmission technology. However, multi-satellite and multi-beam cooperation, imperfect CSI and avoidance of co-frequency interference also bring new challenges to resource management and scheduling. In particular, unlike the acquisition of single satellite CSI, cooperative transmission needs to consider the inter satellite interaction under the common coverage of multiple satellites. With multiple satellites forming large-array antennas, the complexity of satellite-ground channels and the need for inter-satellite cooperation via ISL in joint planning pose significant challenges to resource planning in cooperative transmission. In addition, the high-speed and dynamic satellite network will bring more complex multi satellite selection problems, which make it more difficult to deal with the cooperative transmission under the dynamic satellite selection. The resource allocation between satellites and the task requirements between multiple points need more effective planning and matching.

4 Resource management modeling for CCS

This section primarily presents an overview of the resource management model in the CCS system, including communication traffic and computing task modeling, channel modeling, resource modeling and its performance metrics, which also form the foundation for subsequent optimization methods in resource management.

4.1 System modeling and analysis

4.1.1 *Communication traffic and computing task modeling*

Communication services can be classified into two main categories: delay-sensitive GBR (guaranteed bit rate) services (like real-time voice services), and non-delay-sensitive NGBR (non-guaranteed bit rate) services (like most non-real-time data services), and priority is set for different service categories [83]. Furthermore, modeling methods such as two-state Markov processes, truncated Lognormal distributions, and truncated Pareto distributions are employed for various traffic types including VoIP, FTP, web browsing, video streaming, and gaming [84]. In addition to the small-scale packet-level service models, a large-scale traffic model in terms of spatial distribution and temporal variation can be developed by considering factors such as the prevalence of satellite services, economic development, and human activities [85].

In the context of extensive computational tasks within a CCS system, which typically encompasses parameters such as data volume, the CPU cycle count necessary for each bit of input, and the maximum time threshold for task completion [86], the computational efficiency can be enhanced by decomposing a complex task into several sub-tasks, thereby creating a task flow [87] that can be allocated to the processor for concurrent processing. Since the computation of some tasks requires the result of the preceding tasks, and there is a priority constraint relationship among tasks, the scheduling issue is represented through a tree diagram of a directed acyclic graph, comprehensively considering the constraints such as communication delay and processor resources [87]. In addition, the allocation about independent tasks can be regarded as a 0/1 decision problem, while for the multi-task-computing node matching problem, centralized and distributed algorithms can be used respectively [88]. On the other hand, due to the dynamics of the edge computing scenario, it is necessary to investigate the distributed computing resource scheduling method for edge nodes. The main concept is that when edge devices are unable to handle computational tasks, they broadcast offloading requests. Idle devices respond to these requests and provide feedback regarding their computational resource availability. This approach further integrates the computational resources of idle devices with the latency and power requirements of the tasks that need to be scheduled [89], which facilitates task forwarding and execution to achieve efficient task scheduling. Finally, during the process of task offloading, the collaborative communication and transmission among satellite nodes should also be considered.

4.1.2 *Channel modeling*

Unlike the terrestrial mobile cellular communication system, the long propagation distance of the satellite-ground link, the complexity of signal fading and the high-speed movement of LEO satellites [90] all have a non-negligible impact on the performance of the CCS system. Consequently, these factors must be carefully considered in the management of communication and computation resources. Based on different channel modeling methods, it can be summarized as follows.

Empirical model. The measured data are processed by mathematical fitting methods to obtain a formula for calculating link fading, which is related to factors such as carrier frequency, elevation angle, antenna, and weather conditions of the satellite link [91]. Since the empirical model is a simple channel model based on field measurement data, it is able to describe the sensitivity of important parameters but cannot represent the propagation characteristics of the signal [92], while it is more difficult to measure satellite communications.

Statistical model. The electromagnetic wave propagation theory and statistical theory are used to determine the probability density function and establish the channel model. The single-state model usually assumes that the envelope of the signal at the receiving end obeys a definite probability distribution, such as Rice distribution, Lognormal distribution, Nakagami distribution, etc. [93–95], which is appropriate for narrowband stationary channels. In contrast, the multi-state model is founded on a Markov process, where each channel state is represented by a distinct probability distribution, making it suitable for wideband non-stationary channels.

Stochastic geometric model. The specific communication environment is mapped onto the geometry, which describes the positional relationship between the scatterer and the antenna. And different scenarios can be simulated by changing the shape of the scattering area or the distribution function of the scatterer [96, 97].

AI-based modeling. In order to cope with the increasingly massive and diverse wireless channel data demands, AI-based channel modeling methods have been widely studied in recent years, such as estimating the pathloss parameters and shadow fading using convolutional neural network (CNN) [98], as well as predicting the rain fading and atmospheric attenuation using long short-term memory [99]. This approach not only effectively extracts channel characteristics from complex channel data, but also adapts well to the high dynamics of satellite-ground channels.

4.1.3 *Resource modeling and virtualization characterization*

As previously stated, virtualization technology facilitates the abstraction and pooling of heterogeneous resources, enabling the virtualization of multiple satellites within the CCS system into a singular “large satellite”. This allows for the integrated management and scheduling of multidimensional resources within the system. Consequently, it is essential to develop a cohesive virtualization characterization that encompasses computation, storage, communication resource, and power consumption. Building upon this foundation, it is crucial to distill it into granular metrics such as computational accuracy, computational speed, communication bandwidth, cache size, and both communication and computational power [89]. The advantage of this metric is its ability to provide a standardized and shareable evaluation framework for heterogeneous resources in the CCS system.

Existing methods for virtual network mapping typically focus on static network topology. However, given the highly dynamics of the CCS system, these methods exhibit low resource utilization and necessitate frequent adjustments during network changes, leading to significant migration costs. Consequently, the work of [100] proposed a spatiotemporal association characterization method for SAGIN network resources. This approach aims to capture both the graph structure features and the temporal evolution characteristics of dynamic networks. Additionally, it seeks to identify fine-grained spatiotemporal association relationships, considering both the graph structure and the time series neighborhood of dynamic networks, thereby facilitating accurate modeling of spatiotemporal associations of network resources [101].

4.2 Metric for resource management

This part describes the main metrics for communication and computation resources management in the CCS system.

4.2.1 *Communication metrics*

Communication metrics can be generally categorized into throughput and system capacity, delay, and resource utilization rate [102].

Throughput. Communication throughput refers to the amount of data that can be successfully transmitted by the system within a unit of time, serving as a crucial performance metric for evaluating the efficiency of data transmission [103]. Maximization of throughput is one of the main objectives in the resource optimization research of satellite communication networks. The work of [104] proposed a capacity model for a multi-layer heterogeneous satellite network and analyzed the capacity performance of this network. Another work [105] investigated a frequency resources flexible allocation method to improve the throughput of multi-beam satellite communication system.

Delay. The main factors affecting delay are bandwidth, transmission distance and network quality. In practical, high communication delay can lead to problems such as conversation delay, stuttering and desynchronization, which negatively impact the quality of experience (QoE) for users. Recent studies have primarily concentrated on reducing the long-term average delay of the system, while also addressing the minimization of transmission and processing delay associated with tasks [106, 107].

Resource utilization rate. It typically denotes the proportion of actual resources used versus the total available resources. This metric can be used to evaluate the efficiency of resources utilization, including spectrum [108, 109], time, and power resources [110].

4.2.2 *Computation metrics*

Computation metrics can be primarily categorized into: task completion time, energy consumption, and the joint minimization of weighted total cost.

Task completion time. The work of [111] subdivided the latency into several factors, in which the main influencing factors are the processing time of tasks and transmission time of communication [112]. The processing time depends on the allocation of computing resources and task offloading strategy, while the transmission time depends on the allocation of communication resources and communication distance. The work of [113] modelled the task completion time with a probability distribution function and proposed a task completion efficiency evaluation function.

Energy consumption. This includes the power consumption of computers and cooling/heat dissipation equipment. Furthermore, in edge computing, the energy-constrained MEC servers must maintain normal operation, and this consumption is primarily related to the allocation of power and computational resources, as well as task offloading strategies. The energy consumption of edge computing nodes can be modeled as a single-increasing strictly convex function relative to the amount of computation, and its properties can be inscribed with a quadratic function [114]. Many researches focus on resource allocation with the goal of minimizing system energy consumption. The work of [115] analyzed power consumption with different scenarios such as data downloading, updating, and preloading. In the work of [116], the authors proposed a LEO satellite network with a three-layer offloading architecture for hybrid cloud and edge computing. Building on this framework, they investigate the challenges of resource allocation and task scheduling, aiming to reduce overall system energy consumption.

Weighted total cost. The cost function of a computational system can be defined as a weighted sum of task completion time and energy consumption. A trade-off between these two metrics can be realized. This is done by defining different weights according to application scenarios [50]. For instance, for an MEC-enhanced SAT-IoT network consisting of multiple satellites and multiple satellite gateways, a dynamic mixed-integer planning problem is established with the objective of minimizing delay and energy consumption [48]. Aiming at the problem of minimizing the weighted cost of computing services in the large LEO constellation, the study of [117] proposed a fast convergence algorithm based on game theory.

5 Resource management optimization problems and methods for CCS

Resource optimization methods of CCS can be mainly divided into: traditional optimization methods and AI-based optimization methods, which are summarized and analyzed as follows.

5.1 Traditional optimization methods

For efficiently allocating resources in the CCS system, the traditional mathematical optimization methods such as convex optimization, constrained Markov decision process (CMDP) and Lagrange dual decomposition, which help to keep the network stable and reduce delays. Also, for mixed integer nonlinear programming (MINLP), challenges can be solved by methods such as Lyapunov optimization and game theory. Furthermore, heuristic algorithms, including greedy algorithms, genetic algorithms, and swarm intelligence algorithms, along with resource allocation strategies which facilitate the decomposition of complex problems into several subproblems for dynamic planning and allocation resources.

5.1.1 *Classical mathematical optimization algorithms*

These methods mainly include convex optimization algorithms, CMDP, and Lagrange dual decomposition. The advantage of convex optimization algorithms is that the solution process is simple and reliable, which can ensure the existence of the global optimum, and many non-convex problems can be transformed into convex optimization problems or be approximated as convex optimization problems [116]. As the work in [118], the problem of maximizing the sum rate of the satellite-ground network is transformed into a convex power allocation problem by employing the minimum mean square error norm and log-linearization method. CMDP [119] differentiates from Markov in the aspect that it is required not only to maximize the results, but also to make the optimal decision under resource and scenario limitations. Lagrange multiplier method and linear programming are usually applied to cope with the CMDP problem. In the work of [119], the authors investigated the IoT task offloading problem in SAGIN by modeling the problem as CMDP and solving it by employing linear programming methods. Lagrange dual decomposition can simplify the solution process, which helps to split tasks and allocate resources to devices. It also provides a globally optimal solution in specific optimization problems [46].

5.1.2 *Heuristic algorithms*

Heuristic algorithms represent empirical and intuition-driven approaches. Typically, instead of aiming to discover an optimal solution within a complex problem, they endeavor to acquire a feasible approximate solution via simple and rapid algorithms [120]. Heuristic algorithms are typically appropriate for addressing large-scale, computationally intensive optimization challenges or non-deterministic polynomial (NP) problems that are arduous to be solved by exact algorithms. Heuristic algorithms mainly comprise greedy algorithms, genetic algorithms and swarm intelligence algorithms.

Greedy algorithm makes the best possible choice at each stage of the decision-making process, with the expectation that these local optimal selections will lead to a globally optimal solution [121]. The main steps of this algorithm include sorting, selection, and iteration. For the LEO-satellite based NB-IoT (narrowband Internet of Things) system [121], considering the high dynamics and propagation delay of LEO, a resource allocation model is established in order to maximize the resource utilization, where the problem falls into the 0-1 two-dimensional knapsack problem. An enhanced greedy algorithm is designed to achieve performance that closely aligns with the optimum solution while simultaneously decreasing the complexity of processing.

Genetic algorithm, as a search heuristic algorithm, mimics the process of biological evolution by constructing a simulated environment and then performing operations such as selection, crossover and mutation to continuously generate new solutions to approximate the optimal solution [122]. Through this mechanism, the genetic algorithm can better avoid falling into the local optimum and enhance the global search capability. Considering that a satellite network is a high complexity multi-layer heterogeneous system, the work of [123] employed a genetic algorithm to solve the NP-hard resource allocation problem to optimize QoE. The simulation shows that this algorithm has certain validity and stability for the dynamic allocation of satellite resources. For multi-beam satellite systems, considering the influences of the propagation effect, inter-beam interference and atmospheric attenuation, the work of [124] formulated the optimization problem of joint power and bandwidth allocation, and then combines repair functions and genetic algorithms for its solution. This approach not only guarantees the validity of the solutions but also enhances the convergence rate simultaneously.

Swarm intelligence algorithm represents a category of bio-inspired method that address optimization problems by simulating the collective behaviors observed in animal groups. Ant colony optimization (ACO) [125] and particle swarm optimization (PSO) are two typical swarm intelligence algorithms evolved from this concept [126]. A scheduling model has been developed considering satellites switching due to the movement of LEO, aimed at minimizing resource consumption of the satellites [127]. Building on this framework, the author derived the linear integer programming problem and subsequently employed the PSO algorithm for its resolution. Experimental findings also demonstrated that the performance of this algorithm outperforms that of genetic algorithms across all evaluated scenarios.

5.1.3 *Game theory*

Owing to the heterogeneity and high degree of dynamics within the CCS network, its resource allocation is confronted with significant challenges. Game theory, which treats the participants within a system as rational users, is used to design decentralization mechanisms that can effectively solve the problem of multiple rational participants making decisions regarding their goals [128]. And in these games, each participant or player seeks to optimize their respective goals [129, 130]. This part aims to deliver an examination of game-theoretic methods for resource allocation in the CCS system.

Cooperative game theory emphasizes the establishment of coalitions to collectively attain objectives or advantages. The creation of these coalitions can enhance the overall benefits for the members involved. A fundamental concern within this framework is the equitable allocation of the total benefits among the coalition participants [129]. According to the previous analysis, satellite clusters are inherently unstable due to dynamical network topology and unpredictable link failures. Therefore, by setting up a distributed federation, satellites in the LEO constellation are able to cooperate effectively to enhance the reliability of the CCS system and reduce the network's operation overhead [131]. This collaborative approach is capable of utilizing resources in a more efficient manner and improving the service quality of the CCS system.

For **non-cooperative game**, the current combination of strategies is a Nash equilibrium if no participant can unilaterally change the strategies to obtain a better outcome [129]. The work of [132] decomposed the minimizing total computational cost problem into two stochastic game problems: one for multiuser computing offloading and another for edge server deployment, and proves the existence of at least one Nash equilibrium for each game. For the satellite edge computing system, the work of [47] investigated the task offloading strategy using the response time and energy consumption as the metrics and proposed an iterative algorithm to find the Nash equilibrium.

Matching game is a special type of game that takes into account the orderly preferences of each group of participants over the other group members, and the participants need to be paired according to certain rules in order to achieve some form of cooperation or competition [133]. It has been applied in CCS for problems such as resource allocation, computational task offloading and network selection [123]. In the work of [134], a resource allocation algorithm that takes interference into account has been developed with the aim of optimizing the downlink capacity and data rate of the satellite-ground network. The optimization problem is disassembled into an interference management sub-problem and a user association sub-problem. Specifically, the interference issue is converted into a many-to-one matching game, which focuses on how to rationally distribute satellite resources among ground stations. Meanwhile, the user association sub-problem is transformed into a one-to-one matching game, dealing with the way in which individual users connect to these ground stations. Simulation results demonstrate that this approach can effectively suppress the inter-satellite interference.

5.1.4 *Dynamic programming strategy*

Dynamic programming strategy represents an analytical methodology that resolves intricate problems by dissecting them into more straightforward and tractable subproblems. This approach is particularly well-suited for grappling with issues that exhibit overlapping subproblems and optimal substructures [19].

In the work of [118], the authors tackled the interference challenge in MIMO satellites system. They consider constraints on spectrum and energy resource, and methodically decompose the intricate problem into three distinct sub-problems: hybrid precoding with the separation of analog and digital pre-coders, power allocation within the satellite system, and interference mitigation between the satellite and terrestrial links. These sub-problems are addressed sequentially to achieve an optimal solution. For the LEO-based edge computing network system, the work of [46] processed a complex energy minimization problem by solving two stratified subproblems respectively associated with the space and ground segments. Simulations show that an increase in the number of satellites leads to a higher offloading ratio of computational tasks and a reduction in energy consumption for edge IoT devices, which is an insight for future research on task offloading and resource allocation.

For space-air-ground-integrated electric power IoT, the work of [135] took into account that the joint optimization of task offloading and computational resource allocation is confronted with incomplete information, the curse of dimensionality, and the coupling between the long-term constraints of queuing delay and short-term decision-making. Then it puts forward a learning-based queue-aware task offloading and resource allocation algorithm. This algorithm decomposes the optimization problem into three deterministic sub-problems: task and resource allocation on the device side, resource allocation on the server side, and task offloading. For these three sub-problems, Lagrange dual decomposition, a queue-aware actor-critic based task offloading algorithm, and a greedy low-complexity algorithm are respectively employed.

5.2 AI-based optimization methods

The above methods, based on traditional optimization, provide potential solutions to the resource scheduling problem of the CCS system. However, with the increasing diversity of tasks and heterogeneity of resources in the CCS system, the resource scheduling problem of the system turns out to be extremely challenging to model. Consequently, traditional optimization methods struggle to handle it effectively. Therefore, researchers have begun to explore AI-based optimization methods. Data-driven methods, which build on AI, train neural networks with a large number of high-quality samples. One of its merits is that it can use extensive offline training to save online computational time. However, it is also beset by issues like poor interpretability and limited generalization ability [136]. Therefore, researchers consider the amalgamation of knowledge and data-driven approaches with the help of domain knowledge and expert knowledge to improve the interpretability of data-driven methods, which can also improve the convergence speed in situations with limited training samples. This section categorizes and summarizes AI-based methods and discusses the feasibility of applying some of them to the CCS system.

5.2.1 *Data-driven methods*

Distinct from traditional resource scheduling methods, data-driven methods will modify the original model during the iterative process. Moreover, they necessitate the system to repeatedly provide a substantial quantity of high-quality samples. Subsequently, the model performs self-updating by means of feedback mechanisms, and as the number of iterations rises, the optimal model is ultimately acquired. Data-driven methods include, but are not limited to, deep learning, reinforcement learning and federated learning. The following introduces the data-driven resource scheduling methods for the CCS system.

(1) **Deep learning.** Deep learning (DL) algorithm is proficient in handling complex input-output mapping relationships. Composed of multiple layers for feature extraction and transformation, DL is used for in-depth analysis in complex scenarios with massive data [137]. Considering the scarcity of spectrum resources and intense co-channel interference between satellites and terrestrial systems, intelligent resource allocation algorithms based on DL for high spectral efficiency and low co-channel interference have received extensive attention. To improve spectral efficiency under varying user densities, the work of [109] proposed a hierarchical architecture for satellite-terrestrial spectrum sharing. Based on this architecture, the intelligent resource management scheme of the spectrum management unit (SMU) composed of spectrum sensing, prediction and allocation is studied. The spectrum prediction utilizes a CNN method and the training data of the CNN is derived from historical detection results from spectrum sensing. Simulation results show that the proposed intelligent resource management scheme can achieve lower error detection.

(2) **Reinforcement learning.** Reinforcement learning (RL) is based on the Markov decision process, consisting of an environment and multiple agents. Agents choose from a set of actions according to the

policy and the current state of the environment, thereby causing the environment to transition into the next state. In this process, agents receive rewards as a measure of the quality of their policy. In general, RL can be categorized into multi-agent reinforcement learning, deep Q-learning (deep Q-network) and deep reinforcement learning [138].

Multi-agent reinforcement learning. CCS network is a comprehensive service system where each specific service function is defined as a domain in [139]. However, with the rapid growth of services and the explosion of data [139], the differentiation between domains becomes more significant and the resources demand and supply among domains also become increasingly unbalanced. Highly busy state satellites need to offload tasks to idle state satellites to improve the resource utilization ratio (RUR). The cross-domain resource scheduling (CDRS) problems should be studied to improve the CCS system performance. The work of [140] proposed a hierarchical sparse resource representation (HSRR) scheme to accurately characterize the resource of the networks. Based on the HSRR, a cross-domain dynamic multi-resource scheduling (CD-DMRS) algorithm is proposed using multi-agent reinforcement learning. Simulation results prove that the algorithm achieves higher RUR.

Deep Q-learning. Multi-agent reinforcement learning relies on explicit feature extraction and complete state observation, which is insufficient in high-dimensional state space problems. Deep Q-learning (DQL) employs end-to-end reinforcement learning to directly acquire successful strategies from high-dimensional sensory input, thereby effectively remedying this shortcoming [141].

CCS system aggregates satellite payloads from different orbits to complete various missions such as navigation, communication, remote sensing, and earth observing [20]. To improve the efficiency of the CCS system and avoid the transmission of worthless information, a semantic-aware method [11] is adopted to maximize the utility of information (UoI) and the correlation of semantic information between user devices within a limited connection window, thereby maximizing the UoI per unit energy consumption in the dynamic CCS system. By introducing a multi-agent double and dueling deep Q-learning (MAD3QL) algorithm, a higher average UoI per unit energy consumption is significantly realized [11]. On the other hand, the onboard battery power resource is worth considering. In shaded areas without solar energy, the data transmission of the satellite depends on its onboard batteries. From an economic perspective, battery life also indirectly affects the cost of satellite manufacturing. The life of the LEO satellite battery is influenced by the depth of discharge, which should be reduced per charging cycle. To increase the average lifetime of the LEO satellite battery, the work of [142] proposed a method based on Q-learning to control transmission power. In this manner, the degraded satellite will be allocated a reduced amount of data for processing, thereby extending the lifespan of the satellite battery.

When dealing with massive data processing, CCS system often needs to implement task offloading strategies and use distributed architectures for scheduling [143]. Although these tasks can be achieved by cloud computing, it may lead to severe delay [144]. Therefore, researchers propose MEC, aiming to shift the computation and storage capabilities from cloud computing to local servers. In this way, user-perceived latency and energy consumption are greatly reduced [145]. Furthermore, MEC can provide content caching and storage services [42]. The work of [146] considered a three-node MEC system with partial and binary of offloading models, and a joint computing and communication scheme is proposed to improve the energy efficiency of the nodes. Considering the constraints on system resources, rapid time-variation of channel states, heterogeneity of user terminals and mutual interference in satellite networks, it is impossible to solve complex collaborative tasks with a single optimization objective. Therefore, the work of [48] decomposed the problem into computing and communication resource allocation, as well as joint user association and offloading sub-problems. To collaboratively tackle the two sub-problems, a Lagrange multiplier and deep Q-network (DQN) method is proposed [48].

Deep reinforcement learning. DRL aims to get the optimal strategy by a training process according to the action feedback without prior knowledge. It is a learning framework based on DL and RL, being capable of addressing the issue of large state and action space [147]. Compared to GEO satellites, LEO satellites of the CCS system have limited onboard power, spectrum and channel resources. Thus, the work of [148] proposed an energy-saving channel allocation method. In this method, the dynamic channel allocation problem is modeled as a Markov decision process (MDP) and a DRL algorithm is used to solve the problem. Simulation results show that this method can effectively save system energy consumption.

(3) **Federated learning.** The machine learning methods described above perform centralized training mode [149–151]. Unlike these methods, the federated learning (FL) method is a distributed algorithm, which enables users to collaborate and share models while training and maintaining data on their own devices [152–154].

By introducing the FL method into the inter-satellite link resource scheduling problem, it is possible to significantly reduce system energy consumption, transforming centralized training into distributed training and enhancing system efficiency. Currently, most link scheduling methods mainly focus on optimizing channel capacity, power consumption and resource utilization, while neglecting the impact of actual traffic distribution. This results in suboptimal system delay performance. To achieve the optimal solution and simultaneously reduce the average number of hops among different satellites as well as decrease system energy consumption, a multi-agent deep reinforcement learning (MADRL) approach is employed to decompose the global scheduling of multiple satellites into independent scheduling for individual satellites [155]. Then, the authors propose an inter-orbit plane laser inter-satellite links scheduling algorithm. Based on inter-satellite cooperation, FL is used to implement the scheduling. Simulation results show that the proposed algorithm can effectively reduce the updating cost of the model [155].

Different from researches on the application of FL in satellites [154, 156], which do not consider the resource scheduling that integrates computation and communication, Ref. [157] proposed an energy-efficient communication-computation integrated FL algorithm. This algorithm minimizes total energy consumption by taking into packet error rate and completion time for each user. Compared with other baselines, this method can effectively reduce energy consumption and improve the algorithm efficiency of FL.

5.2.2 *Knowledge-driven methods*

Knowledge can be categorized into rather specialized and formalized scientific knowledge, everyday life's world knowledge, and more intuitive expert knowledge [158]. The knowledge-driven method mainly applies expert knowledge to solve various complex resource scheduling problems. Thus, for resource allocation, knowledge is defined as a summary of logical rules, theoretical algorithms and other content in the resource scheduling process, which can be used to guide or accelerate resource scheduling processes [159]. The most important feature of knowledge-driven machine learning is the dimensionality reduction of the input to the neural network in advance to simplify the learning network. Additionally, specific knowledge can also directly give approximate solutions to problems, thereby improving algorithm reliability [160]. On the one hand, knowledge-driven methods enhance the robustness and interpretability of data-driven methods by specifying logical rules in advance. For example, a protector mechanism is set to adjust the action generated by the agent to get the better system performance. The adjustment is based on information such as expert knowledge [161]. On the other hand, convergence speed is improved through knowledge sharing and transfer [162]. For example, the guiding reward is set in the analytical model that integrates expert knowledge, and transferred to combine with the observed reward collected by the agent from the environment to generate the updating reward [161].

However, it should be noticed that, the application of knowledge-driven methods to CCS systems may also bring the following potential problems and challenges. Firstly, the process of mining and summarizing domain knowledge or expert knowledge is relatively difficult, which is usually not easily obtainable through simple rules or models. Instead, sometimes it can be obtained by researchers or engineers in the relevant field through analysis, summarization, and experience. Secondly, domain knowledge or expert knowledge lacks universality. This means that the knowledge of the CCS system is applicable to the field of satellite communication and may not be directly applied to ground networks or other types of communication networks, resulting in limited knowledge reusability and universality. Conversely, the applicability of expert knowledge from other communication systems in the CCS system will also be greatly limited.

The following introduces knowledge-driven resource scheduling methods, applications and potential research directions in the CCS system.

(1) **Knowledge-assisted neighborhood search algorithm.** Neighborhood search algorithm is an efficient meta-heuristic framework for solving a large number of combinatorial optimization problems [163]. In the resource scheduling and management of large-scale CCS system, the support of satellite-ground links (SGLs) is essential, making the SGL scheduling problem (SGLSP) one of the focal points of research. To be specific, SGLSP focuses on determining which nonconflicting SGLs should be activated to maximize the total linking time. In order to solve this problem, a knowledge-assisted adaptive large neighborhood search algorithm (KA-ALNS) is used [164], integer programming is embedded in the ALNS framework, followed by the application of a data mining method called frequent pattern mining to extract the knowledge from excellent solutions to construct new solutions. Simulation results show that KA-ALNS algorithm, which combines data mining method with integer programming modeling, can effectively solve

SGLSP.

(2) **Knowledge-assisted evolutionary algorithm.** As previously stated, an evolutionary algorithm such as a genetic algorithm is a kind of global optimization algorithm inspired by biological evolution [165, 166], which can well solve the problem of satellite resource scheduling.

In order to improve the operational efficiency of the CCS system, the relay satellite can be employed to transmit command and data between the ground and mission satellites. However, the resources of the relay satellite are limited, which cannot cope with heavy relay tasks. It is urgent to develop a reasonable and effective task scheduling algorithm to improve overall system operational efficiency. Therefore, Ref. [167] proposed a hybrid integer mathematical model for task scheduling of a relay satellite system based on graph structure, utilizing a knowledge-based genetic algorithm (KBGA) to optimize the sequence of relay tasks. Through KBGA, the efficiency of the relay satellite system can be improved significantly.

Considering the optimization problem of multi-satellite resource scheduling, previous solutions assume that each TT&C task only needs to be executed once in a given time range [168], which is unrealistic in a practical system. Thus, considering the limited system resources, continuity of operation, and number of TT&C tasks, a knowledge-guided evolutionary algorithm is proposed [169]. In order to depict the distribution of terrestrial monitoring resources for TT&C operations and orbits, the algorithm adopts a tri-layer encoding technique. In addition, the search process can be accelerated through the guidance of conflict knowledge between tasks, thus enhancing both the quality and efficiency of the system. By using public satellite data, the simulation results demonstrate the superiority of the algorithm.

Although resource scheduling methods that combine knowledge with traditional classical algorithms can accelerate the convergence time and improve the algorithm performance to some extent, there are still shortcomings, such as a complex iterative process and a large calculation amount. Based on the AI approach, the knowledge-intergraded resource scheduling methods are divided into methods based on local knowledge and methods based on knowledge sharing and transferring.

(3) **Resource scheduling method based on local knowledge.** This method utilizes knowledge obtained from the local network training process to guide resource scheduling, reducing the action space dimension to achieve fast convergence.

As previously mentioned, beam hopping technology establishes the foundation for flexible beam management. However, the traditional beam-hopping method does not consider long-term returns, only seeking the optimal solution at the current moment, which leads to a significant increase in system complexity as service demands or the number of beams increase. In order to ensure inter-beam service fairness, minimize real-time service delay and maximize non-real-time service throughput, a model-free multi-objective deep reinforcement learning method is proposed [170]. The multi-action selection method based on double-loop learning is used to extract relevant knowledge and solve the problem of action dimension disaster. Simulation results under practical conditions show that this method can achieve multiple objectives simultaneously and intelligently allocate resources based on user demands and channel conditions.

Secondly, traditional scheduling methods such as proportional fair, round-robin, earliest-deadline-first and maximum throughput are not designed for delay-sensitive services. Since scheduling policy is a deterministic mapping from channel and queue states to scheduling actions, it can be optimized using the deep deterministic policy gradient (DDPG) algorithm. However, the convergence of DDPG is very slow and good QoS cannot be obtained. Therefore, a knowledge-assisted DDPG algorithm is proposed [171]. By leveraging expert knowledge in the design of the schedule, such as the knowledge of QoS, target scheduling policy and the importance of each training sample, the performance of the algorithm can be enhanced.

Finally, in order to overcome the huge losses caused by system crashes during the traditional reinforcement learning process, inspired by the protector mechanism in the field of electrical engineering, an analytical model and scheduling algorithm based on domain knowledge with physical rules are proposed [161]. The convergence speed and reliability of this algorithm are effectively improved by setting execution actions, protectors, and guiding rewards, while maintaining the stability of the search direction. It could be regarded as a prospective research avenue within the CCS system.

(4) **Resource scheduling method based on knowledge sharing and transfer.** Knowledge sharing can better facilitate the update of information and strategies between different networks, while transfer learning can utilize external expert knowledge to guide the target domain and accelerate the learning process. For practical reinforcement learning, the environment dynamics are usually unknown

and the agent can only take advantage of the knowledge in the environment through sufficient interaction and iteration. However, due to partial observability, sparse feedback, and the high complexity of the state and action space, it is generally difficult for agent to obtain sufficient interaction samples. In order to solve this problem, transfer learning (TL), also known as knowledge transfer, has become a hot topic of research [172]. Transfer learning means that a pre-trained model is reused in another task. By using a model applicable to different but related tasks, the search scope of possible models is narrowed in an advantageous way.

Traffic prediction is of great significance for resource management and scheduling of the CCS system. An accurate traffic prediction model can improve resource utilization efficiency. As an explosion of Internet traffic [173,174], several traffic prediction methods using machine learning in terrestrial communications have been provided. Nevertheless, these methods require large amounts of data and transmission bandwidth to train models that are not suitable for satellite systems. Therefore, a lightweight traffic prediction method based on machine learning is proposed to solve this problem, which uses knowledge transfer learning to reduce the consumption of bandwidth [175].

Previous studies based on beam hopping [30,176,177] may suffer some problems such as high algorithm complexity, large computational load and neglect of service delay sensitivity. In order to reduce the training cost and improve the generalization of the model in the dynamic environment, the combination of TL and DRL algorithms is proposed [178]. By leveraging the migratory characteristics of the TL strategy, satellites can quickly obtain optimal resource allocation schemes under conditions of limited samples. This approach is suitable for the dynamic topology of the CCS system. Through the migration of deep reinforcement learning algorithm, the average delay is effectively reduced and the system throughput is improved while guaranteeing QoS requirements. More importantly, simulation results prove that by introducing TL, convergence speed and the effect of the model can be improved.

6 Conclusion and future directions

The advent of large-scale LEO satellite constellations has brought increased attention to CCS systems. This paper aims to integrate insights from both academia and industry to provide a comprehensive analysis and recommendations for managing heterogeneous resources in CCS. It first discusses the evolution and components of CCS systems and then introduces a native AI architecture along with a hierarchical structure for resource scheduling. It also covers the modeling and metrics of resource management based on emerging technologies and applications brought by CCS systems, followed by a detailed review of resource management methods in CCS. Notably, this work is the first exhaustive review focusing on knowledge-driven resource scheduling strategies in CCS.

As CCS systems develop and task requirements increase, the interference of clustered satellites brought by spectrum sharing should be further considered. Moreover, the synchronization error between satellites, multimission planning and collaboration, and AIaaS serve as future research directions for resource scheduling. The details are as follows:

Anti-interference orientation. Current satellite systems are mainly affected by interference caused by complex electromagnetic environments. However, with their vast number of satellite nodes, complex network topology, extensive service coverage, and limited spectrum resources, CCS systems are susceptible to various types of interference, including interbeam, intersatellite, and even cochannel interference between different low-orbit systems. Meanwhile, direct-to-cell has inevitably become the development trend of CCS systems. It can improve spectrum resource utilization through spectrum sharing technology, alleviating the scarcity of spectrum resources in integrated space-ground systems. However, this also introduces interference between CCS and terrestrial mobile communication systems. Above all, interference in future CCS systems is becoming increasingly severe, making interference avoidance or anti-interference resource management and scheduling a critical research issue. Although traditional resource scheduling and management optimization solutions, such as power control, spectrum allocation, and beam management, can help avoid interference, the interference situation is more complex and dynamic for CCS systems. Thus, multisatellite cooperative interference avoidance or suppression methods must be implemented, including, but not limited to, the following: designing resource management architectures focused on spectrum sensing and anti-interference to achieve refined resource management and optimized configuration, proposing coordination mechanisms to enable effective collaboration among nodes, and facilitating interference information sensing and sharing, collaborative decision-making, and resource allocation. Therefore, anti-interference-oriented resource management and scheduling in CCS systems is emerging as a significant future research direction.

Impact of synchronization errors on resource scheduling. Compared with single-satellite systems, a distinctive feature of CCS systems is their aggregation of the service capabilities of conventional payload-integrated single-satellite platforms to form an ultracohesive virtual satellite. Therefore, synchronization has a great impact on the efficient use of resources in CCS systems. First, the synchronization of transmissions between satellites requires the precise alignment of time, frequency, and phase across distributed nodes [16]. Any deviation in synchronization can lead to significant communication quality degradation, such as increased bit error rate, signal distortion, or even communication interruptions. Moreover, these deviations can diminish the system's overall capacity and efficiency, leading to resource wastage, reduced transmission efficiency, and heightened multiaccess interference. From the analysis presented in Section 3 regarding the technologies within CCS systems, synchronization emerges as the fundamental enabler for collaborative computation, transmission, beam hopping, and observation. However, several factors, such as satellite orbit perturbations, signaling and data transmission delays, and highly dynamic changes in network topology, pose significant challenges to synchronization in tasks. These challenges must be considered in future research on CCS resource management. For instance, in cooperative beam hopping, the design of beam hopping patterns for multiple satellites must consider the beam scheduling desynchronization among satellites caused by the delay of control signaling. It must optimize the allocation of power and spectral resources to avoid interference caused by the simultaneous operation of beams of the same frequency in the neighboring wavelengths. In cooperative transmission, the amplitude-phase differences of phased arrays among multiple satellites caused by orbit perturbations must be thoroughly considered, as these differences can impact the performance of distributed MIMO and cooperative beamforming. For cooperative computation, transmission errors of computational data between satellites and the ground station must be addressed, and the reliability of data transmission and high efficiency of computation must be guaranteed through the joint scheduling of communication and computational resources. Thus, the consideration of the effects of synchronization errors in CCS resource management and scheduling becomes an important future research direction.

Multimission planning and collaboration. Mission planning in traditional single-satellite systems is relatively simple and uses centralized resource allocation. However, in CCS systems, mission realization usually relies on the collaborative operation of multiple satellites distributed in the same or different orbits, including weather prediction, scientific experiments, collaborative transmission, and remote sensing, which may lead to competition between multidimensional heterogeneous resources and uneven network load. Therefore, how to multitask planning and collaboration to efficiently utilize heterogeneous resources has become a worthwhile issue for resource management in CCS. Its research points include, but are not limited to, the following: for ground-satellite task offloading and collaboration, a dynamic load balancing algorithm can be designed on the basis of the state of the resource pool to address the issue of resource competition and uneven load distribution caused by massive tasks in the network. Meanwhile, the distributed satellite architecture of CCS systems is utilized to efficiently handle ground task offloading.

For intersatellite task offloading, the current multitask computing node matching problem is primarily categorized into centralized and distributed algorithms. Because of the dynamic nature of the edge computing environment in the CCS system, distributed resource scheduling methods must be developed. This approach must consider the available computational resources and the delay and power consumption requirements of the tasks to be scheduled [89]. For computing task collaboration, the dynamic changes in network topology necessitate a comprehensive consideration of computational and transmission resources, as well as the evolving service demands. Additionally, balancing resource efficiency with service quality and designing intelligent routing algorithms are crucial. Consequently, multitask planning and coordination have emerged as vital research directions for future CCS resource management and scheduling.

AI as a service. With the rise of AI large-model services, future CCS systems can also be considered to provide AI services. Satellite AIaaS represents the convergence of AI capabilities with satellite infrastructure, delivering on-demand AI services to users. Using application programming interfaces (APIs) or cloud interfaces, users can access powerful AI models hosted on satellites to obtain real-time services, such as data analysis and image processing [179]. Unlike traditional ground-based AIaaS, satellite AIaaS can extend coverage to remote regions, including oceans and mountainous areas, providing uninterrupted, all-weather, and full-coverage AI services. To adapt large AI models to the limited processing capabilities of satellites, techniques such as model pruning and distillation are essential. Additionally, the collaborative computation enabled by the CCS can compensate for the limitations of individual satellites, thereby enhancing the overall capacity for AI services. However, this also introduces new challenges to the resource management and scheduling of CCS systems. First, AI services demand substantial computational power, so a unified framework for metrics and management must be implemented. Given the dynamic characteristic of computational resource distribution in CCS, the status of available resources must be continuously updated and monitored. Second, intelligent routing of computational tasks is crucial. This involves a comprehensive assessment of the computational capacity and QoS of each satellite node to ensure efficient task distribution. Consequently, a system can guarantee the QoS and optimal user experience for AI services. In summary, ensuring the seamless delivery of satellite AIaaS has elevated CCS resource management and scheduling to a critical research direction for the future.

Acknowledgements This work was supported by National Key R&D Program of China (Grant No. 2022YFB2902600) and National Natural Science Foundation of China (Grant Nos. 61901230, U21A20450).

References

1 Centenaro M, Costa C E, Granelli F, et al. A survey on technologies, standards and open challenges in satellite IoT. IEEE Commun Surv Tut, 2021, 23: 1693–1720
2 Na D H, Park K H, Ko Y C, et al. Performance analysis of satellite communication systems with randomly located ground users. IEEE Trans Wireless Commun, 2021, 21: 621–634
3 Giordani M, Zorzi M. Non-terrestrial networks in the 6G era: challenges and opportunities. IEEE Netw, 2020, 35: 244–251
4 Sheng M, Zhou D, Bai W G, et al. Coverage enhancement for 6G satellite-terrestrial integrated networks: performance metrics, constellation configuration and resource allocation. Sci China Inf Sci, 2023, 66: 130303
5 Saad M M, Tariq M A, Khan M T R, et al. Non-terrestrial networks: an overview of 3GPP release 17 & 18. IEEE Int Things M, 2024, 7: 20–26
6 3GPP. Study on Management Aspects of Internet of Things (IoT) Non-Terrestrial Networks (NTN) Enhancements. Technical Report TR 28.841. 2023
7 3GPP. Study on Integration of Satellite Components in the 5G Architecture. Technical Report TR 23.700-29. 2024
8 Barbara N H, Lizy-Destrez S, Guardabasso P, et al. New GEO paradigm: re-purposing satellite components from the GEO graveyard. Acta Astronaut, 2020, 173: 155–163
9 Liu G P, Zhang S. A survey on formation control of small satellites. Proc IEEE, 2018, 106: 440–457
10 Saeed N, Elzanaty A, Almorad H, et al. CubeSat communications: recent advances and future challenges. IEEE Commun Surv Tut, 2020, 22: 1839–1862
11 Xu L, Jiao J, Jiang S Y, et al. Semantic-aware coordinated transmission in cohesive clustered satellites: utility of information perspective. Sci China Inf Sci, 2024, 67: 199301
12 Jung D H, Im G, Ryu J G, et al. Satellite clustering for non-terrestrial networks: concept, architectures, and applications. IEEE Veh Technol Mag, 2023, 18: 29–37
13 Wang P, Zhang J, Zhang X, et al. Convergence of satellite and terrestrial networks: a comprehensive survey. IEEE Access, 2019, 8: 5550–5588
14 Li X T, Xu S, Zhao Z P, et al. A survey on computing offloading in satellite-terrestrial integrated edge computing networks. In: Proceedings of the 15th International IEEE Conference on Communication Software and Networks (ICCSN), Shenyang, 2023. 172–182
15 Chen Q, Guo Z, Meng W, et al. A survey on resource management in joint communication and computing-embedded SAGIN. IEEE Commun Surv Tut, 2025, 27: 1911–1954
16 Marrero L M, Merlano-Duncan J C, Querol J, et al. Architectures and synchronization techniques for distributed satellite systems: a survey. IEEE Access, 2022, 10: 45375–45409
17 Wang S, Li Q. Satellite computing: vision and challenges. IEEE Int Things J, 2023, 10: 22514–22529
18 Sharif S, Zeadally S, Ejaz W. Space-aerial-ground-sea integrated networks: resource optimization and challenges in 6G. J Netw Comput Appl, 2023, 215: 103647
19 Liang H, Yang Z, Zhang G, et al. Resource allocation for space-air-ground integrated networks: a comprehensive review. J Commun Inf Netw, 2024, 9: 1–23
20 Deng R, Di B, Song L. Ultra-dense LEO satellite based formation flying. IEEE Trans Commun, 2021, 69: 3091–3105
21 Azari M M, Solanki S, Chatzinotas S, et al. Evolution of non-terrestrial networks from 5G to 6G: a survey. IEEE Commun Surv Tut, 2022, 24: 2633–2672
22 Goto D, Shibayama H, Yamashita F, et al. LEO-MIMO satellite systems for high capacity transmission. In: Proceedings of IEEE Global Communications Conference (GLOBECOM), Abu Dhabi, 2018. 1–6
23 Sulli V, Giancristofaro D, Santucci F, et al. An analytical method for performance evaluation of digital transparent satellite processors. In: Proceedings of IEEE Global Communications Conference (GLOBECOM), Washington, 2016. 1–7
24 Kodheli O, Lagunas E, Maturo N, et al. Satellite communications in the new space era: a survey and future challenges. IEEE Commun Surv Tut, 2020, 23: 70–109
25 Angeletti P, de Gaudenzi R, Lisi M. From "bent pipes" to "software defined payloads": evolution and trends of satellite communications systems. In: Proceedings of the 26th AIAA International Communications Satellite Systems Conference (ICSSC), San Diego, 2008. 1–10
26 Sulli V, Santucci F, Faccio M, et al. Performance of satellite digital transparent processors through equivalent noise. IEEE Trans Aerosp Electron Syst, 2018, 54: 2643–2661
27 Gupta R K. Communications satellite RF payload technologies evolution: a system perspective. In: Proceedings of Asia Pacific Microwave Conference (APMC), New Delhi, 2016. 1–4
28 Whitefield D, Gopal R, Arnold S. Spaceway now and in the future: on-board IP packet switching satellite communication network. In: Proceedings of IEEE Military Communications Conference, Washington, 2006. 1–7
29 Fenech H, Sonya A, Tomatis A, et al. Eutelsat quantum: a game changer. In: Proceedings of the 33rd AIAA International Communications Satellite Systems Conference and Exhibition, Queensland, 2015. 1–7
30 Lei L, Lagunas E, Yuan Y, et al. Beam illumination pattern design in satellite networks: learning and optimization for efficient beam hopping. IEEE Access, 2020, 8: 136655
31 Pecorella T, Fantacci R, Lasagni C, et al. Study and implementation of switching and beam-hopping techniques in satellites with on board processing. In: Proceedings of IEEE International Workshop on Satellite and Space Communications, Salzburg, 2007. 206–210
32 Freedman J B, Marshack D S, Kaplan T, et al. Advantages and capabilities of a beamforming satellite with a space-based digital processor. In: Proceedings of the 32nd AIAA International Communications Satellite Systems Conference, San Diego,

2014. 1–7
33 Morris I, Clarke O, Wheatley N, et al. Airbus Defence and Space: Ku band multiport amplifier powers HTS payloads into the future. In: Proceedings of the 33rd AIAA International Communications Satellite Systems Conference and Exhibition, Queensland, 2015. 4340
34 Bai B, Zhang Y, Lu Z, et al. A Novel space-based computing platform system based on an open architecture. J China Acad Electron Inf Technol, 2020, 15: 900–904
35 Lu S, Liang H, Liu D. Thoughts on the status quo and development of localized on-board computer technology. Mobile Inf, 2018, 40: 126–129
36 Yu Z, Feng X, Dai T, et al. Space edge computing: requirement, architecture and key technique. J Electron Inf Technol, 2022, 44: 4416–4425
37 Liao X. Review on networking and control technologies of space satellite network. Space-Integrated-Ground Inf Netw, 2023, 4: 48–58
38 Bi Y, Han G, Xu S, et al. Software defined space-terrestrial integrated networks: architecture, challenges, and solutions. IEEE Netw, 2019, 33: 22–28
39 Cao B, Zhang J, Liu X, et al. Edge-cloud resource scheduling in space-air-ground-integrated networks for Internet of Vehicles. IEEE Int Things J, 2021, 9: 5765–5772
40 Tang Q, Xie C, Liu X, et al. MEC enabled satellite-terrestrial network: architecture, key technique and challenge. J Commun, 2020, 41: 162–181
41 Cao S, Sun X, Wang H, et al. Review of intelligent routing technology in integrated satellite-ground network. Space-Integrated-Ground Inf Netw, 2021, 2: 11–19
42 Zhang Z, Zhang W, Tseng F H. Satellite mobile edge computing: improving QoS of high-speed satellite-terrestrial networks using edge computing techniques. IEEE Netw, 2019, 33: 70–76
43 Cheng X, Lyu F, Quan W, et al. Space/aerial-assisted computing offloading for IoT applications: a learning-based approach. IEEE J Sel Areas Commun, 2019, 37: 1117–1129
44 Zhang Y D. Research on edge computing architecture of mobile satellite communication network. Dissertation for the Master Degree. Chengdu: University of Electronic Science and Technology of China, 2020
45 Han J, Wang H, Wu S, et al. Task scheduling of high dynamic edge cluster in satellite edge computing. In: Proceedings of IEEE World Congress on Services, Beijing, 2020. 287–293
46 Song Z, Hao Y, Liu Y, et al. Energy-efficient multiaccess edge computing for terrestrial-satellite Internet of Things. IEEE Int Things J, 2021, 8: 14202–14218
47 Wang Y, Yang J, Guo X, et al. A game-theoretic approach to computation offloading in satellite edge computing. IEEE Access, 2020, 8: 12510–12520
48 Cui G, Li X, Xu L, et al. Latency and energy optimization for MEC enhanced SAT-IoT networks. IEEE Access, 2020, 8: 55915–55926
49 Yu S, Gong X, Shi Q, et al. EC-SAGINs: edge-computing-enhanced space-air-ground-integrated networks for Internet of Vehicles. IEEE Int Things J, 2021, 9: 5742–5754
50 Chen Q, Meng W, Quek T Q S, et al. Multi-tier hybrid offloading for computation-aware IoT applications in civil aircraft-augmented SAGIN. IEEE J Sel Areas Commun, 2022, 41: 399–417
51 Yu X, Chen Y, Liu H, et al. Strategy of joint resource allocation and computation offloading in LEO satellite edge computing scenario. J Nanjing Univ Posts Telecommun, 2021, 41: 1–9
52 Xie R, Tang Q, Wang Q, et al. Satellite-terrestrial integrated edge computing networks: architecture, challenges, and open issues. IEEE Netw, 2020, 34: 224–231
53 Cheng S, Li H, Bai W, et al. Resource scheduling method for integration of TT&C and observation based on multi-agent deep reinforcement learning. Space-Integrated-Ground Inf Netw, 2023, 4: 12–22
54 Zhang X. Multi-satellite collaborative computing method based on time-varying network. Dissertation for the Master Degree. Changsha: Central South University, 2023
55 Guo X, Ren Z, Cheng W, et al. Inter-satellite cooperative computing scheme driven by business graph in LEO satellite network. Space-Integrated-Ground Inf Netw, 2021, 2: 35–44
56 Ma B, Ren Z, Li Z. Multi-dimensional gradient based high-reliability collaborative computing method for satellite network. ZTE Technol J, 2021, 27: 36–42
57 Chen H, Chen X, Ma R, et al. Privacy preserving scheme of federated learning for remote sensing data (in Chinese). J Comput Appl, 2025, 45: 506–517
58 Wu W, Li M, Qu K, et al. Split learning over wireless networks: parallel design and resource management. IEEE J Sel Areas Commun, 2023, 41: 1051–1066
59 Mu X, Liu Y, Guo L, et al. Heterogeneous semantic and bit communications: a semi-NOMA scheme. IEEE J Sel Areas Commun, 2023, 41: 155–169
60 Li B, Fei Z, Zhou C, et al. Physical-layer security in space information networks: a survey. IEEE Int Things J, 2020, 7: 33–52
61 Shi L, Yang F, Wu W, et al. Load balancing and remaining visible time based handover algorithm for LEO satellite network. In: Proceedings of the 8th International Conference on Computer and Communications, Chengdu, 2022. 391–395
62 He S, Wang T, Wang S. Load-aware satellite handover strategy based on multi-agent reinforcement learning. In: Proceedings of IEEE Conference and Exhibition on Global Telecommunications, Taipei, 2020. 1–6
63 Lin Z, Ni Z, Kuang L, et al. Satellite-terrestrial coordinated multi-satellite beam hopping scheduling based on multi-agent deep reinforcement learning. IEEE Trans Wireless Commun, 2024, 23: 10091–10103
64 Lin Z, Ni Z, Kuang L, et al. Multi-satellite beam hopping based on load balancing and interference avoidance for NGSO satellite communication systems. IEEE Trans Commun, 2023, 71: 282–295
65 Shi T, Liu Y, Kang S, et al. Angle-based multicast user selection and precoding for beam-hopping satellite systems. IEEE Trans Broadcast, 2023, 69: 856–871
66 Zhu J, Sun Y, Peng M. Beam management in low earth orbit satellite communication with handover frequency control and satellite-terrestrial spectrum sharing. IEEE Trans Commun, 2025, 73: 5247–5263
67 Zhao X, Wang C, Cai S, et al. Multi-satellite cooperative load-balancing scheme based on dynamic beam coverage for LEO beam hopping systems. IEEE Wireless Commun Lett, 2024, 13: 2892–2896
68 Tang J, Bian D, Li G, et al. Optimization method of dynamic beam position for LEO beam-hopping satellite communication systems. IEEE Access, 2021, 9: 57578–57588
69 Wang Y, Chen Y, Qiao Y, et al. Cooperative beam hopping for accurate positioning in ultra-dense LEO satellite networks. In: Proceedings of IEEE International Conference on Communications Workshops (ICC Workshops), Montreal, 2021. 1–6
70 Wang Y. Research on cooperative spectrum sensing and sharing technology for LEO satellites. Dissertation for the Doctoral Degree. Nanjing: Nanjing University of Posts and Telecommunications, 2022
71 Chang I, Park S Y, Choi K H. Nonlinear attitude control of a tether-connected multi-satellite in three-dimensional space. IEEE Trans Aerosp Electron Syst, 2010, 46: 1950–1968
72 Serrani A. Robust coordinated control of satellite formations subject to gravity perturbations. In: Proceedings of the

American Control Conference, Denver, 2003. 302–307
73 Dang Z, Zhang Y. Control design and analysis of an inner-formation flying system. IEEE Trans Aerosp Electron Syst, 2015, 51: 1621–1634
74 Hodge D B. An improved model for diversity gain on earth-space propagation paths. Radio Sci, 1982, 17: 1393–1399
75 Wang H, Ye N, An J. Multi-satellite cooperative signal detection for low earth orbit constellations. ZTE Commun, 2021, 27: 12–17
76 Hofmann C, Storek K U, Schwarz R T, et al. Spatial MIMO over satellite: a proof of concept. In: Proceedings of IEEE International Conference on Communications (ICC), Kuala Lumpur, 2016. 1–6
77 Yang H, He Y. Correlated channel capacity of virtual MIMO based on distributed satellites. Comput Sci, 2015, 42: 276–278
78 Schwarz R T, Knopp A, Lanki B, et al. Optimum-capacity MIMO satellite broadcast system: conceptual design for LOS channels. In: Proceedings of the 4th Advanced Satellite Mobile Systems, Bologna, 2008. 66–71
79 Zhang L, Zhu L, Ju C. Generalized MIMO channel model and its capacity analysis in formation flying satellite communication systems. In: Proceedings of the 6th International ICST Conference on Communications and Networking in China (CHINACOM), Harbin, 2011. 1079–1082
80 Xu L, Jiao J, Zhang Q. Code-domain cooperative transmission technology for cohesive clustered satellites. ZTE Commun, 2024, 30: 24–29
81 Jiao L, Li W, Tong J, et al. Satellite phased arrays for direct-to-handset satellite: key technologies and future vision. Telecommun Sci, 2024, 40: 30–42
82 Hong T, Liu R, Liu Z, et al. An asynchronous collision-tolerant ACRDA scheme based on satellite-selection collaboration-beamforming for LEO satellite IoT networks. Sensors, 2023, 23: 3549
83 Kong Z, Kwok Y K, Wang J Z. A low-complexity QoS-aware proportional fair multicarrier scheduling algorithm for OFDM systems. IEEE Trans Veh Technol, 2008, 58: 2225–2235
84 3GPP. LTE Physical Layer Framework for Performance Verification Radio Access Network. TS R1-070674. 2007
85 Hu X, Liu S, Wang Y, et al. Deep reinforcement learning-based beam hopping algorithm in multibeam satellite systems. IET Commun, 2019, 13: 2485–2491
86 Fang H, Zhao Y, Gao Y, et al. A satellite edge computing resource allocation and offloading algorithm with task dependence. Comput Eng Sci, 2022, 44: 1951–1958
87 Qu K, Zhuang W, Ye Q, et al. Dynamic flow migration for embedded services in SDN/NFV-enabled 5G core networks. IEEE Trans Commun, 2020, 68: 2394–2408
88 Meng J, Tan H, Li X Y, et al. Online deadline-aware task dispatching and scheduling in edge computing. IEEE Trans Parallel Distrib Syst, 2020, 31: 1270–1286
89 Zhu S Q, Xu Q Q, Li X T, et al. Computational measurement and task scheduling: a study on IoT edge device strategies. Telecommun Sci, 2024, 40: 122–138
90 Baeza V M, Lagunas E, Al-Hraishawi H, et al. An overview of channel models for NGSO satellites. In: Proceedings of the 96th IEEE Vehicular Technology Conference (VTC2022-Fall), London, 2022. 1–6
91 Moraitis N, Milas V, Constantinou P. On the empirical model comparison for the land mobile satellite channel. In: Proceedings of the 65th IEEE Vehicular Technology Conference (VTC), Dublin, 2007. 1405–1409
92 Al-Hourani A, Guvenc I. On modeling satellite-to-ground path-loss in urban environments. IEEE Commun Lett, 2021, 25: 696–700
93 Sharma P K, Yogesh B, Gupta D, et al. Performance analysis of IoT-based overlay satellite-terrestrial networks under interference. IEEE Trans Cogn Commun Netw, 2021, 7: 985–1001
94 Clemente M C, Paris J F. Closed-form statistics for sum of squared Rician shadowed variates and its application. Electron Lett, 2014, 50: 120–121
95 Abdi A, Lau W C, Alouini M, et al. A new simple model for land mobile satellite channels: first- and second-order statistics. IEEE Trans Wireless Commun, 2003, 2: 519–528
96 Jung D H, Ryu J G, Byun W J, et al. Performance analysis of satellite communication system under the shadowed-Rician fading: a stochastic geometry approach. IEEE Trans Commun, 2022, 70: 2707–2721
97 Bai L, Wang C X, Goussetis G, et al. Channel modeling for satellite communication channels at Q-band in high latitude. IEEE Access, 2019, 7: 137691
98 Ates H F, Hashir S M, Baykas T, et al. Path loss exponent and shadowing factor prediction from satellite images using deep learning. IEEE Access, 2019, 7: 101366
99 Bai L, Xu Q, Huang Z, et al. An atmospheric data-driven Q-band satellite channel model with feature selection. IEEE Trans Antennas Propagat, 2022, 70: 4002–4013
100 Mai T, Yao H, Xin X, et al. Spatiotemporal correlation representation based precise resource management in space-air-ground integrated network. Space-Integrated-Ground Inf Netw, 2024, 5: 34–42
101 Sankar A, Wu Y, Gou L, et al. Dynamic graph representation learning via self-attention networks. 2018. ArXiv:1812.09430
102 Fang X, Feng W, Wei T, et al. 5G embraces satellites for 6G ubiquitous IoT: basic models for integrated satellite terrestrial networks. IEEE Int Things J, 2021, 8: 14399–14417
103 Jia M, Zhang X, Gu X, et al. Joint UE location energy-efficient resource management in integrated satellite and terrestrial networks. J Commun Inf Netw, 2018, 3: 61–66
104 Jiang C, Zhu X. Reinforcement learning based capacity management in multi-layer satellite networks. IEEE Trans Wireless Commun, 2020, 19: 4685–4699
105 Kawamoto Y, Kamei T, Takahashi M, et al. Flexible resource allocation with inter-beam interference in satellite communication systems with a digital channelizer. IEEE Trans Wireless Commun, 2020, 19: 2934–2945
106 Wang G, Zhou S, Niu Z. Radio resource allocation for bidirectional offloading in space-air-ground integrated vehicular network. J Commun Inf Netw, 2019, 4: 24–31
107 Feng X, Sun Y, Peng M. Distributed satellite-terrestrial cooperative routing strategy based on minimum hop-count analysis in mega LEO satellite constellation. IEEE Trans Mobile Comput, 2024, 23: 10678–10693
108 Yan Y, An K, Zhang B, et al. Outage-constrained robust multigroup multicast beamforming for satellite-based Internet of Things coexisting with terrestrial networks. IEEE Int Things J, 2021, 8: 8159–8172
109 Jia M, Zhang X, Sun J, et al. Intelligent resource management for satellite and terrestrial spectrum shared networking toward B5G. IEEE Wireless Commun, 2020, 27: 54–61
110 Ruan Y, Jiang L, Li Y, et al. Energy-efficient power control for cognitive satellite-terrestrial networks with outdated CSI. IEEE Syst J, 2021, 15: 1329–1332
111 Brogi A, Forti S, Guerrero C, et al. How to place your apps in the fog: state of the art and open challenges. Softw Pract Exp, 2020, 50: 719–740
112 Zhang S, Cui G, Long Y, et al. Joint computing and communication resource allocation for satellite communication networks with edge computing. China Commun, 2021, 18: 236–252
113 Gorlatova M, Inaltekin H, Chiang M. Characterizing task completion latencies in fog computing. 2018. ArXiv:1811.02638
114 Deng R, Lu R, Lai C, et al. Optimal workload allocation in fog-cloud computing towards balanced delay and power consumption. IEEE Int Things J, 2016, 3: 1171–1181

115 Jalali F, Hinton K, Ayre R, et al. Fog computing may help to save energy in cloud computing. IEEE J Sel Areas Commun, 2016, 34: 1728–1739
116 Tang Q, Fei Z, Li B, et al. Computation offloading in LEO satellite networks with hybrid cloud and edge computing. IEEE Int Things J, 2021, 8: 9164–9176
117 Jia M, Zhang L, Wu J, et al. Joint computing and communication resource allocation for edge computing towards Huge LEO networks. China Commun, 2022, 19: 73–84
118 Peng D, Li Y, Chatzinotas S, et al. Hybrid analog-digital precoding for mmWave coexisting in 5G-satellite integrated network. In: Proceedings of the 31st Annual International Symposium on Personal, Indoor and Mobile Radio Communications, London, 2020. 1–6
119 Zhou C, Wu W, He H, et al. Delay-aware IoT task scheduling in space-air-ground integrated network. In: Proceedings of IEEE Global Communications Conference, Waikoloa, U2019. 1–6
120 Desale S, Rasool A, Andhale S, et al. Heuristic and meta-heuristic algorithms and their relevance to the real world: a survey. Int J Comput Eng Res Trends, 2015, 351: 2349–7084
121 Kodheli O, Maturo N, Chatzinotas S, et al. NB-IoT via LEO satellites: an efficient resource allocation strategy for uplink data transmission. IEEE Int Things J, 2022, 9: 5094–5107
122 Kramer O. Genetic Algorithms. Berlin: Springer International Publishing, 2017
123 Deng X, Zhu L, Tan Q, et al. Multi-layer satellite network resource management based on genetic algorithm. In: Proceedings of International Symposium on Networks, Computers and Communications (ISNCC), Dubai, 2021. 1–6
124 Paris A, Del Portillo I, Cameron B, et al. A genetic algorithm for joint power and bandwidth allocation in multibeam satellite systems. In: Proceedings of IEEE Aerospace Conference, Big Sky, 2019. 1–15
125 Picchi R, Chiti F, Fantacci R, et al. Towards quantum satellite internetworking: a software-defined networking perspective. IEEE Access, 2020, 8: 210370
126 Wang D, Tan D, Liu L. Particle swarm optimization algorithm: an overview. Soft Comput, 2018, 22: 387–408
127 Pachler N, Crawley E F, Cameron B G. Beam-to-satellite scheduling for high throughput satellite constellations using particle swarm optimization. In: Proceedings of IEEE Aerospace Conference (AERO), Big Sky, 2022. 1–9
128 Wang S, Hu Z, Deng Y, et al. Game-theory-based task offloading and resource scheduling in cloud-edge collaborative systems. Appl Sci, 2022, 12: 6154
129 Marden J R, Shamma J S. Game theory and control. Annu Rev Control Robot Auton Syst, 2018, 1: 105–134
130 Li F, Lam K Y, Chen H H, et al. Spectral efficiency enhancement in satellite mobile communications: a game-theoretical approach. IEEE Wireless Commun, 2020, 27: 200–205
131 Liu J, Zhang X, Zhang R, et al. Reliable and low-overhead clustering in LEO small satellite networks. IEEE Int Things J, 2022, 9: 14844–14856
132 Ning Z, Yang Y, Wang X, et al. Dynamic computation offloading and server deployment for UAV-enabled multi-access edge computing. IEEE Trans Mobile Comput, 2023, 22: 2628–2644
133 Ren J, Xia F, Chen X, et al. Matching algorithms: fundamentals, applications and challenges. IEEE Trans Emerg Top Comput Intell, 2021, 5: 332–350
134 Zhang L, Liu J, Sheng M, et al. Interference-aware resource allocation in satellite integrated terrestrial networks. In: Proceedings of IEEE/CIC International Conference on Communications in China (ICCC), Foshan, 2022. 654–659
135 Liao H, Zhou Z, Zhao X, et al. Learning-based queue-aware task offloading and resource allocation for space-air-ground-integrated power IoT. IEEE Int Things J, 2021, 8: 5250–5263
136 Li W J, Li J H, Zhang C, et al. A survey on data and knowledge-driven intelligent resource scheduling for LEO satellite. Space Electron Technol, 2023, 20: 42–51
137 Pham Q V, Ruby R, Fang F, et al. Aerial computing: a new computing paradigm, applications, and challenges. IEEE Int Things J, 2022, 9: 8339–8363
138 Hu J, Zhang H, Song L, et al. Reinforcement learning for a cellular Internet of UAVs: protocol design, trajectory control, and resource management. IEEE Wireless Commun, 2020, 27: 116–123
139 Zhou D, Sheng M, Wu J, et al. Gateway placement in integrated satellite-terrestrial networks: supporting communications and Internet of Remote Things. IEEE Int Things J, 2022, 9: 4421–4434
140 Bao C, Sheng M, Zhou D, et al. Toward intelligent cross-domain resource coordinate scheduling for satellite networks. IEEE Trans Wireless Commun, 2023, 22: 9610–9625
141 Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning. Nature, 2015, 518: 529–533
142 Tsuchida H, Kawamoto Y, Kato N, et al. Efficient power control for satellite-borne batteries using Q-learning in low-earth-orbit satellite constellations. IEEE Wireless Commun Lett, 2020, 9: 809–812
143 Wang Y, Sun Q. Task offloading and resource allocation in satellite terrestrial networks: a deep deterministic policy gradient approach. In: Proceedings of International Conference on Big Data Intelligence and Computing, Denarau Island, 2022. 365–378
144 Zhang W, Zhang Z, Chao H C. Cooperative fog computing for dealing with big data in the internet of vehicles: architecture and hierarchical resource management. IEEE Commun Mag, 2017, 55: 60–67
145 Mao Y, You C, Zhang J, et al. A survey on mobile edge computing: the communication perspective. IEEE Commun Surv Tut, 2017, 19: 2322–2358
146 Cao X, Wang F, Xu J, et al. Joint computation and communication cooperation for energy-efficient mobile edge computing. IEEE Int Things J, 2019, 6: 4188–4200
147 Seid A M, Boateng G O, Anokye S, et al. Collaborative computation offloading and resource allocation in multi-UAV-assisted IoT networks: a deep reinforcement learning approach. IEEE Int Things J, 2021, 8: 12203–12218
148 Zhao B, Liu J, Wei Z, et al. A deep reinforcement learning based approach for energy-efficient channel allocation in satellite Internet of Things. IEEE Access, 2020, 8: 62197–62206
149 Chen M, Challita U, Saad W, et al. Artificial neural networks-based machine learning for wireless networks: a tutorial. IEEE Commun Surv Tut, 2019, 21: 3039–3071
150 Sun Y, Peng M, Zhou Y, et al. Application of machine learning in wireless networks: key techniques and open issues. IEEE Commun Surv Tut, 2019, 21: 3072–3108
151 Liu Y, Bi S, Shi Z, et al. When machine learning meets big data: a wireless communication perspective. IEEE Veh Technol Mag, 2020, 15: 63–72
152 Wang X, Han Y, Wang C, et al. In-edge AI: intelligentizing mobile edge computing, caching and communication by federated learning. IEEE Netw, 2019, 33: 156–165
153 Saad W, Bennis M, Chen M. A vision of 6G wireless systems: applications, trends, technologies, and open research problems. IEEE Netw, 2020, 34: 134–142
154 Yang Z, Chen M, Saad W, et al. Energy efficient federated learning over wireless communication networks. IEEE Trans Wireless Commun, 2021, 20: 1935–1949
155 Wang G H, Yang F, Song J, et al. Federated reinforcement learning with constellation collaboration for dynamic laser inter-satellite link scheduling. In: Proceedings of IEEE International Conference on Communications, Denver, 2024. 3815–3820

156 Yang H H, Liu Z, Quek T Q S, et al. Scheduling policies for federated learning in wireless networks. IEEE Trans Commun, 2020, 68: 317–333
157 Feng Q, Sun J, Zhang K, et al. Energy efficient federated learning joint communication and computation framework over wireless networks. In: Proceedings of the 4th International Conference on Computer Engineering and Intelligent Control, Guangzhou, 2023. 351–355
158 Vonrueden L, Mayer S, Beckh K, et al. Informed machine learning—a taxonomy and survey of integrating prior knowledge into learning systems. IEEE Trans Knowl Data Eng, 2023, 35: 614–633
159 Sun R J, Wen T S, Yang H, et al. Knowledge-driven resource management for 6G networks: a survey. Radio Commun Technol, 2022, 48: 630–637
160 Li D, Xu Y, Zhao M, et al. Knowledge-driven machine learning and applications in wireless communications. IEEE Trans Cogn Commun Netw, 2022, 8: 454–467
161 Zhao H, Zhao J, Qiu J, et al. Cooperative wind farm control with deep reinforcement learning and knowledge-assisted learning. IEEE Trans Ind Inf, 2020, 16: 6912–6921
162 Zheng Y, Chen H, Duan Q, et al. Leveraging domain knowledge for robust deep reinforcement learning in networking. In: Proceedings of the IEEE Conference on Computer Communications, Vancouver, 2021. 1–10
163 Liu X L, Xu H Y, Chen J M, et al. Neighborhood combination search for single-machine scheduling with sequence-dependent setup time. J Comput Sci Technol, 2024, 39: 737–752
164 Liu Z, Liu J, Liu X, et al. Knowledge-assisted adaptive large neighbourhood search algorithm for the satellite-ground link scheduling problem. Comput Indust Eng, 2024, 192: 110219
165 Slowik A, Kwasnicka H. Evolutionary algorithms and their applications to engineering problems. Neural Comput Applic, 2020, 32: 12363–12379
166 Li W, Wang R, Zhang T, et al. Reinvestigation of evolutionary many-objective optimization: focus on the Pareto knee front. Inf Sci, 2020, 522: 193–213
167 Song Y, Xing L, Wang M, et al. A knowledge-based evolutionary algorithm for relay satellite system mission scheduling problem. Comput Indust Eng, 2020, 150: 106830
168 Zhang J, Xing L. An improved genetic algorithm for the integrated satellite imaging and data transmission scheduling problem. Comput Oper Res, 2022, 139: 105626
169 Yao X, Pan X, Zhang T, et al. Knowledge-guided evolutionary algorithm for multi-satellite resource scheduling optimization. Future Generation Comput Syst, 2024, 156: 130–141
170 Hu X, Zhang Y, Liao X, et al. Dynamic beam hopping method based on multi-objective deep reinforcement learning for next generation satellite broadband systems. IEEE Trans Broadcast, 2020, 66: 630–646
171 Gu Z, She C, Hardjawana W, et al. Knowledge-assisted deep reinforcement learning in 5G scheduler design: from theoretical framework to implementation. IEEE J Sel Areas Commun, 2021, 39: 2014–2028
172 Zhu Z, Lin K, Jain A K, et al. Transfer learning in deep reinforcement learning: a survey. IEEE Trans Pattern Anal Mach Intell, 2023, 45: 13344–13362
173 Zhou I, Makhdoom I, Shariati N, et al. Internet of Things 2.0: concepts, applications, and future directions. IEEE Access, 2021, 9: 70961–71012
174 Abkenar F S, Ramezani P, Iranmanesh S, et al. A survey on mobility of edge computing networks in IoT: state-of-the-art, architectures, and challenges. IEEE Commun Surv Tut, 2022, 24: 2329–2365
175 Tamada K, Kawamoto Y, Kato N. Bandwidth usage reduction by traffic prediction using transfer learning in satellite communication systems. IEEE Trans Veh Technol, 2024, 73: 7459–7463
176 Wang L, Hu X, Ma S, et al. Dynamic beam hopping of multi-beam satellite based on genetic algorithm. In: Proceedings of IEEE International Conference on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking, Exeter, 2020. 1364–1370
177 Lei L, Eva L, Yuan Y, et al. Deep learning for beam hopping in multibeam satellite systems. In: Proceedings of the 91st Vehicular Technology Conference, Antwerp, 2020. 1–5
178 Chen Q, Ma S, Duan R, et al. A novel beam hopping resource allocation scheme of low earth orbit satellite based on transfer deep reinforcement learning. J Electron Inf Technol, 2023, 45: 407–417
179 6GANA. 6G AlaaS Requirements Whitepaper. 6GANA TG1. 2023

Appendix A

Table A1 List of abbreviations.

Abbreviation	Definition	Abbreviation	Definition
3GPP	3rd generation partnership project	KA-ALNS	Knowledge-assisted adaptive large neighborhood search
5G	Fifth generation	KBGA	Knowledge-based genetic algorithm
6G	Sixth generation	LEO	Low earth orbit
ACO	Ant colony optimization	MIMO	Multiple-input multiple-output
AI	Artificial intelligence	MIPS	Microprocessor without interlocked pipeline stages
AIaaS	AI as a service	MEC	Mobile edge computing
ALNS	Adaptive large neighborhood search	MINLP	Mixed integer nonlinear programming
API	Application programming interface	MAD3QL	Multi-agent double and dueling deep Q-learning
ARM	Advanced RISC machines	MDP	Markov decision process
CCS	Cohesive clustered satellites	MADRL	Multi-agent deep reinforcement learning
CPU	Central processing unit	NFV	Network functions virtualization
CSI	Channel state information	NTN	Non-terrestrial networks
CMDP	Constrained Markov decision process	NCC	Network control center
CNN	Convolutional neural network	NP	Non-deterministic polynomial
CDRS	Cross-domain resource scheduling	NB-IoT	Narrowband Internet of Things
CD-DMRS	Cross-domain dynamic multi-resource scheduling	PowerPC	Performance optimization with enhanced RISC-performance computing
DSS	Distributed satellite system	PSO	Particle swarm optimization
DTP	Digital transparent processor	QoS	Quality of service
DSP	Digital signal process	QoE	Quality of experience
DPS	Dynamic point selection	RL	Reinforcement learning
DL	Deep learning	RUR	Resource utilization ratio
DQL	Deep Q-learning	S&F	Store and forwarding
DQN	Deep Q-network	SAGIN	Space-air-ground integrated network
DRL	Deep reinforcement learning	SAGSIN	Space-air-ground-sea integrated network
DDPG	Deep deterministic policy gradient	SDN	Software defined network
EIRP	Equivalent isotropic radiated power	SSO	Sun-synchronous orbit
FPGA	Field programmable gate array	SPARC	Scalable processor architecture
FL	Federated learning	SL	Split learning
FTP	File transfer protocol	SAT-IoT	Satellite Internet of Things
GEO	Geostationary earth orbit	SMU	Spectrum management unit
GPU	Graphics processing unit	SGLs	Satellite-ground links
GW	Gateway station	SGLSP	Satellite-ground link scheduling problem
GBR	Guaranteed bit rate	TT&C	Tracking, telemetry and command
HPSC	High-performance space computing	TL	Transfer learning
HSRR	Hierarchical sparse resource representation	UE-SAT-UE	User equipment–satellite–user equipment
ISL	Inter-satellite links	UE	User equipment
IoT	Internet of Things	UoI	Utility of information
JCC-SAGIN	Joint communication and computing-embedded space-air-ground integrated network	VoIP	Voice over Internet protocol